A new modeling methodology combining engineering and statistical modeling methods: A semiconductor manufacturing application

by

Vikas Sharma

M.S., Massachusetts Institute of Technology (1991)
B.Tech., Indian Institute of Technology (1989)

submitted to the department of Mechanical Engineering in partial fulfillment of the requirements for the degree of

Doctor of Science in Mechanical Engineering

at the

Massachusetts Institute of Technology

September 1996

Signature of Author_____
Department of Mechanical Engineering
August 26, 1996

Certified by_____
Prof. Roy E. Welsch
Professor of Statistics and Management Science
Thesis Supervisor

Certified by_____
Prof. Steven D. Eppinger
Associate Professor of Management Science
Thesis Supervisor

Accepted by_____
Professor A. A. Sonin
Chairman, Graduate Thesis Committee

A new modeling methodology combining engineering and statistical modeling methods: A semiconductor manufacturing application

by
Vikas Sharma

Submitted to the department of Mechanical Engineering
on August 26 1996 in partial fulfillment of the requirements for the degree of
Doctor of Science in Mechanical Engineering

Abstract

This thesis aims to develop a new methodology to combine physics-based engineering modeling methods with data-driven statistical modeling methods. The use of engineering models alone or statistical models alone poses serious limitations in modeling a large system. Engineering models can model only small manufacturing operations (e.g., orthogonal cutting). Used alone, they are severely limited in modeling complex manufacturing systems, such as a whole manufacturing line. On the other hand, statistical methods can model large and complex manufacturing systems. However, statistical models may fail to capture true functional dependence, which is necessary for process control and process improvement. It is important to combine the two types of models to characterize a large system.

The methodology developed in this thesis research helps model an end-of-line (EOL) output parameter as a function of important in-line process parameters in a manufacturing line. The resulting model will help process engineers take proactive control action (feedback and feedforward), respond appropriately to special-cause signals in control charts, and provide improved understanding to develop future technologies.

There are two parts to the methodology. The first part focuses on the identification of important process steps for a given EOL output. It does so by exploiting the multivariate nature of EOL inspection variables. EOL inspection variables are the measurements taken on the final product. The second part of the methodology focuses on the development of a large-scale model for the EOL output as a function of parameters from important process steps. The framework of multivariate adaptive regressions splines (MARS) is used to develop the final model. The model combines piece-wise models from physics, empirical understanding about different process steps and data from regular production. As such, the final model should provide more predictability and improved process control than the one developed using only data or only physics-based models.

Using the methodology, we have modeled end-of-line (EOL) channel length as a function of in-line process parameters for a specific family of microprocessors at Intel Corporation. (Channel length is an electrically measured distance that an electron, or a hole, travels from the source to the drain of an MOS transistor. It is a key factor influencing device speed.) The standard error of our model is about a third of the acceptable error whereas the standard error of a leading multivariate statistical methods is more than the acceptable error. This means that while one in three predictions by the leading method may be off by more than the acceptable limit, that mis-prediction number using our model is only one in 322 predictions. These results are especially encouraging because our model predicts EOL channel length from

in-line process parameters. Equally importantly, our models help identify regions of process operation that result in largely good parts, and those that result in largely bad parts. Identification of such regions directly addresses the question of how to run a manufacturing process to improve the quality of the final parts. We believe that the methodology is applicable to many diverse manufacturing processes and output parameters.

Thesis Committee:
Prof. Duane S. Boning
Prof. Steven D. Eppinger (Advisor)
Dr. Robert D. Gordon
Prof. David E. Hardt (Chair)
Dr. Russ Sype
Prof. Roy E. Welsch (Advisor)

Key words: models, engineering, statistical, physics based, data driven, empirical models, hybrid models, non-parametric, local models, process control, dependent spec limits, manufacturing, semiconductor, MOS transistor, microprocessor, channel length, device physics, multivariate adaptive regression splines, MARS

Acknowledgments

This research was conducted mostly at the Rio Rancho site of Intel Corporation (Intel), and partly at the Massachusetts Institute of Technology (MIT). Support for the work at MIT was provided by the MIT's Leaders For Manufacturing (LFM) program and through a generous contribution from Intel's Research Council. Dr. Gene Meieran arranged for my initial travel funding from Intel's Research Council. I appreciate Dr. Meieran's confidence in me, and Dr. Rob Gordon, Prof. Roy Welsch and Prof. Dave Hardt's tireless efforts in finding sources for travel funds.

Dr. Rob Gordon coordinated the continuation of support for the work since June 1995 through the Yield Engineering Department, Fab9, Intel Corporation. I sincerely appreciate Cara Taylor and Larry Deshler's faith in my work. Without their support for this project through Fall 1995 and part of Spring 1996, this work would just not have been possible. I appreciate Larry's encouragement in helping me learn more about Intel's Q&R efforts. I am also very thankful to Dr. Russ Sype in taking on my supervisor's role at Rio Rancho after the departure of Dr. Rob Gordon to STTD Portland in September 1995.

