Statistical Modeling of a Fluorescent Tube Coating Process

Michael J. Rauh

Department of Electrical Engineering
Carnegie Mellon University

Charles P. Neuman

Department of Electrical Engineering
Carnegie Mellon University

Francis C. Wimberfy

The Robotics Institute
-Carnegie Mellon University
Pittsburgh, Pennsylvania

May 1983

The research **was supported** by «<Jram(romt;,c ,/5sli,!ghouse CorporaWon

Table of Contents

1. The Coating Process 5
   1.1. Introduction 5
   1.2. Overview of the Report 6
   1.3. The Coating Process 6
      1.3.1. The Wash Process 7
      1.3.2. The Paint Coating Process 8
      1.3.3. Drying Process 9
      1.3.4. Lehring Process 10
      1.3.5. Inspection 11
2. Statistical Study of the Process 13
   2.1. Introduction 13
   2.2. Study Design 14
      2.2.1. Process Variables 14
         2.2.1.1. The Output Variables 14
         2.2.1.2. Input Variables 15
      2.2.2. Study Plan 15
   2.3. Data Collection 17
      2.3.1. Output Variable Measurement 17
      2.3.2. Input Variable Measurement 17
   2.4. Interpolation 19
   2.5. Summary 20
3. Data Analysis 21
   3.1. Introduction 21
   3.2. Descriptive Statistics 21
   3.3. Optical Density as a Function of the Process Inputs 22
      3.3.1. Bivariate Correlation 22
      3.3.2. Linear Regression Analysis 23
         3.3.2.1. Finding the Model from Experimental Data 24
         3.3.2.2. Testing the Significance of the Regression Coefficients 26
         3.3.2.3. Forward Entry Regression Procedure 27
         3.3.2.4. Results of Linear Regression Analysis for November Data 29
         3.3.2.5. Interpretation of the SPSS Results 31

3.4. Visual Defects as a Function of the Process Inputs 34
3.4.1. Discriminant Analysis 34
3.4.1.1. Obtaining the Discriminant Function 34
3.4.1.2. Ranking and Testing the Discriminant Functions 36
3.4.1.3. Ranking die Independent Variables in the Discriminant Function 37
3.4.2. The SPSS DISCRIMINANT Subprogram 37
3.4.3. Discriminant Analysis of the Data 38
3.4.3.1. Dichotomized Visual Defects 39
3.4.3.2. Visual Defects by Groups 39
3.4.3.3. Interpretation of Discriminant Analysis Results 40
3.5. Summary 41

4. Design of a Controlled Experiment 43

4.1. Hypothetical Model of the Coating Process 43
4.1.1. Optical Density 43
4.1.2. Visual Defects 44
4.2. Design of the Experiment 44
4.2.1. Classifying the Independent Variables 44
4.2.2. Measured Variables 45
4.2.3. Experiment Design 46
4.2.3.1. Full Factorial Design 46
4.2.3.2. Minimal Incomplete Design 47
4.2.3.3. Fractional Factorial Design 50
4.3. Carrying out the Experiment 51
4.3.1. Hardware 51
4.3.1.1. Sensors 52
4.3.1.2. Data Logging 53
4.3.2. Scheduling the Experiment 54
4.4. Analysis of the Results of the Controlled Experiment 55
4.4.1. The Optical Density Experiment 55
4.4.2. The Visual Defects Experiment 55
4.5. Summary 56

5. Summary 59

5.1. Objectives of the Project 59
5.2. The Study of the Coating Process 59
5.3. Analysis of the Study Data 60
5.3.1. Variance in the Data 60
5.3.2. Optical Density 60
5.3.3. Visual Defects 61
5.4. The Controlled Experiment 61
5.5. Interpretation of the Results of the Controlled Experiment 62

List of Figures

Figure 1-1:	The Coating Process	7
Figure 1-2:	The Wash Process	8
Figure 1-3:	The Paint Coating Process	9
Figure 1-4:	The Drying Process	10
Figure 1-5:	The Lehring Process	11
Figure 2-1:	Luminosity vs. Optical Density	14
Figure 3-1:	Percentage of variation in Y explained by X_1 and X_2 where X_1 and X_2 are correlated.	32
Figure 4-1:	A Two Level Experiment	48
Figure 4-2:	Responses in a Two Level Experiment	50

List of Tables

Table 2-1: Types of Visual Coating Defects 16
Table 2-2: Variables Thought to Determine Coating Properties 17
Table 2-3: Time Delays for Input Variables 20
Table 3-1: Percentage of Variation for each Study Variable 21
Table 3-2: Correlation of Input Variables with Optical Density 24
Table 3-3: Forward Entry Results 29
Table 3-4: Reduced Model for Optical Density 30
Table 3-5: Correlation Matrix For November 6 Data 33
Table 3-6: Frequency of Occurrence of Visual Defects 38
Table 3-7: Dichotomized Visual Defects 39
Table 3-8: Analysis Results With Six Groups 40
Table 4-1: Classes of Independent Variables 45
Table 4-2: Treatment Levels 46
Table 4-3: Eight Observation Experimental Design 47
Table 4-4: 32 Observation Design 50
Table 4-5: Suggested Sensors for Automatic Data Acquisition 53

Abstract

The coating of a fluorescent lamp with fluorescent paint is an example of a complex industrial process. Improved control of this process could lead to reduction in the cost of producing a lamp. Modeling the process is necessary for improved control. As a first step, a study of the coating process at the Westinghouse Fairmont Works in Fairmont, West Virginia has been made. The study included two criterion, or dependent, variables and 12 predictor, or independent variables. Analysis of the study data has produced a linear regression model with five independent variables which accounts for 58% of the variation in coating thickness. Also, a set of linear classification functions has been found which correctly classify 92% of visual defects from 12 input variables, using the training data.

These preliminary models have been used to design a controlled experiment. The controlled experiment will allow the significance of seven independent variables in determining optical density and visual defects to be established conclusively.

Acknowledgements

Thanks are due to Professor Luc Tierney of the CMU Statistics Department who suggested the fractional factorial design described in Chapter 4. George Preston and Charles Trushell of the Westinghouse Lamp Division provided background information about the coating process. George Preston acted as the contact with the production personnel at the Fairmont Works. Bill Tarleton and Charles Moore of the Fairmont Works provided information and assistance in carrying out on-site work. Finally, George Preston, Henry Stone, John Schlag, and Ray and Kathy Horton deserve special mention for putting in two very long days collecting data.

Chapter 1
The Coating Process

1.1. Introduction

The Carnegie-Mellon University Robotics Institute research effort to study the fluorescent lamp coating process at the Fairmont Works of the Westinghouse Electric Company is aimed at modeling and developing a control system for the process and providing data for an Intelligent Management System. It is believed that improved control of the process could result in substantial benefits including reduction of material loss, improved performance of the finished lamps, and reduced training costs for machine attendants [Wimberly 81]. The objective of this project is to develop a statistical input-output model from production data. This model is intended to be static rather than dynamic. Such a model is appealing because detailed knowledge of the physical processes involved is not required. Standard statistical methods are applied to develop the model from records of the input and output variables. The model is designed to serve as the basis for the development of a dynamic model and computer controlled system.

The modeling effort entails three steps. The first step is to study the process and identify the relevant input and output variables. The second step is to design an experiment to provide the production data. This step includes both the planning of the experiment and the specification of the instruments required to carry it out. The third step is to conduct the experiment and analyze the results. This report covers the work done from September 1981 to May 1982, and summarizes completion of the first two steps.

The work described in this report contributes to the Robotics Institute study of the coating process in several ways. First, it provides a written description of the process. It describes the use of standard statistical methods to make a static input-output model. It identifies several input variables which should be investigated by a controlled experiment. An experimental design is suggested which offers a good compromise between simplicity and thoroughness. Finally, off-the-shelf instrumentation required for the experiment is described. It is the author's hope that this report shows that

experimental modeling of the coating process is practical, and that others who may work on making such a model will find the information provided here useful.

The ultimate goal of this work is a control system for the coating process. The static model which is the topic of the present work can be used to determine set-points for such a control system. The process variables would then be maintained to these set points by individual closed-loop systems. The set points are derived from the regression model developed in Chapter 3. The regression model does not imply that any combination of the process variable values is acceptable; rather, the model is valid over only a limited range of values. The nominal value of the process variable over which the regression model is calculated can be taken as the starting point. These values can be adjusted gradually until the regression model is satisfied. The process variable values which satisfy the model can then be used as set-points. Closed-loop control requires a dynamic model of the process. Such a model can be developed from observations of the input and output variables of the process. The static model described in subsequent chapters can suggest which variables arc significant, paving the way for future development of a dynamic model and feedback control system. The issues are amplified in [Box 70].

1.2. Overview of the Report

In this chapter a description of the. coating process on Line 1 at Fairmont, West Virginia is presented. Chapter 2 describes a preliminary study of the process made in November of 1981. Chapter 3 contains a review of the statistical methods used to analyze the study data, and the preliminary models derived arc presented. In Chapter 4 a design for a controlled experiment and the requirements for hardware to carry it out are set forth. Conclusions and a summary of the report are found in Chapter 5.

1.3. The Coating Process

The coating of a fluorescent lamp with fluorescent paint is a multi-step process. A block diagram of the process is shown in Figure 1-1. Each of the steps is discussed in turn. The process begins with an tincoated glass tube, open at each end The tubes are suspended vertically on a conveyor (chain) by an operator. From this point the tubes are washed, coated with paint, dryed, etched with a trademark, and baked The glass is treated with sulfur dioxide gas to reduce friction between tubes* die ends of the tubes are brushed, and the tubes are then inspected visually for defects. Except for mounting and inspection, each step of the process is completely automatic. Tubes pass through the process at a rate of several thousand per hour.

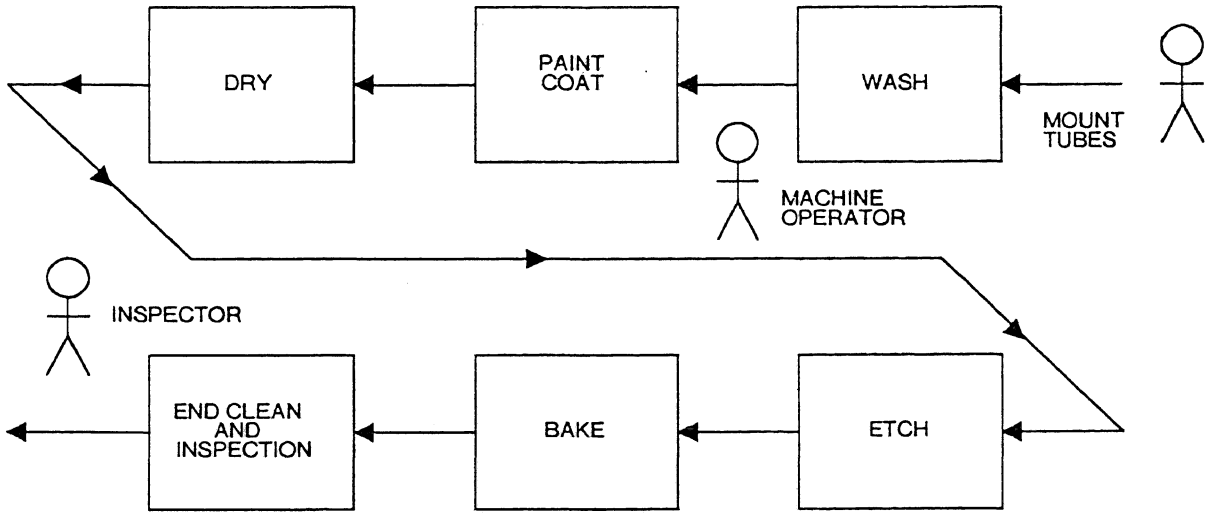


Figure 1-1: The Coating Process

The quality of the coated tubes is accessed by coating thickness and freedom from visual defects. Coating thickness can vary from tube to tube, and from end to end in an individual tube. Coating which is too thick or too thin results in poor performance of the completed lamp, and, in the case of a thick coating, waste of fluorescent paint. The presence of visual defects in the coating, such as uncoated areas, bubbles, or veils is unacceptable.

1.3.1. The Wash Process

In the first step of the coating process, the tubes are washed to remove dust that may have accumulated during their storage. The tubes enter the wash enclosure end up (Figure 1-2). Hot wash water (190 °F) is discharged from nozzles above the tubes. The wash water is made up of de-ionized water and a surfactant. De-ionized water is added from time to time to make up lost volume. After flowing over the tubes, the wash water falls into a tank where it is re-heated and pumped back through the nozzles. As time passes, impurities collect in the water. The water is changed daily to limit the buildup. The tubes pass through the wash in about 15 seconds and then pass to the paint coating area. The passage takes about three minutes. During this time, excess water drips out and the tubes are at least partially dry by the time they reach the paint coating turret. A blast of compressed air is directed onto the tube hanger to remove water caught between the hanger and tube. Impurity of the wash water is monitored by a conductivity meter. Should the conductivity rise above a setpoint, the machine operator adds de-ionized water to the wash. Wash water temperature is monitored by a thermocouple and controlled by a comercially available controller.

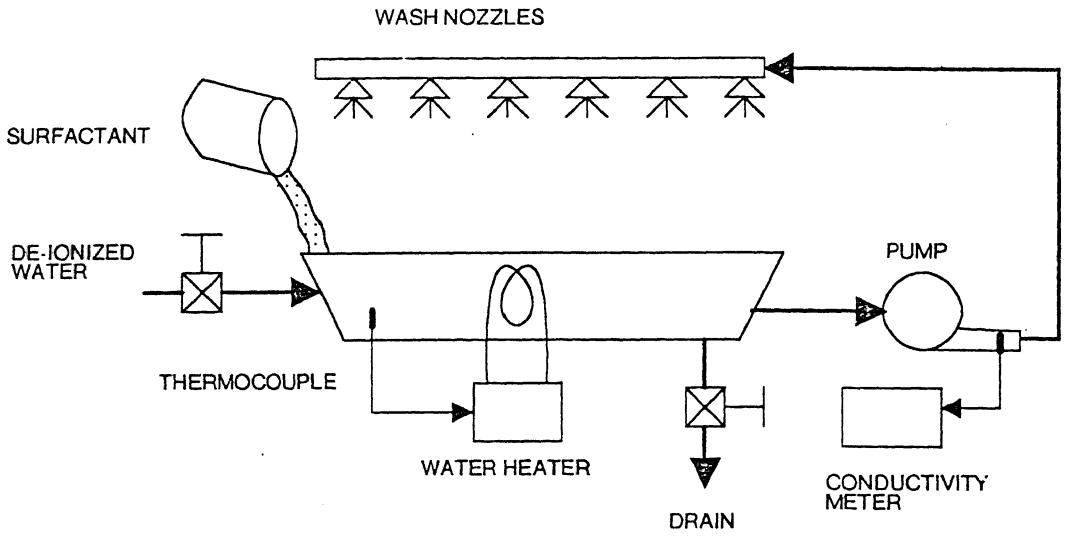


Figure 1-2: The Wash Process

1.3.2. The Paint Coating Process

The paint coating system shown in Figure 1-3 consists of a mixing tank, a line tank, a delivery pump, a paint turret and a recovery basin. The paint is mixed in 100 gallon lots in a mixing tank. The paint is made up of de-ionized water and a water-soluble lacquer. The lacquer is added to produce the desired viscosity. The fluorescent material is added as a powder. The paint then consists of fluorescent paint particles held in suspension in the water-lacquer mixture. Paint is periodically transferred to the line tank from the mixing tank by an electric pump.