It has been my privilege to work with one of the best minds and hardworking people at Intel. Their support and cooperation for this project is exemplary. I sincerely thank Ann Nelson, Jon Tetzloff and Mark Isenberger for doing my "academic exercises" during busy weeks. I appreciate Matt for providing "real-time" help with unexpected computer problems, and Paul's long hours that often ran late into evenings to get the data-extraction program ready. Marleigh and Neil helped restore data bases for data extraction ahead schedule during very busy weeks. I appreciate Brian's help with MMPC. He has agreed to use my routines to reproduce my results independently, and to try them on integration problems. It has been a sincere pleasure to work with each of the following who have contributed directly to this work: Reza Jazayeri, Carl Geisert, Tom Sciorilli, Ray Carey, Mike Westphal, Sandy Smiley, Debbie Rettke, Colleen Felsch, Mark Sorrell, Dr. Sang Kim, Peter Stoll, Patrick Dishaw, Melinda Hoppe, and Jay Mark. I look forward to working in future with all of them in the next part of this project.

I am also thankful to Gopal Rao for arranging my talk at the Intel's Engineering Lecture Series Seminar in September 1995. I am also thankful to Rick Chin, Steff Chanat, Tim Riddle and Divyesh Patel for their interest in this work.

My doctoral committee has been very supportive. Besides helping coordinate funds and logistics, Dr. Rob Gordon and Dr. Russ Sype have been always available for guidance during the course of this work. I have learned immensely through Dr. Rob Gordon's mentorship at Intel. My advisors Prof. Roy Welsch and Prof. Steven Eppinger's belief in my abilities to provide direction to this research effort is very much appreciated. Their timely helpful interventions in technical and administrative matters has helped me maintain a prospective course. I admire Steve's encouraging supervision, and cherish his friendship. Prof. Duane Boning's expertise in device and process physics has been of great importance in the last phase of the work that focused on modeling and prediction. I also express gratefulness to Prof. Dave Hardt for his support in making this work possible.

During the course of the work, many of my friends have been supportive of me. Sreeram, Pratyush and Srikrishna have always been a reassuring figures for me. I can never forget Donna Kaiser's help over the years with my writing, with discussions and refinement of ideas, and with providing a friendly support at many moments of despair during the research. Ujjwal and Atul's help in many small but important things will always be appreciated. Mike Peterson and Geoff made my stay in the small MIT office an agreeable one. The several timely helps of the secretarial staff at MIT has been just exemplary. While I cannot narrate all such helpful occasions, the prominent names are Leslie Regan, Joan, Joanne, Josh, Kim, Kathy, Jessica and Ana.

My family has been a source of inspiration and guidance throughout. My dad and mom's blessings and their faith helped me even thousands of miles away. My aunt and uncle in Rhode Island were always there for me when I needed them. Their kids, Anish, Anita and Kush helped me forget many of my frustrating moments during the research with their simple but enjoyable games. I appreciate the help and encouragement of my close family friends, Ajay and Rajul. Without their support and an enjoyable stay at their house in Albuquerque for almost a year, this work would not have been possible. Ajay and Rajul's sincerity for work, a positive attitude toward life and belief in hard work will be guiding factors for me in future too. I enjoyed spending time with their kids, Amit and Ankur.

Finally, I appreciate the patience of my wife, Pratima, who often stayed up almost till dawn to give me company as I feverishly scrambled to finish my thesis writing. I am also thankful for her effort in taking care of many administrative things, and admire her finesse of doing them even though she had just arrived in a new country. I hope I will be able to reciprocate her kindness in future when she will work on her thesis.

Table of Contents

List of Figures

List of Tables

List of Equations

1. Introduction

Recent consumer products provide many more functions with significantly improved accuracy. For example, microprocessors previously used mainly for scientific computations are now rapidly entering multi-media applications as a household communications tool. Besides improving computation accuracy and speed, they are now performing multi-tasking at an affordable price. Consumer cars now provide many more safety features (anti-lock brakes, air bags, etc.), comfort features (dual temperature zones maintained simultaneously) and improved functionality (greater tolerances on car body prevent leakage and enhance the life of the transmission system).

The multiplicity of product functions, and the need for improved product accuracy have resulted in relatively complex product designs. These in turn have made current manufacturing processes increasingly complex. They now comprise many more operations, and the desire for high precision in the final product demands a good understanding of each operation. Equally importantly, current manufacturing processes now also demand an improved understanding of the interaction between several operations. These interactions significantly influence the quality of today's high precision products.

Models of manufacturing processes help us understand these influences. Most existing modeling techniques are broadly of two types: first, the physics-based **engineering modeling methods**, and second, the data-driven **statistical modeling methods**. A third type of modeling method attempts to combine the two approaches, and gives **combined or hybrid models**. However, current modeling methods have proved inadequate in modeling large complex manufacturing lines.

This thesis aims to develop a new modeling methodology which combines physics-based and data-driven modeling methods to provide better predictability, improved controllability and an improved understanding of manufacturing processes.

In this chapter, we introduce the concept of hybrid models, and their pertinence to current manufacturing processes. We discuss a typical current manufacturing situation and the state-of-the-art engineering models, statistical models and existing concepts in combining the two approaches. We motivate the need for improved hybrid models, and thus a new modeling methodology. We relate the current limitations in hybrid models to the needs in a typical manufacturing environment, and show that a solution to those limitations can help address the needs in manufacturing. We then present the problem statement for this thesis, and then conclude this chapter with an overview of the organization of this thesis.