When the paint reaches the coating line, a portion is bled off to the drip basin, and an additional sample is drawn away to a viscosimeter. The remainder is passed through a mesh filter and flows to a reservoir directly above the paint turret.

The paint turret consists of a number of nozzles, each fitted with a cam actuated valve. As the turret rotates, a tube, still in vertical position, is transferred from the conveyor to the turret. The valve actuator senses that a tube is in place, and the valve is opened for a fixed period of time. Paint flows from the valve into the top of the tube, and then flows by gravity, coating the entire interior surface of the tube. The valve is closed again, and the turret rotates back to the conveyor. The tube is transferred back to the conveyor and passes to the drying hood.

Beneath the turret and drying hood lies the drip basin. Excess paint drips from the tube and is

strikes the tube. This is done by varying the slot width and placing wire mesh in the slot and results in air velocities which vary over a range of a few hundred feet per minute. The purpose of varying the velocity is to achieve uniform end-to-end thickness of the paint, which tends to creep down the walls of the tube. The drying process is depicted in Figure 1-4.

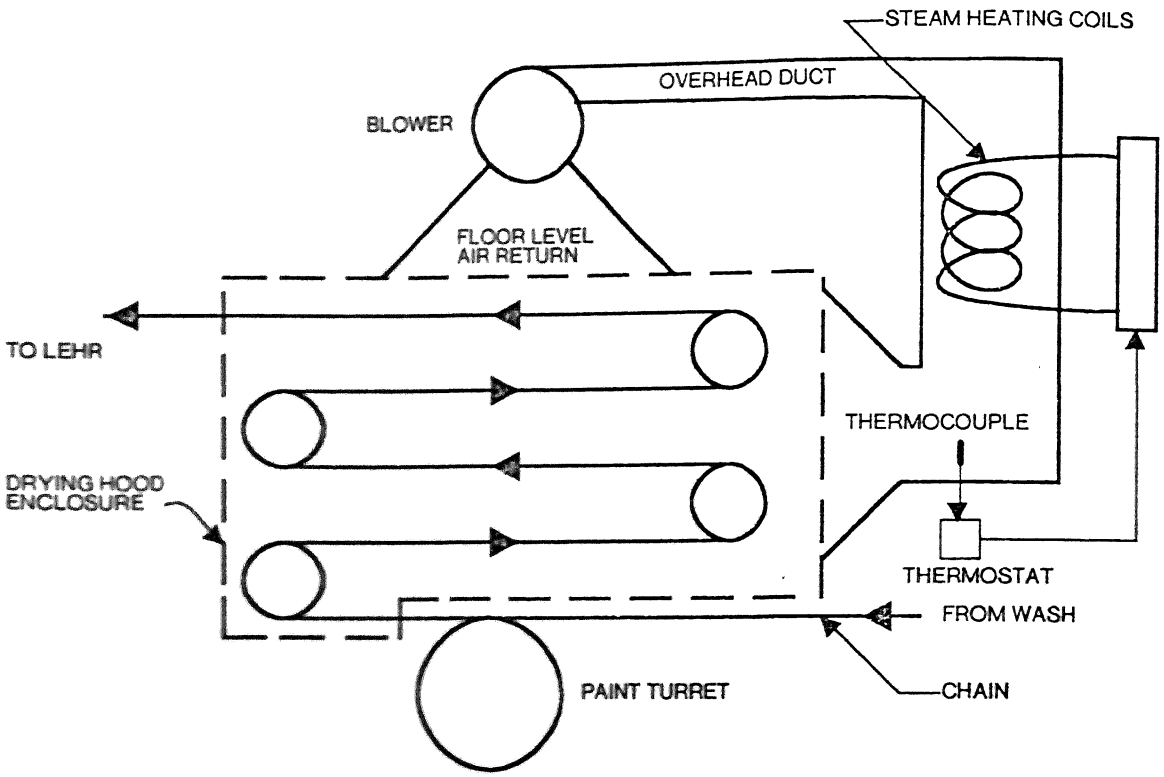


Figure 1-4: The Drying Process

The drying hood is partially enclosed. A return duct gathers the air from the floor of the hood area. The air is then propelled by a large blower through a duct where it is heated. It then passes overhead to a set of manifolds and ultimately back to the slots. The duct work is fairly air tight, but there are openings under the hood where the tubes pass in and out and these openings allow some outside air to enter.

1.3.4. Lehring Process

The Lehring Process is depicted in Figure 1-5. When the tube exits the drying hood, the paint is solidified. The tubes are released from the chain, and fall onto a ramp. From the ramp, they roll onto a conveyor which carries them past the etching station where one end of the tube is marked with a trademark and trade-name.

-uftcs 'he tube TTiis is done by varying the slot width and placing wire mesh in the slot and rest •r air velocities which vary over a range of a few hundred feet per minute. The purpose of varyi Ac velocity is to achieve uniform end-to-end thickness of me paint, which tends to creep down < Jails of the tube. The drying process is depicted in Figure 1-4

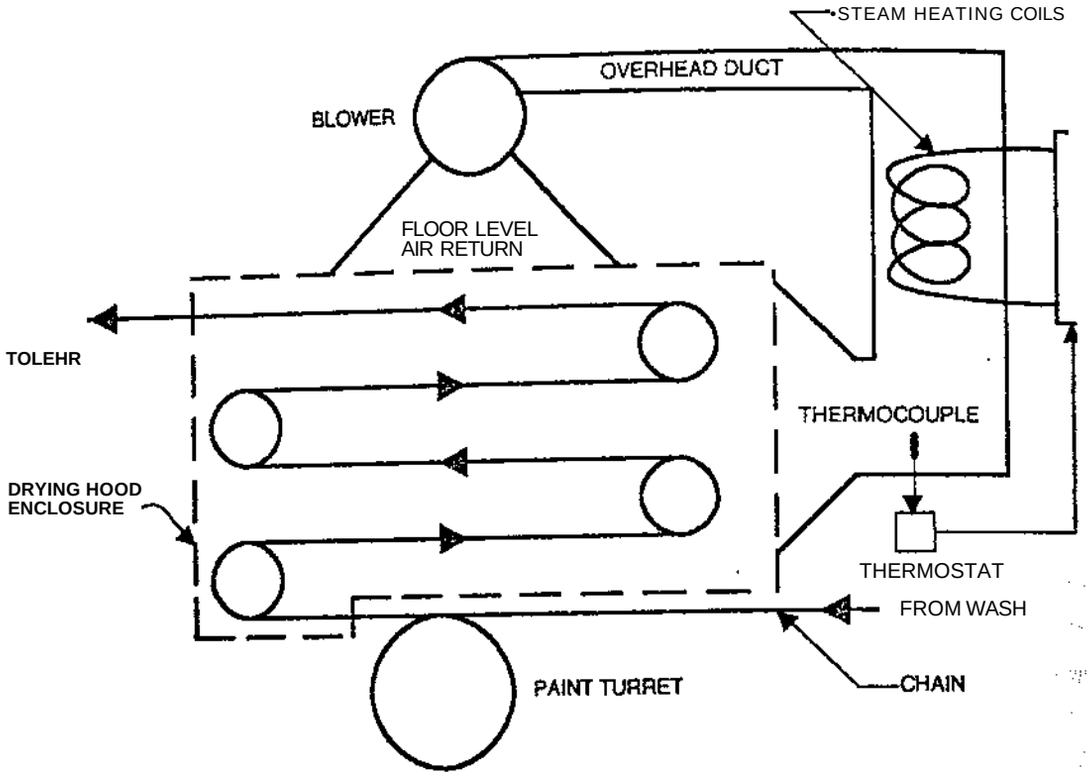


Figure 1-4: The Drying Process

The drying hood Is partially enclosed. A return duct gathers the air from the floor of the hood am The air is then propelled by a large blower through a duct where it is heated. It then passes oteritead to a set tf masifolds and ultimately back to the slots,. The duct work is fairly air tight, but. 3ere arc epeningi under the hood where the tubes pass in and out and these openings allow some &&&&£ airio enter.

1*3,4, ithring Process

The ielrisg Process» depicted la Figure 1-5. When the tube exits the drying hood, the paint is ¥M€\tL Be tubes are released from the chain, and fall onto a ramp. From the ramp, they rol Mto a aonleyor *hich tarries them *pm* the etching station where one end of the tube is marked will i trademark *mi* trade-name

The tubes then enter the Lehr oven. In the oven they are subjected to temperatures of over 1000 °F. Air is blown into the ends of the tubes to create an oxidizing atmosphere. Binders in the paint are burned away, leaving only the fluorescent material. (This oxidation is separate from that of the combustible gas fuel, which is carbureted before entering the oven.) Near the end of the Lehr, sulfur dioxide gas is introduced. The gas reacts with the glass on the surface of the tube, imparting a quality of slipperiness called "lubricity" to it.

Upon leaving the Lehr, the tubes are cooled. They then pass by a series of brushes. The brushes contact the ends of the tubes and remove the paint from the inner surface around the edge, or "collar", to present a clean surface to be fused to the filament mount.

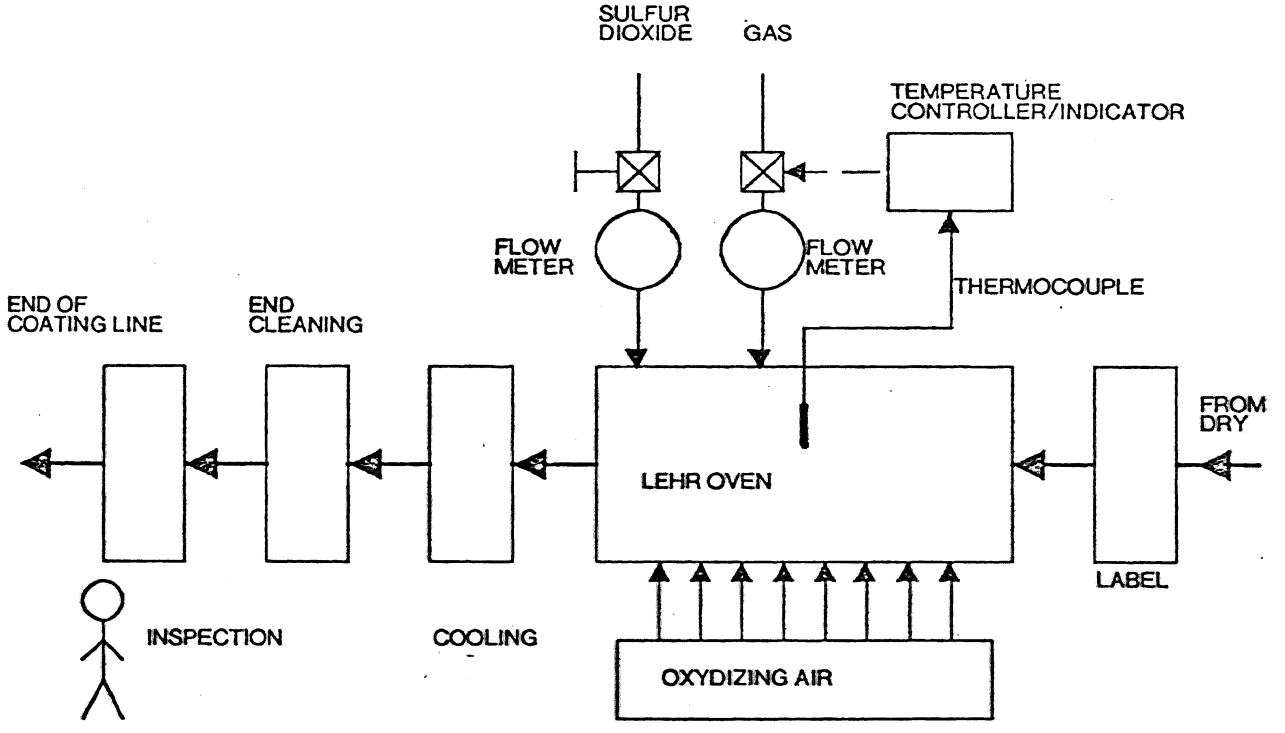


Figure 1-5: The Lehring Process

1.3.5. Inspection

At the end of the conveyor, the tubes roll over a lighted table. An inspector checks the tubes for defects and removes those tubes in which defects are visible. The tubes then pass onto another conveyor where they are transferred to another line for further processing.

Chapter 2
Statistical Study of the Process

2.1. Introduction

The study of the coating process is designed to answer the following questions:

1. What is the range and variability of the input and output variables?

2. What precision is necessary to measure significant changes in the variables?

3. Which of the input variables are significant in determining the output variables, and therefore merit further study?

A study differs from a controlled experiment in that no attempt is made to control the input variables and operating conditions. Instead, the input and output variables arc observed and their numerical values are recorded. The data are then statistically analyzed to seek out relationships between the input and output variables.

A study is less powerful than a controlled experiment because:

1. A study cannot guarantee that all possible values of the variables are encountered, and therefore the resulting model may be restricted in its range of validity.

2. A study cannot guarantee that the independent variables are uncorrelated and is therefore subject to the problems associated with multicollinearity.

A study of the process is useful as an initial effort because it is much easier to carry out The study can then serve as the basis for choosing the variables and treatment levels for a controlled experiment *In* this chapter the study design, including the identification of the variables to be measured* the method of measurement, and special considerations for time delays in the process are considered. The study data are analyzed in Chapter 3.

2.2. Study Design

Design of the study includes identification of the process variables to be measured and formulati of the plan for carrying out those measurements.

2.2.1. Process Variables

2.2.1.1. The Output Variables

There are two output variables of interest: optical density and visual defects. Optical density is measure of the thickness of the coating of fluorescent paint on the inside of the tube. It is of intere because it affects both the luminosity and the lifetime of the fluorescent lamp. The effect of optic density on initial luminosity is shown in Figure 2-1.

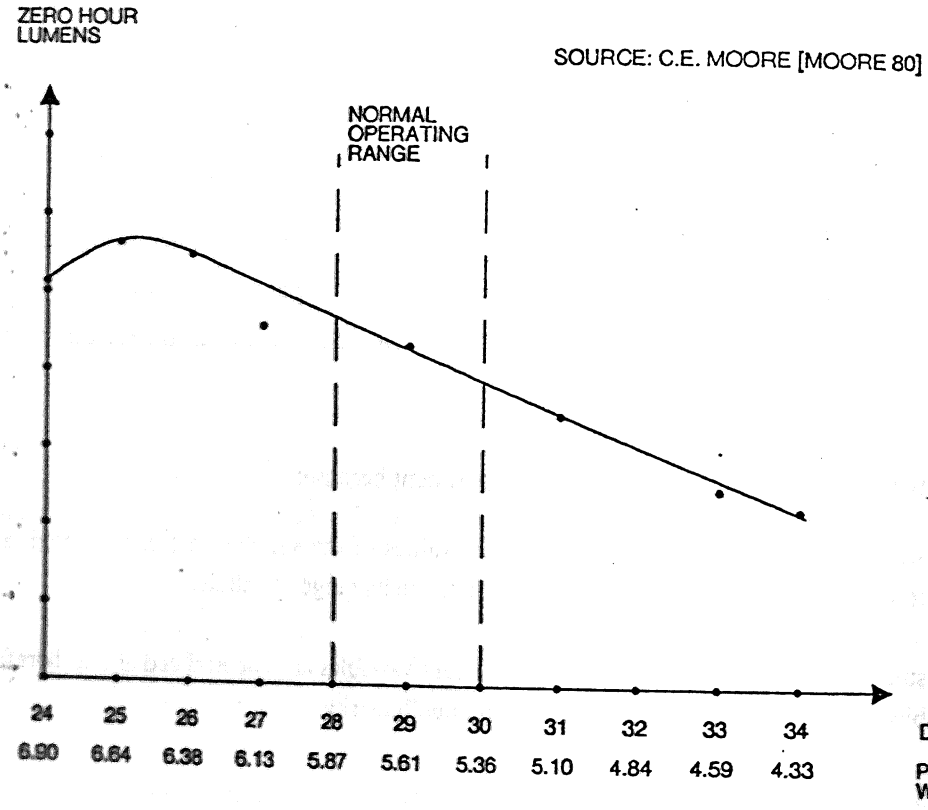


Figure 2-1: Luminosity vs. Optical Density

Optical density is measured with an instrument made by Westinghouse. It consists of an incandescent lamp and a photocell enclosed in a box. The tube to be measured is inserted manually and a reading is taken from a panel meter. The output is the current through the photocell as

measured by a milliammeter. The higher the reading, the thinner the coating; the instrument actually measures the transmittance of the tube. The scale is arbitrary and although the reading is in milliamperes, optical density is a dimensionless number. The correlation between the weight (amount) of fluorescent powder in the tube and its optical density as measured on the Westinghouse instrument is known. (Analysis of data for a typical set of tubes produced the relation

$$\text{powder weight} = -0.25646 \text{ optical density} + 13.05165$$

where powder weight is in grams and optical density is in optical density units. The data were supplied by C.E. Moore [Moore 80].) The correlation varies among paint lots. A correction factor, known as the mill factor, can be applied to the correlation to compensate for the variation[1]. The nominal value of optical density is 29 ± 1 optical density units.

"Visual defects" is a term used to describe a number of problems which make the tube unacceptable for use. They are detected visually by a human inspector at the end of the coating line. The various types of visual defects are enumerated in Table 2-1, along with their suspected causes.

2.2.1.2. Input Variables

The input variables are the process variables which are thought to affect the outcome (optical density and visual defects). Some are measured and controlled during normal operation; others are not. The list of input variables in Table 2-2 was chosen in consultation with Westinghouse personnel. Table 2-2 lists all of the variables which are believed to affect the outcome.

2.2.2. Study Plan

The study was carried out by a group of CMU students and researchers, and Westinghouse engineers. The study is of the sample survey type [Neeter 78]. With the coating line in normal operation, tubes are chosen randomly at a rate of about one per minute from the end of the coating line. The optical density of the tube at one point near its center is read and recorded, along with the time of the observation. Concurrently, measurements of the input variables are made and the measured values recorded. A data table of the optical density reading and corresponding input variable values is then built up.

Visual defects are studied concurrently with optical density. Each tube rejected by the inspector at

---

[1] The existence of the mill factor was pointed out by Westinghouse Engineer George Preston in a meeting at CMU on July 2, 1982.

Table 2-1: Types of Visual Coating Defects

CODE NUMBER	VISUAL DEFECT	SUSPECTED CAUSE
1	Bubbles	Obstruction in paint pump vanes Agitator on in Mill Room Low volume in Mill Room Improper head closing on turret Too much surfactant (small bubbles) Too little surfactant (large bubbles) Insufficient bubble breaking air Insufficient bubble breaker
2	Streaks	Wash water dirty Too much surfactant Drying conditions not proper Tubes not hanging straight
3	Short Coat (top)	Insufficient paint in system Dirty filters
4	Texture	Improper milling Not enough surfactant or defoaming agent Too much surfactant or defoaming agent
5	Hanger Marks	Too much heat Air blast too weak to blow wash water off hanger (water marks)
6	Partial Coat	Improper head adjustment Paint too heavy Paint too thin Dirty wire mesh filter
7	Density	Line stops Paint too thin Paint too heavy Drying pattern
8	Thin End (top)	Bubble breaking air Drying pattern

the end of the coating line is noted along with the reason for and time of the rejection. The defec and the corresponding input variable values are added to the data table. This process continued fo several hours on each of two days, until the data table contained several hundred observations.

At the end of the study, the data are analyzed statistically. What is sought are relationship

Table 2-2: Variables Thought to Determine Coating Properties

VARIABLE	MEASUREMENT (During Normal Production)	CONTROL
1. Wash Water Temperature	Continuously	Automatic
2. Wash Water Conductivity	Continuously	Manual (Corrected Daily)
3. Paint Viscosity	Continuously	Semi-Automatic
4. Paint Specific Gravity	Periodically	Manual
5. Paint pH	Not Measured	None
6. Drying Air Temperature	Continuously	Automatic
7. Drying Air Velocity	Not Measured	None
8. Drying Air Humidity	Not Measured	None
9. Percent Excess Oxygen in Lehr Gas-Air Mixture	Not Measured	None
10. Natural Gas Flow Rate	Continuously	None
11. Lehr Oven Temperature	Continuously	Automatic
12. Sulfur Dioxide Flow Rate	Continuously	None

between the input and output variables. The range and mean value of the variables in the survey, useful for specifying measuring instruments and judging the physical, as opposed to statistical, significance of observed variation, are also computed.

2.3. Data Collection

2.3.1. Output Variable Measurement

The methods used to measure each of the output variables in the study were as follows:

1. Optical Density: Tubes were selected randomly at the end of the coating line at one minute intervals. Optical density was measured with the Westinghouse instrument and recorded with a precision of 0.1 optical density units.

2. Visual Defects: All visual defects noted by the inspector were recorded.

2.3.2. Input Variable Measurement

The methods used to measure each of the input variables in the study were as follows:

1. Wash Water Temperature: An existing thermocouple in the wash water tank was used. Readings were taken at the rate of one every two minutes on the first day and one every five minutes on the second day with a precision of 0.5 °F.

2. Wash Water Conductivity: The existing conductivity meter was used. The probe was mounted in the return pipe from the pump. Readings were taken every two minutes on the first day and every five minutes on the second day. A precision of 0.5 /xS/cm was used.

3. Paint Viscosity: The on-line viscosimcter was used. The probe was mounted in a chamber filled in parallel to the paint turret. Readings were taken about every nine minutes the first day, and every three and one-half minutes the second with a precision of 0.1 cP.

4. Paint Specific Gravity: A hydrometer with a precision of 0.005 s.g.u. (gm/cm^3) was used. The sample was taken from the drip basin oscillator once every nine minutes the first day and once every six minutes the second day.

5. Paint pH: A Fisher Accumet Model 525 Digital pH Meter, with a precision of 0.001 pH, was used. Readings were recorded to 0.01 pH. The sample drawn for the specific gravity measurement was used. Readings were taken once every nine minutes the first day and once every six minutes the second day.

6. Drying Air Velocity: The instrument was an Anemotherm model 60 hot tip anemometer. Equipped with a hand-held probe, it is precise to 25 ft/min. The sample was taken at the first vent of the drying hood. Readings were taken once every two and one-half minutes the first day and once every five and one-half minutes the second.

7. Drying Air Temperature: A mercury-filled glass thermometer was read to a precision of 0.5 °C Air temperature was sampled at the same spot as air velocity ever)' two and one-half minutes the first day and once every five and one-half minutes the second day.

8. Drying Air Humidity: The wet-bulb dry-bulb method was used. Wet-bulb temperature was measured with a mercury-filled glass thermometer with its bulb covered with a dampened cloth. Readings were recorded with a precision of 0.1 °C. Wet bulb temperature was sampled at the same place and rate as drying air temperature.

9. Percent Excess Oxygen in Lehr Air-Gas Mixture: A Thermox was temporarily Installed in the fuel line to the Lehr oven. It was read every 2 minutes the first day and every one and three-quarter minutes the second day to a precision of 0.01% excess O_2-

10. Combustible Fuel Gas Flow Rate: A ball and tube type gage permanently installed o& the Lehr gas line was used. Readings precise to 25 ftVhr were recorded every two minutes the first day and every one and three-quarter minutes the second day.

11. Lehr Temperature: A thermocouple located inside the Lehr oven provides a temperature

signal for the oven temperature controller. The controller provides a panel meter read-out of temperature which can be read to 1.0 °C. Readings were recorded every two minutes the first day and every one and three-quarter minutes the second day.

12. Sulfur Dioxide Flow Rate: A ball and tube type gage which can be read to 0.005 ft^3/hr is permanently installed in the supply line to Lehr. Readings were recorded every two minutes the first day and every one and three-quarter minutes the second day.

Measurements were made and recorded with pencil and paper by the observers. One person was assigned to record optical density and visual defects; one to Lehr temperature, gas flow, SO_2 flow, and % excess O_2; one to humidity, air temperature and air velocity (and specific gravity on the second day); one to paint pH, viscosity, and specific gravity (and wash water temperature and conductivity the second day); and one to wash water conductivity and temperature. Each observer also recorded the time of each measurement to the nearest second. Time measurement was by wristwatches synchronized at the beginning of each day.

2.4. Interpolation

Because the process of coating a tube takes place over a 21 minute time period, the inherent process time delays must be taken into account. For example, the optical density of a tube is measured when it reaches the end of the line, 21 minutes after it is subjected to the wash process. Thus the current optical density reading must be paired with the wash water temperature and conductivity measurements made 21 minutes earlier.

To determine these time delays, measurements of the time interval from the point where the various input variables impinge on the tube to the end of the process were made. The average of two sets of measurements are listed in Table 2-3.

A second delay is introduced by the time required for the material (paint, wash water and drying air) whose qualities are measured to travel from the point of measurement to the point of impingement on the tube. These time delays range from a fraction of a second to several seconds. The effect of these delays on the study was judged to be negligible because the measurement times are likely to have errors of the same order of magnitude as these delays. Therefore, they are ignored in this study.

The data table described in Section 2.2.2 consists of a series of 457 cases. Each case consists of a

Table 2-3: Time Delays for Input Variables

PROCESS STEP	VARIABLES	TIME TO END OF PROCESS SECONDS	MINUTES
Enter Wash	Conductivity, Temperature	1213	20.2
Enter Painting	Viscosity, pH, Specific Gravity	1047	17.5
Enter Drying	Air temp., Humidity, Velocity	1006	16.8
Mid-Lehr	Gas, O_2, Temperature	141	2.4
End Lehr	SO_2 flow	96	1.6

dependent (output) variable and its associated independent (input) variables. The value of the independent variable is the value that existed *when that variable impinged on the tube.* The time of impingement is found by subtracting the delay time for that variable from the time of measurement of the output variable.

The value that existed when the input variable impinged on the tube is not necessarily recorded, since the period of each measurement was not controlled. More likely, the input variable impinged on the tube at a time which falls between the times of two recorded values. The input variable value for the data table is determined by a linear interpolation between the two recorded values whose measurement times bracket the time of impingement. The independent (input) variable values used in the analysis are therefore interpolated values.

2.5. Summary

A sample survey study of the coating process on Coating Line 1 at the Westinghouse Lamp Works at Fairmont, West Virginia was made on November 4thand 6thof 1981. Two output variables and twelve input variables were measured. The recorded values were used to calculate, by time shifting and interpolation, the values of the input variables at the time of impingement on a tube for which the output variables were measured. The study data are analyzed in Chapter 3.

Chapter 3
Data Analysis

3.1. Introduction

The objectives of this chapter arc to review and apply statistical methods to analyze the study data. Separate treatment of optical density and visual defects is necessary because the former is a continuous, ratio-level variable, arid the latter is a discrete, nominal-level variable. A standard statistical package, *Statistical Packagefor the Social Sciences* (SPSS) [Nie 75] is used for the analysis.

3.2. Descriptive Statistics

Because the variables were not controlled in the study, it is of interest to sec how much variation occurred. The percentage of variation is given by $\%V = oVJ\dot{x}lOO\%$ where $\bar{X}$ is die mean and a is the variance of the variable. The percentage of variation for each variable is shown in Table 3-1.

Table 3-1: Percentage of Variation for each Study Variable

VARIABLE	%V, NOV. 4	%V, NOV. 6
Optical Density	2.77	**3.08**
Wash Temperature	0.09	**0.32**
Wash Conductivity	3,96	**8.49**
Paint Specific Gravity	0.29	**0.15**
Paint Viscosity	2.62	**1.29**
Paint pH	0.18	**0.56**
Drying Air Velocity	4.32	**1.32**
Drying Air Temperature	1.25	**1.30**
Temperature Difference (Humidity)	4.62	**1.79**
% Excess Oxygen	2.49	**2.09**
Lehr Temperature	0.31	**0.44**
Lehr Gas Flow Rate	2.14	**2.31**
Sulfur Dioxide Flow	0.11	**0.73**

3.3. Optical Density as a Function of the Process Inputs

3.3.1. Bivariate Correlation

Investigation of the relationship between optical density and the independent variables begins with computing the simple correlation for each. The linear relation of a single independent variable X to a single dependent variable Y is often measured by Pearson's correlation coefficient

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

where

$$S_{xy} = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

$$S_{xx} = \sum_{i=1}^{n} (X_i - \bar{X})^2$$

$$S_{yy} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

and

$\bar{X}$= mean of $X = \frac{1}{n}\sum_{i=1}^{n} X_i$,
$\bar{Y}$= mean of Y,
X_i= i^{th}observation of X,
Y_i= i^{th}observation of Y, and
n= number of observations of X and Y. [Nie 75]

The correlation coefficient r takes on values from $+1$ to -1. The magnitude of r indicates the strength of the relation and the sign its sense.