1.1 A note on terminology

Figure 1 shows a simple three-step canonical manufacturing line on which we define the following terms.

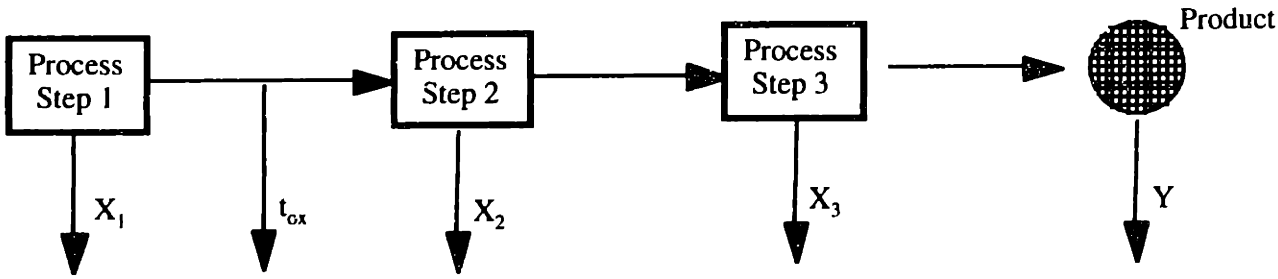

Figure 1. A typical manufacturing line

Process step: A machine (or a set of machines) which takes raw material(s), and changes its geometry and/or material properties. Examples include a diffusion furnace, an oven, a cluster of lithography machines (including a bake oven, a wafer spinning machine, a stepper, and a developer), an etch machine, an implant machine, a lathe, a milling machine, etc. Figure 1 shows three process steps. Material processed by a process step can serve as (part of) raw material(s) for the next process step. A process step is also called an **operation**.

Manufacturing line: More than one process step working in series, in parallel, or a combination of both. Figure 1 shows a manufacturing line consisting of three process steps working serially. A manufacturing line is also called **production line**.

Input: A variable used to manipulate a process step [27]. These variables are typically **machine parameters**. Examples include current, voltage, time of etch, intensity of UV light in lithography, feed rate in turning or milling operation. There are two kinds of inputs to a process. **Set points** are inputs which cannot be manipulated in real time. Examples include machine stiffness in milling, gas-inlet position in diffusion. **Controllable inputs** are inputs which can be manipulated in real time. Examples of controllable inputs are feed rate in turning or milling, time of etch, time of diffusion, beam energy in ion implant, etc.

Process variable: If we know the value of an input, it is called a process variable or **process parameter**, **input variable**, or **in-line**. A process parameter can be **categorical** (species A vs. species B chosen for doping single crystal silicon) or **continuous**, and can include set points and controllable inputs. Specific examples include pressure and temperature in a diffusion furnace, wavelength of UV light in lithography, etc.

Intermediate product measurement: A variable measured (or we wish to measure) on the product (or a surrogate product like a test wafer) while it is transitioning from one process step to another. Examples of intermediate product measurements are thickness of oxide grown in a diffusion furnace (measured after process step one in Figure 1), profile of carbon concentration in a gear after heat treatment, number of particles and the depth of material etched after an etch operation (measured after process step two in Figure 1).

End-of-line (EOL) inspection variable: When the product reaches the end of a manufacturing line, many measurements are taken on the product to test its functionality and appearance. The set of measurements made on the product (or a surrogate product like a test wafer) at the end of a manufacturing line are called EOL inspection variables. EOL inspection variables are a **multivariate response**. They are also signatures of the process steps that made the final product.

Output: An important characteristic of the final product that we want to model as a function of influential process parameters. Typically, one of the EOL inspection variables is the output. However, sometimes it is a function of a few EOL inspection variables. "Y" is an EOL output in Figure 1. Other examples of output are EOL E-tests, bin-splits, horsepower of an engine, etc. EOL output is also called **output variable**, **output parameter**, **response variable**, or **response**.

Often, the output is difficult to define in many manufacturing environments. Two types of difficulties arise in defining an output:

1. The variable of interest is sampled several times (or at several places) on a product. Examples of such variables include electrical tests on a wafer and thickness of photographic films. Since the output is considered to be a single number, an appropriate statistic on the sampled data has to be treated as the output. The statistic chosen for this thesis research is the median of the sampled data on the variables of interest.
2. A single variable measured on the product may not alone be an appropriate output characterizing the product. For example, after a car body is welded together, several measurements are taken on it. None of those measurements alone characterize the car body completely. However, all those variables together (and perhaps many other parts of the body on which no measurements are taken) characterize the car body more completely. Here, the individual variables should be combined to form a geometric feature about the car body, such as taper, shear, etc. An example of such a measure is the body-in-white (BIW) ruler developed by the author and York, and reported in [92]. The measured value of the geometric feature (such as the BIW ruler) can be then considered as the output for the purpose of modeling. Part of this aspect is discussed in Section 5.3.1 on modeling multiple important characteristics about a product.