The probability that the sample for which the correlation is computed is drawn from a population in which the true correlation is zero is called the significance of r, and is found by computing the t statistic [Nie 75]

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

The t statistic is distributed as the Student's t, $f(t,\nu)$, where ν is the number of degrees of freedom ($n-2$ in this case). The t distribution is similar to the normal distribution in shape, being slightly flatter and broader. As ν increases, the t distribution becomes indistinguishable from the normal distribution. The *significance* of r is

$$1 - \int_{-|t|}^{|t|} f(\tau,\nu)\, d\tau$$

which is called the two tailed test of significance; it does not assume that t is positive or negative. The significance is the probability that the population correlation $\rho = 0$ even though the sample correlation $r \neq 0$; i.e., that there is no systematic relation between the dependent and independent variable in the population at large even though such a relation was found in the sample.

Table 3-2 shows the bivariate correlation between optical density and each of the input variables, for each of the two days. The correlations were obtained by an SPSS *SCATTERGRAM* analysis. The correlations range from 0 (completely uncorrelated) to 0.6785 (moderately correlated). The results are nearly all significant to the 0.001 level; i.e., there is less than 1 chance in 1000 that the true correlation is 0. The high degree of significance is due to the large number of cases examined.

There is considerable disagreement in the results of the two days, both in magnitude and sign of the correlation coefficient. It is doubtful that the relationships actually changed so much over the space of two days. More likely, the data have serious systematic error. This error is most likely to have occurred on the first day, when the data takers were learning their jobs. The first day's data are therefore dropped in the subsequent discussion, and henceforth only the second day's data are considered.

3.3.2. Linear Regression Analysis

The next step in the analysis of the study data is a regression analysis. The analysis allows the independent variables to be ranked according to their *marginal* value in predicting optical density. That is, we can find the contribution of a variable to explaining variation in optical density over and above the variation already explained by other variables. This allows us to concentrate future efforts on those variables which appear to be most useful as predictors. Ultimately, the development of a regression model also allows us to predict the optical density which will result from a set of input variable values.

Table 3-2: Correlation of Input Variables with Optical Density

Independent Variable	Correlation Coefficient / (Significance)		
	Nov. 4 (53 Cases)	Nov. 6 (213 Cases)	Both (266 Cases)
Water Temp.	Insufficient Variation	0.0241 (0.726)	0.0800 (0.193)
Conductivity	-0.3978 (0.003)	0.4911 (0.000)	0.3630 (0.000)
Specific Gravity	-0.4001 (0.003)	-0.3219 (0.000)	-0.0962 (0.118)
Viscosity	-0.5158 (0.000)	-0.4102 (0.000)	-0.4023 (0.000)
pH	0.3715 (0.006)	-0.5109 (0.000)	-0.4283 (0.000)
Air Velocity	0.2422 (0.081)	0.0947 (0.168)	0.0970 (0.115)
Dry Bulb Temp.	0.0087 (0.951)	-0.6785 (0.000)	-0.1792 (0.003)
% Excess O_2	-0.5352 (0.000)	0.3251 (0.000)	0.1352 (0.027)
Lehr Temp.	0.2779 (0.044)	-0.1158 (0.092)	-0.0249 (0.686)
Gas Flow	0.5813 (0.000)	-0.2745 (0.000)	-0.0766 (0.213)
SO_2 Flow	0.0000 (1.000)	0.1517 (0.027)	0.0920 (0.134)
Temp. Difference (Humidity)	0.4882 (0.000)	-0.3144 (0.000)	-0.0118 (0.849)

3.3.2.1. Finding the Model from Experimental Data

In multiple linear regression [Neeter 74], the dependent variable Y is assumed to be linearly related to the independent variables $X_1, X_2, \ldots, X_n$ according to the regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{ip-1} + \varepsilon_i$$

where Y_i is the response to the i^{th}observation, $\beta_0,\beta_1,\ldots,\beta_{p-1}$ are parameters (the regression coefficients), $X_{i1},X_{i2},\ldots,X_{ip-1}$ are the known values of the i^{th}observation, and ε_i are independent $\mathcal{N}(0,\sigma^2)$ random errors.

The regression function is

$$E[Y]=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_{p-1}X_{p-1}$$

and estimated regression function is

$$Y_i'=b_0+b_1X_1+b_2X_2+\cdots+b_{p-1}X_{p-1}$$

where Y' is the value of the estimated regression function at the values of $X_1,X_2,\ldots,X_{p-1}$ and $b_0,b_1,\ldots,b_{p-1}$ are the sample estimates of the parameters.

The sample estimates, or regression coefficients, are chosen by the method of least squares. Let

$$Q=\sum_{i=1}^{n}(Y_i-Y_i')^2$$

be the sum of n squared deviations. The least squares estimators are chosen to minimize Q by setting the partial derivative of Q with respect to the p least squares estimators equal to zero. The result is a set of p simultaneous equations, called the *normal equations.*

The normal equations for a multivariate linear regression model can be expressed concisely in vector notation. Let $\mathbf{b}=[b_0,b_1,\ldots,b_{p-1}]^T$ be the vector of least squares estimator regression coefficients, and $\mathbf{Y}=[Y_1,Y_2,\ldots,Y_n]^T$ be the vector of n observed responses. Let $\mathbf{X}$ be the $[n\times p]$ matrix of known independent variable values, and define $X_{i0}\equiv 1$, so that the i^{th}row of $\mathbf{X}$ is

$$[1, X_{i1}, X_{i2}, \ldots, X_{ip-1}]$$

The normal equations are then

$$(\mathbf{X}^T\mathbf{X})\mathbf{b}=\mathbf{X}^T\mathbf{Y}$$

and the vector of regression coefficients is

$$\mathbf{b}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Once $\mathbf{b}$ is computed, the estimated regression function can be written as

$$\mathbf{Y}'=\mathbf{X}\mathbf{b}$$

where $\mathbf{Y}'=[Y_1',Y_2',\ldots,Y_n']^T$ is the vector of fitted values. The equation for Y' in terms of $\mathbf{X}$ and $\mathbf{b}$ is a linear model obtained from the experimental data.

3.3.2.2. Testing the Significance of the Regression Coefficients

The test of the significance of the simple correlation coefficient r is made using the t statistic. A similar test of the regression coefficients b_k is made using the F^* statistic. The F^* statistic is distributed as the F ratio, a tabulated probability function with parameters ν_1, the numerator degrees of freedom, ν_2, the denominator degrees of freedom, and α, the level of significance. The F distribution $f(F)$ is asymmetric and of domain $0 \leq F \leq \infty$.

The level of significance α is the probability that the population regression coefficient β_k is equal to zero given the sample estimate regression coefficient b_k. The significance is given by

$$1 - \int_0^{F^*} f(F)dF$$

The tables of F distributions list values of $F(\nu_1,\nu_2)$ for common levels of significance α such as 0.05, 0.01 and 0.001.

To test the regression coefficients, the level of significance α is chosen. The statistic

$$F^* = \frac{SSE(\mathfrak{R}) - SSE(\mathfrak{F})}{df(\mathfrak{R}) - df(\mathfrak{F})} \div \frac{SSE(\mathfrak{F})}{df(\mathfrak{F})}$$

is calculated as follows:

$SSE = \sum_{i=1}^{n} (Y_i' - Y_i)^2$ is the error sum of squares. A separate error sum of squares is calculated for both the full and reduced model.

The parameter $\mathfrak{R}$ denotes the reduced model; i.e., the model excluding those variables whose regression coefficients are to be tested.

The parameter $\mathfrak{F}$ denotes the full model; i.e., the model with all variables included.

$df(\mathfrak{R}) = n - p + q$ is the number of degrees of freedom of the reduced model error sum of squares, where n is the number of observations, p is the number of parameters in the full model and q is the number of variables removed from the reduced model.

$df(\mathfrak{F}) = n - p$ is the number of degrees of freedom of the full model error sum of squares.

The *critical value* of F, $F(\alpha,\nu_1,\nu_2)$, is found from the table of F distributions given the significance level α, numerator degrees of freedom $\nu_1 = df(R) - df(F)$, and denominator degrees of freedom $\nu_2 = df(F)$. If $F^* > F(\alpha,\nu_1,\nu_2)$, then it is more than $(1-\alpha)\times 100\%$ sure that $\beta_k \neq 0$, and the sample estimate b_k is said to be significant at the α level.

When the test is applied to only one regression coefficient β_k at a time, $q=1$ and the test is of the marginal contribution of the variable X_k given that the remaining $p-2$ variables are in the model. The test can also be applied to all of the coefficients at once. In this case, the hypothesis under test is that $\beta_1=0$ and $\beta_2=0$ and ... and $\beta_{p-1}=0$. Thus, $q=p-1$. If $F^* > F(\alpha,\nu_1,\nu_2)$ the conclusion is that there is a regression relation between the variables $X_1,X_2,\ldots,X_{p-1}$ taken together and the dependent variable Y. Both types of test are necessary when there is intercorrelation in the variables, since the marginal contribution of each variable may not be significant, but the contribution of some or all of the variables taken at once (and hence their regression coefficients β_k) may be significant.

3.3.2.3. Forward Entry Regression Procedure

The forward entry regression procedure used by SPSS operates in the following manner [Nie 75]. The user selects three parameters used to control the program. They are n, F, and T where

> n is the maximum number of independent variables to be entered into the equation. The default value is 80.
>
> F is the F-to-Enter value. It is compared to the F ratio computed for a regression coefficient for a variable not yet in the equation. The value computed is the F ratio for that variable if it were brought into the equation on the next step. The F ratio is used in a test of significance for the estimated regression coefficient. The default value is 0.01.
>
> T ($0 \le T \le 1$) is the T-to-Enter value. It is compared to the tolerance of the independent variable being considered for inclusion in the model. The tolerance is the proportion of the variance of that variable that is not explained by the independent variables already in the regression equation. The default value is 0.01.

The procedure also uses the coefficient of multiple determination R^2 as a measure of the amount of variance in the dependent variable explained when there are p parameters in the regression equation. The coefficient of multiple determination is defined as

$$R_p^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

where $SSR = \Sigma(Y_i' - \bar{Y})^2$ is the regression sum of squares, $SSE = \Sigma(Y_i' - Y_i)^2$, is the error sum of squares, and $SSTO = \Sigma(Y_i - \bar{Y})^2$ is the total sum of squares. As p increases, SSE remains the same or decreases. Thus R_p^2 increases as p increases. The larger the increase, the greater the contribution of the last variable added to the explanatory power of the regression equation. [Neeter 74]

The forward entry procedure computes the tolerance T for each variable. The tolerance is

$SSE/SSTO$ for a regression of the $k-1$ independent variables in the equation against the k^{th} independent variable about to be added. The purpose is to check for variables highly correlated with variables already in the equation. If the correlated variables were included, the $\mathbf{X}^T\mathbf{X}$ matrix would become ill-conditioned [Nie 75].

The forward entry procedure begins by computing a simple (one independent variable) regression for each of the p independent variables. The independent variable with the largest R^2 is chosen to be in the model. The F^* statistic

$$F_k^* = \frac{MSR(X_k)}{MSE(X_k)} = \frac{SSTO - SSE(X_k)}{(n-1)-(n-2)} \div \frac{SSE(X_k)}{(n-2)}$$

in which $MSR = \frac{SSR}{p-1}$ is the mean regression sum of squares[2], and $MSE = \frac{SSE}{n-p}$ is the mean error sum of squares, is then calculated for the k^{th} independent variable. If the F^* statistic is greater than the specified F-to-enter, calculation continues. Otherwise the process stops. Next, the tolerance is computed and compared to the T-to-enter value. If the T statistic is greater than the T-to-enter value, the variable is allowed to enter the model. If not, the process stops.

Each possible two-variable regression equation is then found. The F^* statistic

$$F_j^* = \frac{MSR(X_j \mid X_k)}{MSE(X_k, X_j)} = \frac{SSE(X_k) - SSE(X_j, X_k)}{(n-2)-(n-3)} \div \frac{SSE(X_k, X_j)}{(n-3)}$$

is calculated, as is T. If F_j^* and T exceed F-to-enter and T-to-enter, the variable with the largest increment in R^2 is allowed to enter the equation.

The process continues until either

(1) n is exceeded, or

(2) $F^* \leq$ F-to-enter, or

(3) $T \leq$ T-to-enter, or

(4) all of the independent variables are in the equation. [Neeter 74], [Nie 75]

---

[2] In the case where the reduced model has no variables, there is only one parameter, the constant β_0. Then $Y_i' = \bar{Y}$ and $SSE = SSTO$.

3.3.2.4. Results of Linear Regression Analysis for November Data

The data from November 6 were submitted to analysis by the SPSS forward entry regression procedure. The results are summarized in Table 3-3.

Table 3-3: Forward Entry Results

INPUT VARIABLE	R_p^2 CHANGE	BETA	MARGINAL F
1. Drying Air Temperature	0.46035	-0.38651	12.129
2. Paint pH	0.04858	-0.11303	1.229
3. % Excess Oxygen	0.02230	0.06319	0.658
4. Paint Specific Gravity	0.01337	-0.33275	21.795
5. Wash Water Conductivity	0.02274	0.29077	9.587
6. Drying Air Velocity	0.01431	-0.13811	6.649
7. Lehr Oven Gas Flow	0.00479	-0.11486	2.264
8. Dry-Wet Bulb Difference	0.00407	0.12172	3.401
9. Sulfur Dioxide Flow	0.00338	-0.06794	0.911
10. Paint Viscosity	0.00160	-0.05714	0.542
11. Lehr Temperature	0.00118	-0.03782	0.560
12. Water Temperature	0.00009	-0.01243	0.046
Unexplained Variance	0.40324		

OVERALL F: 24.665 SIGNIFICANCE: 0.000

In the table, the column labeled R_p^2 Change represents the marginal decrease in residual variance in optical density. This statistic may be interpreted as the fraction of the variance explained in optical density by each independent variable as it is added to the regression model. Thus, drying air temperature accounts for about 46% of the variance in optical density when it is added to the equation first. Paint pH accounts for an additional 4.9% when it is added to the model and drying air temperature is already in the model, and so on.

The F^* statistic shown is also marginal in the sense that it is a measure of the significance of an individual regression coefficient β_k given that all of the other independent variables are already in the equation. The significance of β_k is tested thus: If $F^* > F(0.01,12,200)=2.18$ then $\beta_k \neq 0$. Thus, if $F^* > 2.18$, we are 99% sure that β_k is not zero, given that all of the other variables are in the model. This test only tells us about the significance of β_k in the particular model for which the F^* statistic was computed. Therefore, we cannot use the marginal F^* statistic to reject more than one variable in Table 3-3. Rather, we would reject one variable, recompute the F^* statistics for the resulting eleven variable model and use them to reject the next variable, until all of the remaining variables in the model are significant. This method of eliminating variables from the model because their regression coefficients are not statistically significant is called *backward elimination.* A backward elimination was carried out in the same SPSS run as the forward entry procedure, and the reduced model shown in Table 3-4 was obtained.

Table 3-4: Reduced Model for Optical Density

VARIABLE	R_p^2 CHANGE	BETA	MARGINAL F
Drying Air Temperature	0.10135	-0.36677	36.071
Paint Specific Gravity	0.18024	-0.33168	39.138
Wash Water Conductivity	0.24120	0.39465	38.435
Drying Air Velocity	0.03337	-0.14096	7.807
Lehr Oven Gas Flow	0.02605	-0.16805	12.909
Unexplained variance	0.41779		

OVERALL F: 57.693 SIGNIFICANCE: 0.000

Tables 3-3 and 3-4 also show the standardized regression coefficients, Beta. These coefficients are the regression coefficients found when the variables have been transformed so as to have a mean of zero and a variance of one. The resulting coefficients are all to the same scale, and so can be compared to assess their relative importance, under the assumption that the corresponding input variables are uncorrelated. The transformation is given by

$$\beta_k' = (\frac{s_k}{s_Y})\beta_k$$

where

s_k is the standard deviation of variable X_k

s_Y is the standard deviation of dependent variable Y

β_k' is the standardized regression coefficient for variable X_k.

With standardized coefficients, the intercept of the regression model is always zero; i.e., $\beta_0' = 0$.

The overall F and associated level of significance in Tables 3-3 and 3-4 is for the regression equation as a whole.

The reduced model may be written with unstandardized coefficients as

$$Y = 205.958 - 0.392X_1 - 102.949X_2 + 0.067X_3 - 0.014X_4 - 0.002X_5$$

where Y is optical density, X_1 is drying air temperature ($^\circ$ C), X_2 is specific gravity (s.g.u.), X_3 is wash conductivity (μS/cm), X_4 is air velocity (ft/min), and X_5 is lehr oven gas flow (ft^3/hr). This equation can be used to predict optical density from the five input variable values. If this were done, the observed value of optical density might deviate considerably from the predicted value, for reasons

discussed in the next section. The principal value of the analysis of the data taken in November is in suggesting which variables should be studied further to try to produce a reliable model.

33.2.5. Interpretation of the SPSS Results

Table 3-4 lists the variables found to be significant in the present study. It is fair to ask whether they can be ranked by the methods discussed in Section 3.3.2.4, and whether it is certain that the other variables arc not significant. The answer to both questions is no. This is because of intcrcorrclations and limited range in the data collected for the study.

Variables may be ranked by the percentage of variance explained, or by the size of their regression coefficients only if the data are free from moderate to strong intcrcorrclations between the independent variables. When the independent variables are correlated, the regression coefficients are influenced by the correlated variables. When the independent variables are uncorrclatcd, each of the regression coefficients is independent of the others. A single multivariatc regression or a series of bivariate regressions would yield identical coefficients. [Younger 79]

The problem of generalizing the regression coefficients to the population also comes up when the independent variables are correlated. According to Cooley and Lohncs [Cooley 71], the regression coefficients found may fluctuate wildly from one sample to the next When we calculate

$$\beta = (X^T X)^{-1} X^T Y,$$

correlated variables make the X^TX matrix ill-conditioned, and the regression coefficients become numerically unstable [Neeter 74].

The ranking of importance of the variables by changes in $R\}\backslash$ that is, the percentage of variance explained as the variable is added to the regression equation is also affected by multicollincarity. In Figure 3-1 the case of two independent variables is illustrated Even though X_l and X_1 both explain a large portion of the variance in F, the increase in explained variance when X_2 is added is small because the two variables arc correlated. If X_x and X_2 are uncorrclatcd, their regions on the diagram would be disjoint, and the incremental increase in R^2 would be the same as the total variance explained by each.

In the correlated case, a term that is added to the equation first gets all the credit, and any subsequent tenm get no credit for variance already explained by preceding terms. Thus we can make JKi important and X_2 unimportant by the R^2 change criterion, or vice-versa, just by changing the order

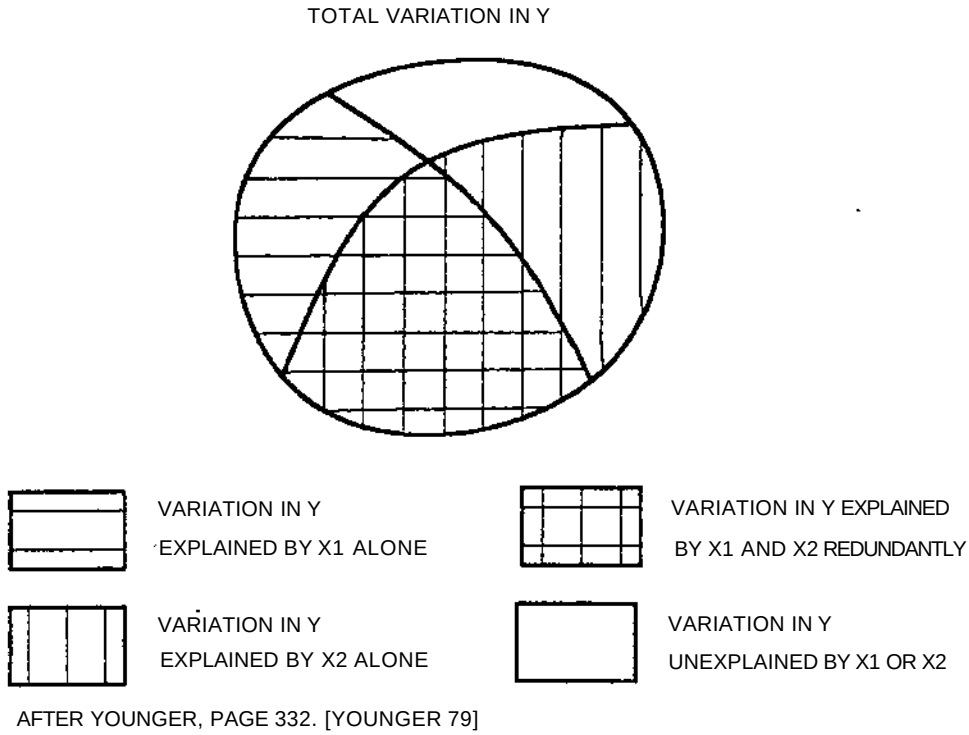


Figure 3-1:
Percentage of variation in Y
explained by X_x and X_2
where X_x and X_x are correlated.

in which they arc added to the regression equation. Tables 3-3 and 3-4 illustrate this phenomenon. The variable drying air temperature accounted for 46% of the variance in optical density in the forward entry results, but only 10% of the variance in the reduced model The difference is simply the order in which the variables were added to the model.

There is significant iniercorrelation in the study data. A simplified correlation matrix, showing only those entries with moderate to strong ($r \geq 0.5$) correlations, is displayed in Table 3-5.

The second problem with the data is the small range of values taken on by some of the variables. The input variables found insignificant in the study may still have a significant effect The problem is that the range of values observed for some variables is so small that the effect might be masked by measurement error.

TTie study results for optical density therefore are *not* conclusive, but rather are suggestive. A controied experiment is needed to produce conclusive results. A design for such an experiment k presented in Chapter 4.

Table 3-5: Correlation Matrix For November 6 Data

VARIABLES

	1	2	3	4	5	6	7	8	9	10	11	12
1	1											
2		1		-0.6	-0.7		-0.5					
3			1								-0.5	
4		-0.6		1			0.6					
5		-0.7			1		0.6					0.5
6						1						
7		-0.5		0.6	0.6		1					0.5
8								1		-0.8		
9									1			
10								-0.8		1		
11			-0.5								1	
12					0.5		0.5					1

KEY:

1 Wash Water Temperature
2 Wash Water Conductivity
3 Paint Specific Gravity
4 Paint Viscosity
5 Paint pH
6 Drying Air Velocity
7 Drying Air Temperature
8 Sulfur Dioxide Flow Rate
9 Percent Excess Oxygen
10 Lehr Temperature
11 Lehr Gas Flow Rate
12 Drying Air Humidity

3.4. Visual Defects as a Function of the Process Inputs

In analyzing the effect of the process input variables on the occurrence of visual defects, we would like to identify the variables that are good predictors of visual defects, and to be able to predict the occurrence of visual defects from the input variable values. The SPSS discriminant analysis procedure is suitable for analysis of data where the independent variables are continuous variables and the dependent variable is a discrete, nominal level variable.

3.4.1. Discriminant Analysis

In discriminant analysis [Tatsuoka 71], we undertake to write a linear combination of independent variables that shows large differences in the means of observations associated with the dependent variable categories. The linear combination is called a *discriminant function.*

3.4.1.1. Obtaining the Discriminant Function

A discriminant function is of the form

$$Y = \mathbf{v}^T \mathbf{X}$$

where Y is the discriminant score, $\mathbf{v}=[v_1, v_2, \ldots, v_p]^T$ is the vector of weighting coefficients, and $\mathbf{X}=[X_1, X_2, \ldots, X_p]^T$ is the vector of p independent variables.

The object is to find the weighting coefficients $\mathbf{v}$ such that the category, or group, means are separated as far as possible on the dimension Y. The criterion for measuring the difference among several group means against Y is

$$F^* = \frac{SSB/(K-1)}{SSW/(N-K)}$$

where

K is the number of groups,

N is the number of observations,

SSB is the between-the-groups sum of squares, and

SSW is the within-the-groups sum of squares.

The expression for the criterion is rewritten in terms of v as follows. We drop, for convenience, the constant (.V— A") and {*K*— 1) factors. The bctwccn-thc-groups sum of squares *SSB(* K) = v^{7}Bv where B is the bctwccn-thc-groups sum of squares and cross products *(SSCP)* matrix. The elements *by* of B are given by

$$b_{ii} = \sum_{k=1}^{K} n_k(\bar{X}_{ik} - \bar{X}_i)^2 \qquad \text{for } i = i$$

and by

$$b_{ij} = \sum_{k=1}^{K} n_k(\bar{X}_{ik} - \bar{X}_i)(\bar{X}_{jk} - \bar{X}_j) \qquad \text{for } i \neq j,$$

where

$\bar{X}_{ik}$ is the mean value of the i^{th} independent variable in the k^{th} group,

$\bar{X}_i$ is the mean value of the I^{th} variable over all the groups,

n_k is the number of observations in the k^{th} group, and

A'is the number of groups.

Similarly, the within-the-groups sum of squares $SSW(Y) = \mathbf{v}^T\mathbf{W}\mathbf{v}$, where

$$\mathbf{W} = \sum_{k=1}^{K} \mathbf{S}_k$$

is the within-the-groups *SSCP* matrix. The elements $S_{ka}p$ of S* are given by

$$S_{k\alpha\beta} = \left[\sum_{i=1}^{m_k} X_{\alpha ki} X_{\beta ki}\right] - \frac{1}{J\ 'k}\left[\sum_{i=1}^{nir} X_{\alpha ki}\right]\left[\sum_{i=1}^{ftp} X_{\beta ki}\right]$$

The *discriminant criterion* is then written as

$$\frac{SSB(Y)}{SSW(Y)} - \frac{\mathbf{v}^T\mathbf{B}\mathbf{v}}{\mathbf{v}^T\mathbf{W}\mathbf{v}} = \lambda$$

Maximizing the discriminant criterion *X* with respect to v yields the eigenvector-eigenvalue problem

$$(\mathbf{W}^{-1}\mathbf{B} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}$$

in which B is of rank r, where *r=min(K—Lp).* (W"^{1} always can be found if no variable is perfectly correlated with the others [Tatsuoka 71], and no variable has zero variance within the groups,) The solution of the maximization problem yields *r* non-zero eigenvalues *X^* The corresponding

eigenvectors, normalized to modulus one, are the weighting coefficients vectors v. Thus by maximizing the discriminant criterion λ, a set of r discriminant functions, $Y_m = v_m^T X$ for $m = 1,2,\ldots,r$ is obtained.

3.4.1.2. Ranking and Testing the Discriminant Functions

The discriminant criterion λ is proportional to the distance between group means on the associated discriminant dimension. The m^{th} discriminant function can therefore be ranked in discriminating power according to the magnitude of the corresponding λ_m.

The significance of the discriminant functions can be tested using Wilks' Λ. We compute the test statistic

$$\Lambda^* = \frac{|W|}{|T|}$$

where T is the total *SSCP* matrix.[3] Since $\frac{1}{\Lambda^*} = (1+\lambda_1)(1+\lambda_2)\ldots(1+\lambda_r)$, Λ^* can be computed from the eigenvalues of $W^{-1}B$.

Tables of Λ distributions are not used to test the significance of the discriminating functions; rather, we use Λ^* to compute *Bartletts' V* according to

$$V = -\left[N-1-\frac{(p+K)}{2} \right] \ln \Lambda^*.$$

Bartletts' V is approximately χ^2 distributed with $p(K-1)$ degrees of freedom. Thus, we can test the significance of all the discriminating functions taken together by using tables of χ^2 distributions. [Tatsuoka 71]

Each individual discriminant function can be tested by computing

$$V_m = \left[N-1-\frac{(p+K)}{2} \right] \ln(1+\lambda_m)$$

which is distributed with degrees of freedom $(p+K-2m)$.

---

[3] The elements $T_{\alpha\beta}$ of the T matrix are given by

$$T_{\alpha\beta} = \left[\sum_{k=1}^{K} \sum_{i=1}^{n_k} X_{\alpha ki} X_{\beta ki} \right] - \frac{1}{N} \left[\sum_{k=1}^{K} \sum_{i=1}^{n_k} X_{\alpha ki} \right] \left[\sum_{k=1}^{K} \sum_{i=1}^{n_k} X_{\beta ki} \right]$$

The residual discrimination after the first discriminant function is accepted is $V-V_1$ with $(p-1)(K-2)$ degrees of freedom. If this statistic meets the confidence level, we accept the second discriminant function and compute the residual discrimination $V-V_1-V_2$ with $(p-2)(K-3)$ degrees of freedom, then $V-V_1-V_2-V_3$ with $(p-3)(K-4)$ degrees of freedom, etc., until the residual is smaller than the confidence limit percentile point desired.