Disturbance: Any unknown inputs and unknown variation in known inputs [27]. Examples include machine wear, unknown variation in beam intensity in ion implant, wafer position in three dimensions during lithography.

Engineering models: These are relationships between inputs and output, or between one input and several other inputs, derived using **physics of the process**. As such, they incorporate **causality relationships** between inputs and output.

Statistical models: These are purely **data-driven models** which exploit the **correlation** between inputs and outputs to relate the output to inputs. They are also called **empirical models**.

Hybrid models: These models use engineering models, empirical models and data. They are also called **combined models** or **integrated models**.

1.2 Industrial manufacturing background

This section motivates the need for an accurate multivariate model relating an output to influential inputs in a typical manufacturing line. The extensive use of univariate control charts in this section may superficially appear to stress the need for multivariate control charts. However, multivariate control charts are only one of the several control and monitoring options

made available by an accurate multivariate model between the output and the inputs. In the absence of a reliable multivariate model, manufacturing companies resort to the use of many univariate control charts.

Figure 2 shows a manufacturing line comprising two process steps connected serially. Univariate control charts are kept on process variables from process steps one (X_1) and two (X_2), intermediate product measurement (t_{ox}), and the end-of-line output variable (Y). Figure 2 also shows these control charts. The control limits are shown as dotted lines, and the process is considered to be capable so that the control limits are inside the specification limits (or spec limits). The discussion in the following paragraphs uses spec limits. However, the arguments apply equally well to control limits.

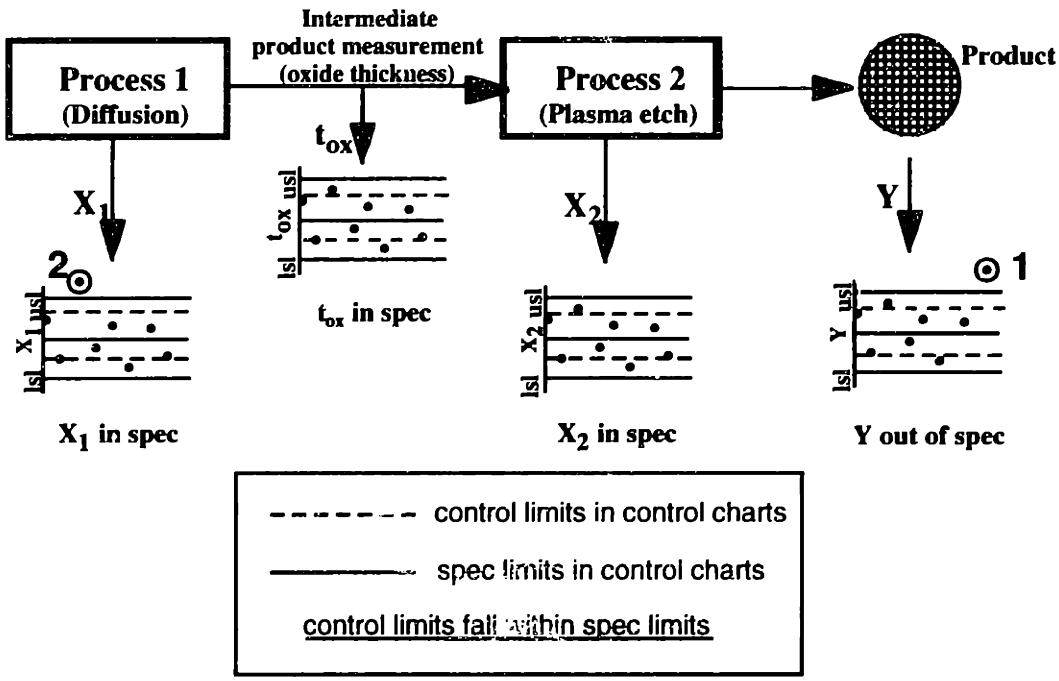

Figure 2. Two typical out-of-control problems in a manufacturing line

As long as all in-line variables (X_1, X_2 and t_{ox}) are in spec and the EOL output (Y) is also in spec, there is no cause for alarm. Such a situation in a large manufacturing line is unrealistically fortuitous. Frequently, two situations arise that raise concerns:

1. **Situation 1**: All measured in-line variables are in spec. However, the EOL output goes out-of-spec, such as the point marked "1" in Figure 2 [48]. The cause for an out-of-spec output cannot be found in the two process steps because X_1 and X_2 are within spec limits.

2. **Situation 2**: When process step one processes raw materials, X_1 goes out-of-spec, as in the point marked "2" in Figure 2. Should the out-of-spec point be ignored here to let the intermediate product go to process step two, or should a control action be taken? If a control action should be taken, what should it be [26, 55]?

The following sections attempt to explain the reasons for these two frequently occurring situations in a large manufacturing line. The definition of a large (and complex) system is discussed later on page 20.

1.2.1 Out-of-spec EOL output despite in-spec in-line process variables

Several reasons can contribute to the first situation where the EOL output is out-of-spec despite all measured in-lines being in-spec, point "1" in Figure 2. These reasons include:

1. **critical process variables not monitored**. If X_1 and X_2 are not critical process parameters for Y, then they have no information about the EOL output (Y). Perhaps some other process variables associated with process steps one and two, but which are still unknown, should be monitored.
2. **poor understanding of multivariate relationship** between EOL output (Y) and process variables (X_1 and X_2) even if X_1 and X_2 are known to be the only influential process variables for Y. For lack of a model relating Y to X_1 and X_2, only univariate control charts are kept on X_1 and X_2. Those univariate control charts cannot monitor the manufacturing process correctly in all situations [48].
3. **stochastic independence** between X_1, X_2 and Y. On a plot of the joint probability distribution between X_1, X_2 and Y, a $+/-3\sigma$ control limits on X_1 and X_2 will still leave a small, but finite, probability of Y being out-of-spec even if X_1 and X_2 are in-spec.
4. **critical process variables monitored improperly**. If the measurements for X_1 and X_2 are highly noisy, then the measurement noise could falsely put them within spec-limits. In fact, X_1 and X_2 may actually be out-of-spec.

The first two reasons appear promising in providing useful answers for process improvement. They are the focus of this thesis research.

Since the final product is made by both process steps, one and two, the EOL output (Y) would most likely depend on both X_1 and X_2. Figure 3 shows that Y depends on X_1 and X_2 jointly. In Figure 3, the x-axis is X_1 and the y-axis is X_2. The curved lines represent the lower spec limit (Y_{lsl}), target (Y_{target}) and upper spec limit (Y_{usl}) for EOL output. The two parallel

vertical lines represent the spec-limits for X_1, and the two parallel horizontal lines represent the spec-limits for X_2.

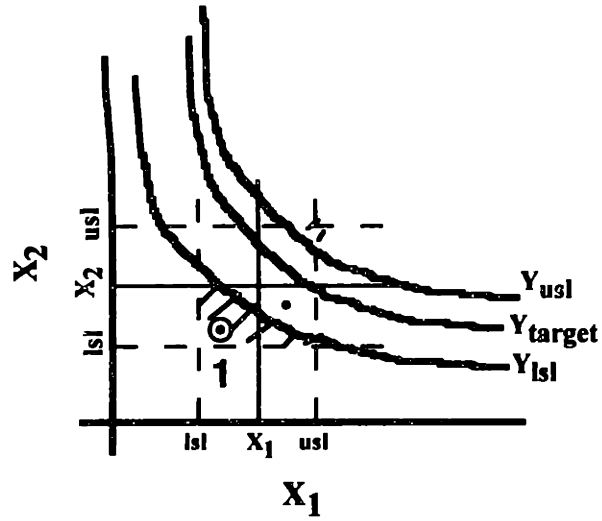

Figure 3. Response: A multivariate function of inputs

Equation 1 shows the mathematical relationship between Y, X_1 and X_2. For lack of prior knowledge about the system, Equation 1 assumes an additive error term, $f(X_1, X_2) + \varepsilon$, rather than a multiplicative error term $f(X_1, X_2)*\varepsilon$, for its simplicity in the model development process. The function (f) relating Y to X_1 and X_2 will include the interaction between X_1 and X_2 as shown in Figure 3.

$$Y = f(X_1, X_2) + \varepsilon, \ \varepsilon \sim N(0, \sigma^2)$$

Equation 1. EOL output: A function of in-line process parameters

The spec-limits for X_1 are not independent of those for X_2 because of the interaction between X_1 and X_2. This is contrary to the assumption in Figure 2. In Figure 3, the point marked as "1" in the cross-hatched area is within the univariate spec-limits for X_1 and X_2. However, it falls out of the spec limits for the EOL output (Y). This depicts the first situation described on page 15.

A **multivariate mathematical model** relating the EOL output (Y) to "influential" in-line process variables, X_1 and X_2, can help determine the cross-hatched regions in Figure 3. (Section 1.2.2 describes the concept of "influential" in-line process variables. The terms "influential" and "critical" are used here synonymously.) The model can be used in many ways,

one of which is the development of multivariate control charts. Such charts can help avoid process operation in the cross-hatched areas of Figure 3 that result in scrap at the end of the line. The multivariate model can also be used to identify robust process settings, to determine feedforward and feedback control actions, and in deciding whether a part should be scrapped earlier in processing if the variation introduced in it is already high.

1.2.2 Out-of-spec in-line process variable

In the second problematic situation, described on page 16, a measurement on an in-line process variable goes out-of-spec. Should the process engineers ignore this out-of-spec measurement, or should they take a control action? If this out-of-spec situation affects the EOL output, then by how much does it affect the EOL output.

Again, a multivariate model relating the EOL output to "influential" process variables can help determine if a response is needed for an out-of-spec measurement for X_1. The process of model development, and the final model will help determine if a particular process variable influences the EOL output. A process variable that influences the EOL output is an **influential process variable** for the EOL output. (This thesis research has developed a methodology to identify influential process variables for a given EOL output, discussed in Section 2.2.1). With a model relating the EOL output to influential process variables, one can handle an out-of-spec measurement for an in-line process variable more methodically. This is explained in the following paragraphs.