3.4.1.3. Ranking the Independent Variables in the Discriminant Function

The relative contribution of the i^{th} independent variable to the m^{th} discriminant function can be judged by comparing the magnitude of its standardized weighting coefficient v_{mi} to those of the other variables in the function. The weighting coefficients v_{mi} are standardized by multiplying them by the square root of the corresponding diagonal element of the within-the-groups sum of squares and cross products matrix **W**:

$$v'_{mi}=\sqrt{W_{ii}}\ v_{mi}$$

The larger coefficients indicate the more important variables. This provides a basis for deciding which variables merit further study.

3.4.2. The SPSS DISCRIMINANT Subprogram

SPSS calculates the discriminant functions by a stepwise procedure. As in linear regression, we can find the marginal contribution to the model as each variable is entered. The stepwise procedure begins by computing the partial multivariate F ratio for each variable. This F^* statistic is used to test the significance of the amount of separation of the group centroids added by the variable being considered. The F^* statistic is compared to the parameter FIN, which has a default value of one. If $F^* > FIN$, and the tolerance computed exceeds the minimum,the variable is eligible for entry into the model. Next, each eligible variable is added to the model, and the entry criterion computed. The variable with the best entry criterion score is retained in the model. Finally, each of the variables in the model is tested with the partial multivariate F again. If $F^* < FOUT$, that variable is removed. It may reenter on a succeeding step if $F^* > FIN$. The procedure repeats until no more variables are eligible or the entry criterion scores are below some minimum.

Several entry criterion are available. This analysis uses the criterion which minimizes the residual variation in a set of dummy variables representing group pairs. This is a way of choosing the next variable so as to maximize the distance between groups. The residual variation is estimated by [Nie 75]

$$R = \sum_{i=1}^{K} \sum_{j=1, j \neq i}^{K} \frac{1}{1+(D_{ij}/4)}$$

where

$$D_{ij} = [(\mu_i - \mu_j)^T \Sigma^{-1} (\mu_i - \mu_j)]^{\frac{1}{2}}$$

is the Mahalanobis distance between groups i and j, μ_i and μ_j are the centroids of groups i and j, and Σ^{-1} is the inverse of the covariance matrix [Duda 73]. (SPSS computes the Mahalanobis distance using sample, rather than population parameters. [Norusis 79])

By observing the change in the residual variation R as the stepwise procedure adds new variables to the model, we can see the marginal contribution of the most recently added variable to the discriminating power of the model, given the other variables that are already in the model.

3.4.3. Discriminant Analysis of the Data

There are eight categories of visual defects, plus the category of good tubes. The frequency of occurrence of observations in each of the nine groups for the data taken on November 6th, 1981 is shown in Table 3-6.

Table 3-6: Frequency of Occurrence of Visual Defects

CATEGORY	NUMBER OF CASES
1. Bubbles	6
2. Streaks	7
3. Short Coat	0
4. Texture	8
5. Hanger Marks	3
6. Partial Coat	54
7. Density	17
8. Thin End	0
9. Good Tubes	225
Total	320

Because optical density is analyzed separately, the density category is dropped.

3.4.3.1. Dichotomized Visual Defects

The results of an SPSS discriminant analysis with the visual defects lumped into two groups, "good" and "bad", is shown in Table 3-7. For this analysis, the stepwise procedure was followed with *FIN* and *FOUT* set equal to one, and the residual variation minimizing criterion was used. Table 3-7 shows the variables listed in order of their discriminating power as measured by % R Change. The standardized coefficients show a similar ranking, with the exception that Lehr gas flow should be first.

The model derived is an eight variable model. The standardized coefficients form the weighting vector v' and the variables form the vector X in the model $Y = v'^{T}X$. The model is statistically significant to the 0.0001 level. The model is not as small as possible since the tolerance, *FIN* and *FOUT* were all set to let almost any variable enter the model. The physical significance of the model is open to question since, of the twelve variables considered for inclusion, only four (wash water conductivity, Lehr oven gas flow, paint viscosity, and drying air velocity) have differences in group means of more than one percent.

Table 3-7: Dichotomized Visual Defects

VARIABLE	% R CHANGE	STANDARD COEFFICIENTS
1. Sulfur Dioxide Flow	37.631	−0.64460
2. Drying Air Velocity	16.192	−0.64716
3. Wash Water Conductivity	6.178	+0.58135
4. Lehr Gas Flow	4.427	+0.72526
5. % Excess Oxygen	2.573	+0.44860
6. Paint Specific Gravity	0.235	−0.27027
7. Paint pH	0.515	+0.25234
8. Paint Viscosity	0.147	−0.11126
9. Unexplained Variance	32.101	

Wash water temperature, Lehr oven temperature, drying air temperature, and dry-bulb wet-bulb temperature difference (humidity) could not meet the *FIN* criterion, and were not entered into the model.

3.4.3.2. Visual Defects by Groups

The results of a SPSS discriminant analysis with all six non-void categories of visual defects is shown in Table 3-8.

In this case there are six groups, so R is $\binom{6}{2}$ or 15, if all of the variance is unexplained. The percentage change in R is then $[(R_i - R_{i+1}) / 15] \times 100\%$.

Table 3-8: Analysis Results With Six Groups

VARIABLE	% R CHANGE	FIRST FUNCTION STANDARD COEFFICIENTS
1. Sulfur Dioxide Flow	27.56	−0.73989
2. Paint Viscosity	14.02	−0.16071
3. Lehr Gas Flow	8.05	+0.92121
4. % Excess Oxygen	7.16	+0.60744
5. Wash Water Conductivity	4.85	+0.50859
6. Drying Air Velocity	3.91	−0.54192
7. Paint Specific Gravity	3.36	−0.25577
8. Lehr Temperature	1.69	+0.00549
9. Drying Air Temperature	1.37	−0.14453
10. Dry-Wet Temp. Difference	1.75	+0.03468
11. Wash Water Temperature	2.12	−0.03884
12. Paint pH	0.59	+0.44010
13. Unexplained Variance	27.57	

The first discriminant function accounts for 80% of the between-groups variance because the eigenvalue associated with the first function is considerably larger than the others.[4] Ranking by the size of the standardized coefficients is roughly the same as by %R change. Viscosity is an exception, but it makes up for lost ground on the remaining four functions. Of the five discriminant functions derived, four are significant to the 0.01 level and are accepted.

3.4.3.3. Interpretation of Discriminant Analysis Results

The ranking of the variables is nearly the same if the visual defects are all lumped into one category or considered separately. Paint viscosity emerges as an important variable in the latter case. Regardless of whether the cases are distributed along a single dimension or the four dimensions found significant in the second analysis, the ultimate goal is to distinguish the operating conditions which result in good bulbs. The choice of functions to do this should be based on a test of their predictive validity. To test the validity of the discriminant functions derived, more data are required. The test consists of using classification functions to classify new data with known group membership. The percentage of correct classifications is a measure of the validity of the classification functions.

---

[4] The percentage of between groups variance is calculated by $(100\lambda_k) / \sum_{k=1}^{r} \lambda_k$ [Norusis 79]

3.5. Summary

A linear regression model for the effect of the input variables on optical density has been found to contain five variables: drying air temperature, paint specific gravity, wash water conductivity, drying air velocity, and Lehr oven gas flow. The model accounts for 58% of the variation in optical density. Because of intercorrelations and limited range in the data, the model derived is preliminary and serves as a basis for the design of a controlled experiment

Discriminant analysis of the effect of the input variables on the occurrence of visual defects yields a model consisting of four discriminant functions, each containing all twelve input variables. The relative importance of input variables, as determined by the percentage of between-groups variance explained, is used in Chapter 4 to choose the variables in a controlled experiment.

Chapter 4
Design of a Controlled Experiment

4.1. Hypothetical Model of the Coating Process

The results of the study are used to propose a hypothetical model for the coating line. The model is used in this chapter to design the controlled experiments. The study involves 12 independent (input) variables. The complexity of an experiment increases at least linearly with the number of variables; for some designs it increases exponentially. From the practical point of view, it is desirable to design the experiment with as few independent variables as possible.

4.1.1. Optical Density

Selecting variables for a reduced optical density model must involve some judgment. The high intercorrclation and limited range of the variables in die study means there is no clear-cut criterion for accepting or rejecting a variable. Of the five variables in the model displayed in Table 3-4, one is a measure of the paint, and two more are measures of the drying process. This finding supports the view that the variables which characterize the paint and drying conditions determine optical density.

The remaining two variables are conductivity and Lehr oven gas flow. Lehr gas flow is related to the effect of the oven on optical density. In fact, gas flow depends upon oven dynamics. When the line coating stops, the oven empties out and the cooling effect of the flow of tubes into the oven is lost The oven temperature controller reduces the gas flow rate, affecting the percentage of excess oxygen. When the line starts again, the oven fills with cool tubes. The gas flow increases and percentage of excess oxygen is changed Thus, changes in gas flow and percentage of excess oxygen are correlated with line stops. When the line stops, the coating thickness is affected because the air velocity profile under the drying hood is altered. Thus there is a correlation between percentage of excess oxygen, gas flow and optical density. By controlling for the line stops, the correlation can be eliminated, Control could be obtained by excluding from the analysis any data taken for 21 minutes (the time it takes one tube to traverse the line) after a line stop.

The optical density experimental hypothesis is then:

Optical Density is determined by the paint variables (pH, viscosity, specific gravity), the drying hood variables (velocity, temperature and humidity), and wash water conductivity.

4.1.2. Visual Defects

From Table 3-8, Lehr gas flow and percent excess oxygen are among the first six variables ranked in terms of the percentage of variance for visual defects explained by each of the independent variables. If these can be removed by controlling for the line stops, the only remaining variable not taken as an independent variable in the optical density experiment is sulfur dioxide flow rate. By adding this variable, and removing one of the other variables used in the optical density experiment, a visual defects experiment can be run concurrently with the optical density experiment with no increase in experimental complexity. As a practical matter, all six of the optical density experimental variables may as well be left in the visual defects experiment, since sulfur dioxide flow is very easily controlled, and the savings in effort obtained by confounding it with one of the other variables is small.

The visual defects experimental hypothesis is:

The occurrence of visual defects is a function of sulfur dioxide flow rate, paint viscosity, wash water conductivity, drying air velocity, and paint specific gravity.

4.2. Design of the Experiment

In Section 4.1 the experimental hypothesis and the independent variables for the experiment are stated. In this section we will decide which variables are to be controlled, the number and values of the treatment levels, and the number of observations which are required.

4.2.1. Classifying the Independent Variables

The independent (input) variables are assigned to the following categories:

1. Constants: These variables are maintained at a fixed value during the experiment. Any effect they may have on the dependent variable is excluded from the experiment.

2. Unmeasured Variables: These variables are not measured or controlled during the course of the experiment. Their effects, if any, on the dependent variable are lumped into the error of the experiment.

3. Measured Variables: These variables are measured and may or may not be controlled during the experiment. We seek to determine the effect of these variables on the dependent variable.

Because experimental complexity is related to the number of controlled variables, it is desirable to reduce the number of such variables. Therefore, as many variables as possible should be assigned to categories 1 and 2. In Section 4.1.1 it was suggested that percent excess oxygen and Lehr oven gas flow could be held constant by controlling for the line stops. Thus, these two variables are placed in category 1. Charles Trushell, a Westinghouse engineer, has suggested a means of holding wash water conductivity constant as well. De-ionized make-up water would be added to the wash tank continuously by means of a set of nozzles at the end of the wash. The tubes would then be subjected to a rinse of pure de-ionized water as they left the wash, thus fixing wash water conductivity at zero. On this basis the variables are classified as shown in Table

Table 4-1: Classes of Independent Variables

FIXED	UNMEASURED	MEASURED
Line Stops (Percent Excess Oxygen, Lehr Gas Flow Rate)	Wash Water Temperature	Paint pH
Wash Water Conductivity	Lehr Temperature	Paint Specific Gravity
Mill Factor		Paint Viscosity
		Drying Air Temperature
		Drying Air Velocity
		Drying Air Humidity
		Sulfur Dioxide Flow Rate

4.2.2. Measured Variables

In a controlled experiment, the measured variables must be held at certain values (called "treatment levels"). The minimum number of levels is two, which yields a first, but not higher, order relation. [Bartee 68] (A first order model is one which contains no powers of the variables.) The values of the treatment levels are set at the limits of normal operation so that the results will apply over that range of values. C. Moore of Westinghouse supplied the limits of normal operation for Cool-White paint on Line 1 at Fairmont which were used to set the levels tabulated in Table 4-2.

Table 4-2: Treatment Levels

VARIABLE	% CHANGE FROM LOW TO HIGH LEVELS	UNITS
1. Paint Specific Gravity	3.0	sgu(gm/cm^3)
2. Paint Viscosity	60	cP
3. Paint pH	7.1	pH
4. Drying Air Temperature	6.3	^oF
5. Drying Air Velocity	20.0	% of nominal
6. Drying Air Humidity	107.7	R.H.
7. Sulfur Dioxide Flow Rate	18.2	ft^3/hr

4.2.3. Experiment Design

Each of the seven variables is to be set at two treatment levels. The next choice is whether to implement a full factorial experiment or an incomplete design. The merits of each are discused in this section.

4.2.3.1. Full Factorial Design

The full factorial design requires measurement of all combinations of high and low values for all of the variables. To run two concurrent experiments with seven variables in each would require $2^7 = 128$ observations. When the full factorial design is used, multicollinearity is eliminated. It is guaranteed that no pair of the independent variables is correlated. With linear regression analysis, this yields three advantages:

1. The regression coefficients are meaningful indicators of the degree of influence of their associated variables on the dependent variable.

2. The regression equation may be reliably generalized to the population.

3. The data have come from the full range of all possible operating conditions, and we can thus be confident of the resulting regression equation's predictive validity under all normal conditions.

The main disadvantage of the full factorial design is that the required number of observations increases exponentially with the number of independent variables. To model the coating process, 128 observations would be required. An experiment of this magnitude would be quite costly and time-consuming.