If the model relating the EOL output (Y) to influential process parameters does not include X_1, then "most likely" X_1 does not influence the EOL output (Y). (The term "most likely" is used here cautiously because of the considerations about causality vs. correlation in model development. These considerations are discussed further in Section 4.6.) An out-of-spec condition for X_1 can be ignored as far as Y is concerned. If Y is the only critical EOL output in our discussion, we could completely remove the sensor for X_1. The sensor is only causing false alarms.

On the other hand, if both X_1 and X_2 are an integral part of the model for Y, and if X_1 goes out-of-control (or out-of-spec), then one or more of the following control actions are plausible:

1. **Discard an intermediate product early** in the manufacturing line if the variation introduced in the intermediate product is already high. This will avoid unnecessary

subsequent value-adding activities if the final product is not expected to meet the specification.

2. **Identify robust set-points** to make the EOL output more robust (or insensitive) to variations in process variables that are hard to **control.**
3. **Tighten spec-limits** to avoid the cross-hatched areas shown in Figure 3.
4. **Feedforward control**. If X_1 goes out-of-control (or out-of-spec), X_2 settings can be changed for that intermediate product [39, 61, 75] to prevent the EOL output (Y) from going out-of-spec for the product.
5. **Feedback control** to avoid large variation in future batches [39, 61, 75].

Section 4.5.1 discussed these control actions in more detail. Here, the main point is that a model relating the EOL output to influential in-line process parameters, such as shown in Equation 1, is necessary in formulating and in evaluating different control strategies. An accurate model will help engineers make a more informed decision. Therefore, the **key is in developing an accurate model for the manufacturing process.**

1.3 Modeling research background

The previous section discussed the need for an accurate model that relates an EOL output to influential in-line process parameters. This section summarizes the state-of-the-art research literature on modeling.

Modeling techniques broadly fall under two headings. First, **physics-based** methods result in **engineering models** or **causal models**; and second, **data-driven correlation-based** methods result in **empirical models** or **statistical models**. A third category attempts to combine the first two types of models, and results in **hybrid models, combined models** or **integrated models**. The following sections describe each category of models.

1.3.1 Engineering models derived from physics

Physics-based engineering models have been developed and used in many disciplines of engineering and pure sciences. Using the physics of the process and product, this approach develops engineering models. The resulting model may help understand the underlying physics better, which would in turn further improve the physics-based engineering model.

Physics-based models do not assume a parametric relationship upfront. Instead, they use process physics and product physics to develop a mathematical formula that connects the relevant variables. An example of a process-physics model is the orthogonal cutting model in

machining/turning [5]. Examples of product-physics models are the several device-physics equations that describe the behavior of an MOS transistor [80].

An accurate physics-based model (or even the physics used to develop it) can help provide a list of influential variables for a given EOL output. The variables that are part of the model certainly influence the output, as discussed in Section 4.6.

Through the use of physics, engineering models provide a good understanding of the base-line connection between the response and the input variables. These models incorporate causality, and are very important for process control, as discussed in Section 1.4.3.

However, engineering models suffer from several limitations. These limitations are listed below:

1. Traditional physics-based modeling approaches do not incorporate **noise**. (Noise can be defined in several different ways. In this thesis, the term noise could refer to the unsystematic part of the signal, unknown dynamics of the system, etc.) The presence of noise is common in most data including manufacturing process data. The sources of noise include measurement error, sensor error, process disturbance in the form of variation in raw materials, machine settings, actuator systems, operator-to-operator variation, etc. Purely physics-based engineering models do not handle noise. Consequently, they fail to explain different output measurements from two similar operations at the same settings [67]. Statistical models can handle noise easily.
2. This modeling approach is **inadequate to model large complex systems**, such as a whole manufacturing line. (A later paragraph in this section presents the definition of a large complex system.) This is due to a large number of input variables, poor physics-based knowledge of the connections among them, and the presence of random noise. Modeling large systems, such as a whole manufacturing line, is a strength of statistical models, as discussed in the next section.
3. The **difficulty in calibrating** a large-scale model. Many constants in a large system depend on factors that are either poorly understood or are hard to measure. Such calibration constants will be difficult to estimate from a purely physics-based approach.

This paragraph discusses the definition of a large complex system. Specifically, it attempts to address the question "**When does a system become large and complex?**". For the purpose of this research, a system becomes large and complex in two situations

1. When there exist multiple steps in a system. Such a system is a multi-stage system with more than one process step, such as the one shown in Figure 2. In this situation,

physics-based understanding is inadequate to develop connections between variables from different steps of the system.

2. When the degree of accuracy needed is higher than that provided by the current physics-based understanding about the system. This situation can arise even with a single step process. Examples of this situation abound in orthogonal cutting, thin-film deposition by the diffusion process in semiconductor manufacturing, etc.

A methodology has been developed in this research which helps develop models for large complex systems. Chapter 2 describes the methodology.

The advantage of engineering models is best had in systems where the connections between all input variables, and their relationship to the output, are well understood and the amount of noise is insignificantly low. The reasons stated above demonstrate that engineering models, used alone, are inadequate to model large and complex manufacturing systems.

1.3.2 Statistical models derived using data

Statistical modeling methods use data to develop models. They consider the system as a black box, and develop an input-output relationship based purely on data from that system. The size of the system in the black box is generally immaterial; a given statistical technique would treat a small machine and a large manufacturing line alike for the purpose of developing a model. All statistical methods need a learning data set before they can be used on test data sets.

Statistical techniques incorporate process noise conveniently. However, all statistical techniques suffer from two general limitations. These limitations are:

1. Poor confidence in predicting outside of the range of values of data used for model development. This can be a serious limitation in the identification of robust settings for the process, especially if the robust settings are outside of the range of the current data.
2. Poor confidence in the first derivative of the output with respect to different inputs (even in the range of current data). This can have serious limitations in developing reliable process control strategies.

Physics-based engineering models do not typically have these limitations because they are derived from causality considerations. Statistical techniques can be further classified into parametric techniques and non-parametric techniques. The following sections explain and critically examine some important techniques.

1.3.2.1 Parametric techniques

Parametric techniques first assume a particular functional relationship (or function) between the output and the input variables. Then, they use data to estimate the numerical values of the coefficients in the assumed relationship with the objective of minimizing some measure of error. Typical estimation techniques include least squares or maximum likelihood methods [59]. From the resulting model, residuals are calculated by subtracting the actual value of the response from the one predicted by the model. The residuals are checked for normality (or that they all come from the same probability distribution), lack of correlation in time and for constant spread at different output (and input) values [15]. The residuals are considered to be stochastic noise if they pass these checks.

Parametric techniques include several regression methods, such as simple regression, forward and backward step-wise regression, etc. [59]. Forward and backward step-wise regression can be used to identify influential process variables for a given EOL output in a manufacturing line. Even if we ignore multi-collinearity (discussed in a later paragraph), all these regression techniques need process variable data to identify influential process variables. These techniques will fail if process variable data are unavailable. A method is presented in Section 2.2.1 to identify influential process steps for a given EOL output from a manufacturing line, even if process variable data from the influential process steps are unavailable.

Most regression methods assume that the data are uncorrelated in time. If the data are correlated in time, Multivariate Time Series Analysis (MTSA) is used to create independent data [8, 10, 83]. Alternatively, every fifth or tenth data point can be used. This is based on an often verified assumption that correlation reduces rapidly at higher lags.

Parametric regression techniques also assume that there is only one functional form characterizing the whole system. These will be called "**global modeling methods**". This assumption may only be valid in a small operating region of the system, and only if the system itself is very small. A large system, such as a whole manufacturing line, has several smaller regions of operation. The behavior of the system in one region could be very different from that in another. A global parametric regression technique would fail to characterize such a system adequately and accurately. (The definition of global modeling methods does not assume any specific error structure.)

Most simple regression methods are prone to high variance in the estimate of model coefficients when the input variables are highly correlated with each other [59]. An assumption of lack of correlation between input variables is an over-simplification for data from most

systems, including that from a manufacturing line. Statisticians term the occurrence of correlation between different input variables as multi-collinearity [59, 60, 87]. Parametric regression techniques are often applied to naturally occurring data, where the problem of multi-collinearity is very common.

Special parametric regression modeling techniques are used to avoid the problem of multi-collinearity. These techniques include ridge regression [59], principal components regression [51], partial least squares, factor analysis [51], design of experiments (DOE) [56, 57], response surface models (RSM) [47], etc. All these techniques work on the principle of creating or using orthogonal data to avoid the problem of multi-collinearity. Ridge regression, factor analysis, partial least squares and principal components regression can work on naturally occurring data. These techniques first identify (a smaller set of) orthogonal axes, and then transform the original data along the new orthogonal axes. On the other hand, DOE is a controlled experiment where the input data are deliberately orthogonalized to remove multicollinearity. A regression model is then developed using orthogonal data assuming a parametric function relating the output to the orthogonal inputs. Note that the orthogonal inputs could be from a DOE or from rotated axes. Several DOE designs exist to suit different systems and the type of inference an analyst wants to obtain about the system.

Designed experiments can be very expensive to run, and often disrupt regular production. Moreover, the final models (from DOE or any other technique discussed in the previous paragraph) are only as good as the assumed parametric function.

An exact functional dependence is difficult to assume a priori when modeling a large system using parametric methods. In the absence of any information about the variables, parametric methods usually assume very simple polynomial relationships, e.g., linear and quadratic with minimal interaction. Parametric regression techniques are simple to use, fast, and usually need little data for model development. They are effective if the assumed function accurately represents the process. Otherwise, they can be highly inaccurate.

The price of simplicity of parametric models is a strong possibility of losing much signal by a poor assumption about the parametric function. In addition, a single function may not be able to globally approximate all the data because of the instabilities of a polynomial over a wide range of the inputs [19].

To overcome the limitations of parametric techniques in modeling complex non-linear systems, and to develop better insight to the underlying phenomena, many non-parametric

modeling techniques are used. The next section discusses and examines a few important non-parametric techniques.