4.2.3.2. Minimal Incomplete Design

A design in which the number of observations is less than k^{p-1}, where k is the number of levels and $p-1$ is the number of independent variables, is called *incomplete* [Bartee 68]. For a regression analysis, the minimum number of observations is equal to p. This is because the observation matrix X (defined in Section 3.3.2.1) must have as many observations as parameters or X^TX is singular and we cannot solve for the regression coefficient vector β. We need $p=8$ observations, since p points define a hyperplane in $(p-1)$-space. This minimal incomplete design has the great advantage that it is much simpler to run than the full design. The design is illustrated in Table 4-3. It can be seen that there is a slight tendency of one variable to be negatively correlated with the others.[5]

Table 4-3: Eight Observation Experimental Design

	VARIABLE				1 = HIGH		0 = LOW
OBSERVATION	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0
3	0	1	0	0	0	0	0
4	0	0	1	0	0	0	0
5	0	0	0	1	0	0	0
6	0	0	0	0	1	0	0
7	0	0	0	0	0	1	0
8	0	0	0	0	0	0	1

[5] The correlation of one variable to another may be tested by Pearson's r, as given in Section 3.3.1.

As stated in Section 4.2.2, the experiment will consist of observations taken at two levels chosen to cover the entire range of normal operation. A two level experiment yields only a first order model (i.e. a model without power terms). A two level experiment with two independent variables is illustrated in Figure 4-1.

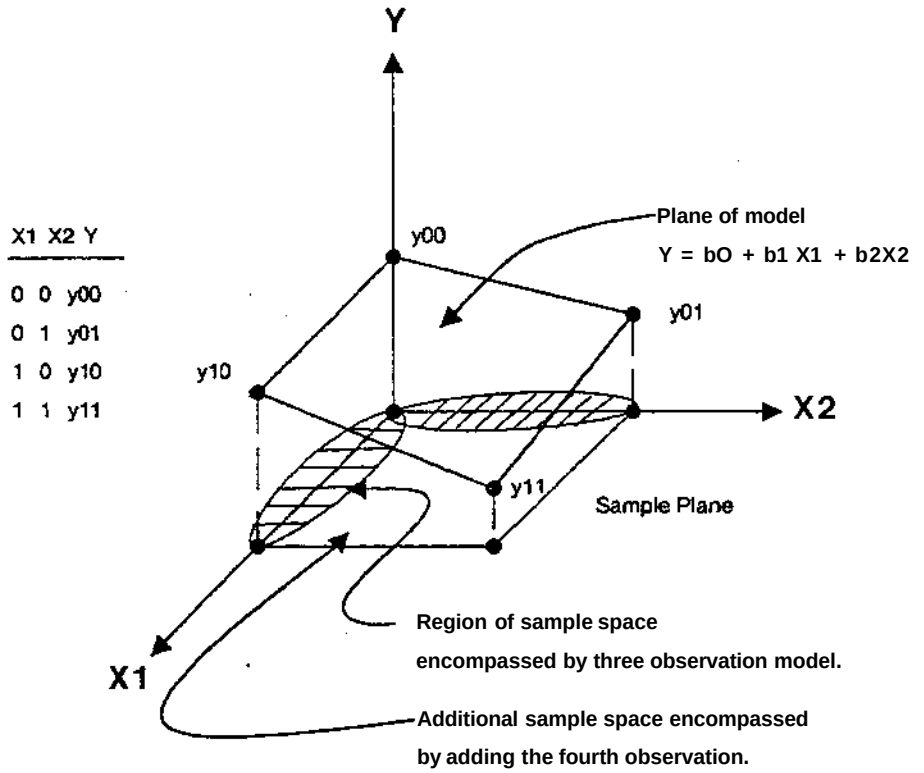


Figure 4-1: A Two Level Experiment

In a full factorial design, the two variable, two level experiment requires four observations. These observations are marked in Figure 4-1. A minimal observation design only requires three observations, but the area of the "sample plane" enclosed is greatly decreased. The resulting model can only be used to predict responses within this smaller range of variable value combinations. The model Is planar; the power or interaction terms of the process are averaged over the range of observations made. There is the danger that the model will be in error in regions of the sample space not investigated.

The "sample plane" In the two variable, two level experiment consists of an area enclosed by four points, the four observations of the ftill factorial experiment The fourth point greatly increases the area enclosed. This point is the observation of the "Interaction effect", and the area enclosed contains information about what happens when both ?ariablc$ simultaneously move through their range The number of interaction effects increases as the number of independent variables is increased. A two

variable system has one two-variable interaction effect, a three variable system has three two-variable effects and one three-variable effect, and in general, a n variable system has $\binom{n}{i}$ i-variable effects. A seven independent variable experiment has 21 two-variable terms, 35 three-variable terms, 35 four-variable terms, 21 five-variable terms, seven six-variable terms, and one seven-variable term. As the order of the interaction effect is increased, it usually diminishes in significance.

The more interaction terms that are to be investigated, the more observations that must be made up to the k^{p-1} observations of the full factorial design, which allows all interactions to be investigated. The need for more observations arises because without them, the interaction terms are intercorrelated and cannot be distinguished from one-another. In linear regression analysis these terms are added to the regression equation. The regression equation for a two variable system becomes

$$Y' = b_0 + b_1 X_1 + b_2 X_2 + b_{12} X_1 X_2.$$

Interaction terms are always correlated with the other variables used to construct them. This fact should be kept in mind when interpreting models which contain interaction terms.

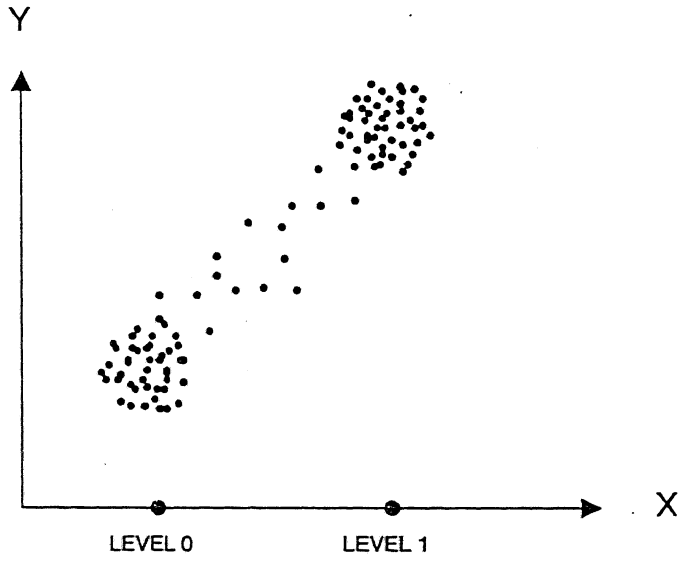
Investigation of power terms requires more than an increased number of observations. The number of levels must be increased to at least three if these non-linear terms are to contribute to the model. The fact is that two data points cannot define a curvilinear line! With only two levels, power terms contribute to the model if we make the necessary observations (remember that the number of observations must be at least as great as the number of parameters in the model), but this contribution cannot be resolved into separate terms.

Throughout the discussion it has been implied that only one observation is made for each variable at each combination of levels. In practice we make several such observations of both dependent and independent variables during the time the particular combination of levels is in effect. We may expect to see some variation in the response recorded due to measurement errors and poor control of the level. The resulting cloud of data points around each level is illustrated in Figure 4-2. As the cloud spreads out, it becomes possible to detect power terms in the regression analysis. In fact, the data are more likely distributed over the range of the variables, since the data are recorded as the variable is moved from one level to the other in the course of the experiment. Use of such cases might allow investigation of power terms. At least, the observations which fall near the treatment levels could be used to construct a first-order model and then intermediate observations used to test its predictive validity.[6]

---

[6]The author is indebted to Prof. Luc Tierney of the Carnegie-Mellon Statistics Department for suggesting this method.

Figure 4-2: Responses in a Two Level Experiment



4.2.3.3. Fractional Factorial Design

A fractional factorial design is a reduced-order design which allows for investigation of some, but not all, of the interaction effects. It also assures that all single-term variables are uncorrelated. A fractional design using 32 observations is shown in Table 4-4. [Cochran 57]

Table 4-4: 32 Observation Design

OBSERVATION	VARIABLE $X_1 X_2 X_3 X_4$	$X_5 X_6 X_7$	OBSERVATION	VARIABLE $X_1 X_2 X_3 X_4$	$X_5 X_6 X_7$
1.	1 1 1 1	0 0 1	17.	0 0 1 1	0 0 1
2.	1 1 1 1	0 1 0	18.	0 0 1 1	0 1 0
3.	1 1 1 0	1 1 0	19.	0 0 1 0	1 1 0
4.	1 1 1 0	1 0 1	20.	0 0 1 0	1 0 1
5.	1 1 0 0	0 0 0	21.	0 1 1 0	0 0 0
6.	1 1 0 0	0 1 1	22.	0 1 1 0	0 1 1
7.	1 1 0 1	1 1 1	23.	0 1 1 1	1 1 1
8.	1 1 0 1	1 0 0	24.	0 1 1 1	1 0 0
9.	1 0 0 1	0 0 1	25.	0 1 0 1	0 0 1
10.	1 0 0 1	0 1 0	26.	0 1 0 1	0 1 0
11.	1 0 0 0	1 1 0	27.	0 1 0 0	1 1 0
12.	1 0 0 0	1 0 1	28.	0 1 0 0	1 0 1
13.	1 0 1 0	0 0 0	29.	0 0 0 0	0 0 0
14.	1 0 1 0	0 1 1	30.	0 0 0 0	0 1 1
15.	1 0 1 1	1 1 1	31.	0 0 0 1	1 1 1
16.	1 0 1 1	1 0 0	32.	0 0 0 1	1 0 0

NUMBER OF CHANGES 1 2 4 8 15 16 23

In this design all of the two variable interactions can be distinguished, with the following exceptions:

1. The interaction $X_4 X_5$ is indistinguishable from the interaction $X_6 X_7$.

2. The interaction $X_4 X_7$ is indistinguishable from the interaction $X_5 X_6$.

3. The interaction $X_4 X_6$ is indistinguishable from the interaction $X_5 X_7$.

These pairs of interaction terms are called *alias pairs.* They are indistinguishable because the observations required to distinguish them have been ommitted from the fractional design and they are therefore intercorrelated. To minimize the possibility that an important effect might be confounded with its alias, a variable, e.g., X_7, is chosen to be that variable which is expected to have the least effect on the experiment. If an alias pair is important in the analysis, it is assumed that the interaction term causing the effect is the one that does not include X_7. Higher order interactions are confounded with the error in this design. Since the analysis of interaction terms is intended for the optical density experiment, it is suggested that the sulfur dioxide flow rate be chosen as X_7, since it is expected that sulfur dioxide flow rate will not affect optical density. The remaining six variables can be assigned according to the effort required to control them, as discussed in Section 4.3.2.

4.3. Carrying out the Experiment

4.3.1. Hardware

To carry out the experiment, an automatic data acquisition system is essential. This is because the data must be gathered on the production line during normal operation. Data gathered manually are expensive, subject to error, and may be recorded too slowly to observe significant changes in the variables. Since the sampling rate cannot be adequately controlled with manual data collection, it becomes necessary to use values interpolated over time. The statistical analysis is then based on calculated, rather than measured data, which is less desirable.

Automatic data acquisition is fast and precise. It allows experiments to be carried out rapidly and unobtrusively with minimal manpower. Sampling times are easily controlled. Once installed, such a system would facilitate both the present and future studies and could serve in a future control and information management system. To design the automatic data acquisition system, the range and precision of each measurement, and the sampling time required must be specified. Once this is done, the means of transmitting and recording the data can be considered.

4.1I.I. Sensors

The range of each sensor must be large enough to accommodate the maximum range of values expected In some cases the range of an instrument affects its precision; in those cases the range is made as narrow as possible consistent with the expected range. The required ranges for the measured variables (which were supplied by C Moore of the Westinghousc Fairmont Works) are given in Table 4-1 These ranges arc for normal operation with cool-white paint on Line 1 at Fairmont.

The desired response time of the sensor is on the order of one second. The instrument should thus register to within 95% of the new value within one second after a change in the variable value. The figure of one second is chosen because the process proceeds with a period of less than one second per robe, and it is unnecessary to resolve changes that happen in less time than it takes to fill a tube.

While, for statistical analysis, absolute accuracy is not as important as precision the instruments should be as accurate as practicable.

Suggested sensors arc shown in Table 4-5. In some cases more than one suggestion is made. The table shows the model manufacturer, and cost of the instrument, along with die range, accuracy, and precision. The dewed precision is based on the results of the study described in Chapter 3; in cases *Jicre this is *not* practical the precision specified is that commonly obtained with available instruiRcnts. The measuring instrument must be precise enough to resolve variation within the expected range*

In *addinm* to the set'en input Yarkble sensors shown in Table 4-5, sensors for the two output variables, optical density and visual defects, will be required. A sensor utilizing a linear Charge-Couplcd Device (CCD) anay has been *the* subjcxt of research done by Mark Handclsman at CMU |Hande!«ran 82), The sensor is said 10 be capable of detecting coating thickness and visual delects *ilh a throughput of 6C0G cubes per htur* The cost of a practical instrument is estimated to be about $20*000* John Murr*y has denwasiraiucl a simple coaling thickness sensor using a phototransistor *hkh cotjfii be feyílf cheaply [Murray I2J. The experiment could be run without output sensors by sampling We aibc* as !he> emerge *dim* the coating line and storing them for later evaluation, This would *be* a ttd;ous *ptmm* however* and would requ're temporary storage of several hundred tubes. The use tff ityStittaiic $C£&>rs 3 tc be preferred*

Table 4-5: Suggested Sensors for Automatic Data Acquisition

Variable	Instrument	Desired Range	Precision	Accuracy	Response	Cost	Maker
Dry Temp.	Thermo-couple	130-150 °F	About 0.2 °F	About 2 °F	0.12 Second	$41	Omega
Humidity	Chilled-Mirror Hygrometer	to 0.0 °C @ 60 °C ambient	0.1 °C	1.0 °C	2.0 °C per Second	About $3000	General Eastern
Paint Density	Vibration type	1.3 to 1.4 s.g.u.	0.001 s.g.u.	Not Specified	Not Specified	About $3500	Yokogawa
	Vibration type	same	0.0001 s.g.u.	0.0005 s.g.u.	Milli-seconds	$8690	Redland
	Gamma Radiation type	same	0.0005 s.g.u.	Not Specified	Not Specified	About $3500	Texas Nuclear
Paint pH	Model 7076-3 pH Meter	7.5 to 9.0 pH	0.01 pH	1 % Full Scale	1 Second	Not Applicable	Leeds & Northrup
Paint Viscosity	Model VTA-100 Viscometer	60 to 95 cP	Not Available	3%	10 Seconds	$450	Brookfield
Air Velocity	Model IM-4 Hot Tip type	600 to 1100 fpm	Not Specified	plus or minus 2%	1 Second	$2800	Anemostat

Sulfur Dioxide Flow Rate: Use of existing ball and tube type instrument recommended.

4.3.1.2. Data Logging

Because the process is to be sampled at a rate of about one observation per second, and there are seven independent variables and two dependent variables to be measured automatically, the data logger must be capable of making nine readings per second. This requirement is well within the capabilities of most data loggers.

With the exception of the thermocouple and the pH meter, the output of the sensors is a 4 to 20 ma current source. A current source permits the use of shielded, twisted-pair cable for the connection between the instrument and the data logger. Such cable is superior in an electrically noisy environment. The current source covers the range of the instrument. The data logger must be equipped with preamplifiers compatible with the sensor output. The preamplifier should have adjustable gain and zero settings to allow maximum resolution.

Table 4-5 indicates that no instrument has a precision which is greater than one one-thousandth of

its range. Thus, a 10 bit D/A converter should be capable of resolving any measurement to its specified precision.

The data logger should also have a clock to keep track of the time each reading was made, to the nearest second. It also should be capable of recording the readings permanently on a medium such as a floppy disk. Because there can be an independent variable reading for every dependent variable reading, there is no need for any interpolation.

4.3.2. Scheduling the Experiment

The goal in formulating a schedule is to reduce the effort required to carry out the experiment. The variables should be ranked in order from the most difficult to the least difficult to change from the high to the low level. The most difficult variable is assigned to be the one changed the fewest number of times and so on. In doing this, the opportunity to minimize bias error by random selection of the order of observation is lost. In this case the difficulty of controlling the independent variables justifies this choice.

The ranking is perhaps best left to those who will carry out the experiment. A tentative ranking follows:

1. Paint pH
2. Paint specific gravity
3. Paint viscosity
4. Drying air humidity
5. Drying air velocity
6. Drying air temperature
7. Sulfur dioxide flow rate

The first three variables are the paint variables. Changing them may require the line to be stopped while new paint is added to the system. The next four probably can be changed "on the fly". Control of drying air humidity might be accomplished by spraying atomized water into the drying air duct.[7] Air velocity could be controlled by venting the duct, reducing the pressure difference and thus the velocity of discharge from the duct. Air temperature is easily controlled by a thermostat; sulfur dioxide flow rate is set by a valve.

---

[7]The measurement of humidity for the study was made using the wet-bulb dry-bulb method which gives poor results in hot, dry conditions. It is not clear why the air does not become saturated, since it is continuously recirculated. Further study of humidity under the drying hood with more precise instruments should be made to determine its variability. Insight gained from this study might suggest a means of controlling humidity.

The 32 observation experiment of Table 4-4 is arranged to show its hierarchical structure. Note that the first variable only changes once in the course of the experiment, the second changes three times, etc. The variables should be assigned in order of decreasing difficulty of control.

4.4. Analysis of the Results of the Controlled Experiment

4.4.1. The Optical Density Experiment

Linear regression models arc built through an iterative process. An excellent discussion of model building is given in Chapter 11 of Necter and Wasscrman [Nccter 74]. The analyst first would try to isolate a reduced set of variables through a stepwisc procedure. Tests for lack of fit (the F ratio tests of Chapter 3) can be made, and residuals examined for lack of fit, outliers, and time dependence in the data. Then the process may be repeated using interaction and power terms. The analyst may try a number of combinations of variables before the best model for the data is found. The model must also be tested for predictive validity. Data taken during the course of the experiment and set aside for that purpose can be used. New data should be obtained at some other time to investigate whether the model remains valid over time. The percentage of variance explained by the model can be used to judge whether there arc other important factors not included in the **experiment.**

Once the model is constructed, tested, and validated, it may be put to use. The uses include:

1. Choice of optimal set points for coating line operation.

2, Suggestion of **compensating** changes when the line drifts from those set points. Since **tome variables** can 'be made to change quickly and easily compared to others, GOfttpeiis^tifig changes, invoked by a control system should enhance the operation and **productivity .of the** process*

4.#,2»Th# Visual O#f«cis Experiment

Like the model for **optical** density, the mixiel for visual defects is built by searching for the best set of **discrimiiuuiiig** variables. The test for predictive **validity** is made using classification functions derived during the discriminant **analysis.**

The most general **classification functions** are Fishers Linear Discriminant Functions

$$Ci = C_n X_x + c_{i2} X_2 + \cdots c_{ip} X_p + c_{i0}$$

in which the C_i is the classification score for group i, and the c_{ij} are classification coefficients. [Nie 75] There is a function for each group; a case is classified into the group on which it has the largest score. Calculation of the coefficients c_{ij} is described in Norusis. [Norusis 79]

This type of classifier is called a linear machine. Duda and Hart discuss how such a classifier can be used to divide the independent variable space into regions in which the case is classified into a given group. [Duda 73] Such a division could be used to provide additional constraints on the range of permissible values for the input variables.

The linear functions are derived under the assumption that the groups are all of multivariate normal distribution, and that the covariance matrices for each group are equal. In the case where the covariance matrices are not equal, the discriminant functions are quadratics. Duda and Hart also discuss the use of quadratic discriminant functions to partition the test space; again, such a partitioning could be used to constrain the process operating limits. Both Tatsuoka [Tatsuoka 71] and Cooley and Lohnes [Cooley 71] discuss how classification can be made by evaluating the probability of group membership. This method is used by SPSS [Norusis 79]. Classification can take into account the *a priori* probability of group membership. In the study data this probability was assumed to be proportional to group size. Better than 90% of the study data were correctly classified. The true test of a classification function requires the correct classification of data other than that from which the function was derived. For such a test, new data are required.

4.5. Summary

On the basis of the study results, the hypothesis is made that optical density is a function of the paint and drying air variables. Visual defects are hypothesized to depend on these and sulfur dioxide flow. The effects of multicollinearity provide the incentive to use an experimental design in which the independent variables are uncorrelated. Inclusion of more observations allows power and interaction effects to affect the model, but more levels are required to model power terms explicitly. To test the hypotheses, it is recommended that a two level, seven variable experiment be run. A fractional factorial, 32 observation design represents a good compromise between experimental complexity and thorough investigation of the test space. It should result in data free from intercorrelation, and sufficient to investigate most two-variable interactions, as well as the main effects.

Automatic measuring instruments should be used to allow the model to be based on measured,

rather than calculated values. A response time on the order of one second is suggested; range and precision sufficient to resolve the range of normal operation arc necessary. The requirements for the data logger are mass storage capability, speed sufficient to process about nine readings per second, ten bit D/A resolution, adjustable prcamps for seven input variable instrument signals, and compatible input(s) for the output variable sensor(s).

The experiment may be most easily carried out by scheduling the observations to require the fewest adjustments for those variables most difficult to change. Analysis of the data generated by the controlled experiment can be done using the methods of Chapter 3. Both models should then be tested against new data to check their predictive validity. The resulting models can then be used to constrain the range of set points for coating line operation.

Chapter 5
Summary

5.1. Objectives of the Project

The objective of this project is to develop two models of the coating process on Line 1 at the Westinghouse Fairmont Works in Fairmont, West Virginia. The first model is a linear regression input-output model of the coating process relating optical density (the output) to the process inputs. The model consists of a linear equation with constant coefficients which allows calculation of the optical density of a tube from the values of the process inputs extant when the tube was made. The second model is a discriminant analysis input-output model of the coating process relating coating defects (the output) to input variable values. The discriminant analysis model consists of a set of classification functions which allow prediction of the presence or absence of a visual defect from the process input variable values.

5.2. The Study of the Coating Process

A preliminary, sample survey type study of the coating process was made on November 4th and 6th, 1981. Twelve input and two output variables were periodically sampled by a group of four to six experimenters using hand instruments. Sampling rates were slow, ranging from one measurement per minute to six measurements per hour. The study data were prepared for analysis by using the input variables to calculate interpolated input variable values for each output sample.

5.3. Analysis of the Study Data

5.3.1. Variance in the Data

The study data were analyzed first to see what variation of values occurred under normal operation. Five of the input variables were found to have a percentage of variation of less than one percent. The remaining seven input variables and the output variable, optical density, exceeded one percent variability. Variability ranged from 8.5 % for wash water conductivity to 0.15 % for paint specific gravity on November 6th. Variability of less than one percent seems too small to show up any effect. Obtaining a large range of values will require deliberate changes in variable values.

5.3.2. Optical Density

Simple correlations between the input variables and optical density were found to fluctuate wildly from one days' data to the next. The data from the first day were judged to be faulty and discarded.

A linear regression model was then found relating optical density to five input variables: drying air temperature, paint specific gravity, wash water conductivity, drying air velocity, and Lehr oven gas flow. The remaining seven input variables were found to be insignificant. There was moderate to strong intercorrelation in the data, making it impossible to find the sensitivity of optical density to changes in any one input variable. Intercorrelation, limited range of some variables, possible correlations to events on the line not measured (such as line stops), and possibly poor quality of data mean that this model should be considered to be preliminary.

The study was of value in that some insight into the process was gained, and one variable, wash water temperature, was found to be insignificant and to merit no further investigation. On the basis of this knowledge, a relation between optical density and some input variables was hypothesized. A controlled experiment to test the hypothesis was designed which will eliminate intercorrelation and limited range. The experimental data would be collected with automatic instruments, assuring good data and eliminating the need to use interpolated values. The hypothesis, experiment, and instruments are described in Section 5.4.

5.3.3. Visual Defects

A set of discriminant functions was derived relating visual defects to the process input variables. The input variables were then ranked according to their discriminating power. Based on this ranking it was hypothesized that the occurrence of a visual defect is a function of certain of the input variables. The hypothesis can be tested by a controlled experiment run concurrently with the optical density experiment.

A set of classification functions were derived which correctly classify 92% of the study data from which they were derived; new data would be needed to test their predictive validity. Because of limited range and possible inaccuracy in the data, the discriminant and classification functions must be regarded as preliminary.

5.4. The Controlled Experiment

The study was useful because it afforded an opportunity to become familiar with the coating process and the computational tools used to build the statistical models. It also showed up the shortcomings of a study in relation to a controlled experiment: limited range of variation and intercorrelated data. The shortcomings of hand-gathered data were also made obvious: the need for interpolated (and perhaps fictitious) values and the possibility of errors made by people doing a tedious job in uncomfortable circumstances.

To overcome these problems a controlled experiment is proposed. The experiment would test the following hypotheses:

1. Optical density is a function of the paint variables (specific gravity, pH, viscosity) and the drying variables (air velocity, humidity, air temperature).

2. The occurrence of a visual defect is predicted by sulfur dioxide flow rate, paint viscosity, drying air velocity, and paint specific gravity.

The experiment would use seven independent and two dependent variables. Three additional variables would be fixed. A two treatment level, 32 observation fractional factorial design is recommended. The experiment would be run without block randomization to reduce the effort needed to control the variable values. For optical density, this design allows investigation of all single-variable and most two-variable effects. Power terms may also be investigated if observations are made at intermediate levels.

An automatic data acquisition system should be used to perform the controlled experiment. Five new input sensors, and two new output sensors would be required. The sensors must be precise enough to resolve changes within the range of expected variable values and fast enough to record changes that occur in the time required to make one tube, whenever possible.

5.5. Interpretation of the Results of the Controlled Experiment

Data gathered from the controlled experiment may be analyzed using the same methods used to analyze the study data. A linear regression model can be built relating optical density to the input variables. The discriminant analysis model would consist of a set of classification functions. Both models should be tested against new data (from data set aside and not used to produce the models, or from new observations) to test their predictive ability.

The results of the analysis will allow prediction of the value of optical density and the presence of visual defects from input variable values. The relative contribution of each variable to the outcome can be found. The models can be used to constrain the set-points for the input variables, reducing the number of defective tubes produced.

References

[Bartee 68] Bartee, E. M.
Engineering Experimental Design Fundamentals.
Prentice-Hall, Englewood Cliffs, New Jersey, 1968.

[Box 70] Box, G. E. P., and Jenkins, G. M.
Time Series Analysis; Forecasting and Control.
Holden-Day, San Francisco, 1970.

[Cochran 57] Cochran, W.G., and Cox, G.M.
Experimental Designs.
John Wiley & Sons, Inc., New York, New York, 1957.
p. 282.

[Cooley 71] Cooley, W. W. and Lohnes, P. R.
Multivariate Data Analysis.
John Wiley & Sons, Inc., New York, New York, 1971.
pp. 55-57, 264-265.

[Duda 73] Duda, R.O. and Hart, P.E.
Pattern Classification and Scene Analysis.
John Wiley & Sons, Inc., New York, New York, 1973.
pp. 24, 27-30.

[Handelsman 82] Handelsman, M.
Automated Coating Inspection of Fluorescent Light Bulbs.
Master's Project Report, Robotics Institute, Carnegie Mellon University, 1982.
Unfinished.

[Moore 80] Moore, C.E.
Bulb-F40CW.
May 19, 1980.
Memorandum, Westinghouse Fairmont Works, Fairmont, West Virginia.

[Murray 82] Murray, J.
The Measurement of the Optical Density of Coated Glass Tubes on a Production Line.
Senior's Project Report, Robotics Institute, Carnegie Mellon University, 1982.
Unfinished.

[Neeter 74] Neeter,J. and Wasserman, W.
Applied Linear Statistical Models.
Richard D. Irwin, Inc., Homewood, Illinois, 1974.
pp. 218, 228, 347, 371, 382.

[Neeter 78] Neeter, J. and Wasserman, W.
Applied Statistics.
Allyn and Bacon, Inc., Boston, Massachusetts, 1978.

[Nie 75] Nie, N. H., Hull, C. H., Jenkins, J. G., Stienbrenner, K., and Bent, D. H.
Statistical Package for the Social Sciences
Second edition, , SPSS, Inc., Chicago, Illinois, 1975.
Published by McGraw-Hill Inc., New York, New York. pp. 280-281, 340, 346, 445, 447.

[Norusis 79] Norusis, M.J.
SPSS Statistical Algorithms, Release 8.0.
SPSS Inc., Chicago, Illinois, 1979.
pp. 75, 76, 81.

[Tatsuoka 71] Tatsuoka, Maurice M.
Multivariate Analysis.
John Wiley & Sons, Inc., New York, New York, 1971.
pp. 160, 165, 219.

[Wimberly 81] F. Wimberly.
CMU/Westinghouse Project on Automated Fluorescent Lamp Manufacture.
Document CMU-RI-81-1, Robotics Institute, Carnegie Mellon University, February, 1981.
p. 7.

[Younger 79] Younger, M. S.
A Handbook for Linear Regression.
Duxbury Press, North Scituate, Massachusetts, 1979.
pp. 322, 332.