1.3.2.2 Non-parametric techniques

Instead of developing a parametric relationship, the overall idea here is to "categorize" data according to some criteria so that "similar" data are grouped in the same category. Within each category a unique approximating function is developed. Here, such methods are termed as **"local modeling methods"** because they assume different model structure in different regions of operation of the system. (There is no assumption about the error structure in the model.) Non-parametric techniques usually need much more data than parametric techniques for model development. The following paragraphs critically examine some important non-parametric tools.

1.3.2.2.1 Cluster analysis

Cluster analysis is a term used for a multitude of algorithms used for classifying numerical and categorical data [1]. Discriminant analysis and nearest centroid sorting methods like Forgy's, Jancy's and K-means start with a pre-specified number of clusters and can work only on numerical data. Methods of transforming categorical data into numerical data are not reliable [1]. Discriminant analysis also assumes that the underlying distribution of the sample and population is normal [37].

While most traditional cluster analysis methods are efficient and simple to use, they lack **interpretability**. Interpretability is defined here as a model's ability to divide the system into several smaller regions, and to **provide the basis of forming those regions**. Section 1.4.3 provides a greater discussion on model interpretability. Lack of interpretability makes cluster analysis difficult to provide (or even incapable of providing) a list of influential process variables for a given EOL output. If cluster analysis could easily provide the reasons for forming (or classifying) regions, then that basis could be used to identify influential process variables.

In and of itself, cluster analysis is of little use in developing models between different variables. Attempts to locally approximate functions in the clusters are generally unsuccessful in high dimensions. In statistics literature, this problem is known as the "curse of dimensionality" [17] (page 822). Most importantly, these methods do not incorporate prior information in the form of engineering models or empirical information. The ability to incorporate prior information is a very important feature of this thesis research, and will be presented in Chapter 2.

1.3.2.2.2 Classification and Regression Trees (CART™)

CART is a hierarchical method of "categorizing" data and then of developing regression models within each resulting category. CART can model Multiple Input Single Output (MISO) systems only [11]. (CART is the trademark of California Statistical Software, Inc.)

Using a learning data set, CART first develops a binary tree for the purpose of classification (or for developing partitions in the region of operation). Figure 4 shows one such binary tree. Note that exactly two branches emerge from each non-terminal node of the tree. Each terminal node (or leaf node) of the tree becomes one of the final categories. The CART tree in Figure 4 has three leaf nodes that produce three categories (or three small regions within the operating space). Each leaf node contains some or all the data falling in that category. However, a data point falls in only one leaf node.

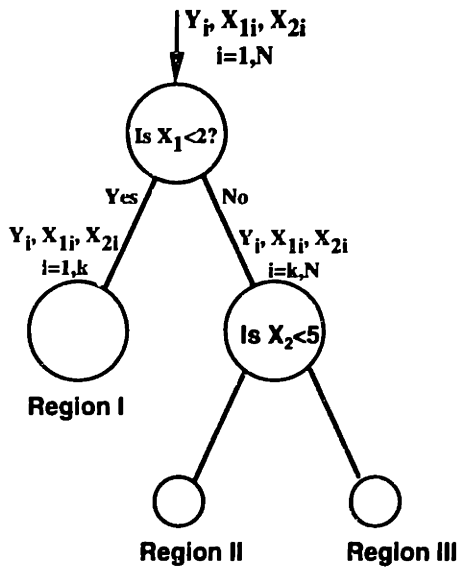

Figure 4. An example CART tree

The learning data set helps find suitable queries at non-terminal nodes (or split nodes). A suitable query is one which gives maximum purity (a pre-defined mathematical criteria) to the two resulting nodes of the binary tree. The classification process starts at the root node. After the selection of a suitable query, the root node results in two children nodes, the left child and the right child, as shown in Figure 4. Each child node now becomes a candidate for further splitting. This process of splitting is called **recursive partitioning**. Recursive partitioning continues until a pre-defined criteria is met. The resulting tree is then pruned to develop parsimony in the model.

Due to recursive partitioning that results in non-overlapping regions in predictor variable space, CART cannot develop **additive models** [12, 29]. **Additive models** are defined here as those that add contributions from several basis functions (or regions of the predictor variable space) to determine the value of the output. By contrast, only one basis function is used to predict a given value of CART's output.

Although a query at each split node uses an input variable, the purity function is based on the response. In this manner, a CART model uses both, the inputs and the response, in the process of model development. This is in contrast to factor analysis and conventional principal components regression (PCR) that determine orthogonal axes with no contribution from the response variable [20]. Lack of contribution from the output can be detrimental during model development because the orthogonal axes may not have any information content for the response, even though the original data may be rich in such information. This research has also modified conventional PCR in two ways: first, by including physics-based models and empirical information with predictor variables in developing regression, and second, by using all principal components and the output in developing regression [84]. To differentiate it from the currently used conventional PCR, the new PCR is called here the modified PCR, and is explained in Section 4.2.2.

Figure 5 shows a typical query at a split node of a CART tree. Currently, CART uses data only. For lack of information about the system that generated the learning data, only simple queries can be asked at split nodes. Examples of such queries include "Is X_1<2?", "Is Flag=true?", etc., where X_1 is an input variable. X_1 is a **split variable** because the CART tree formed a split using X_1. The **knot location** for X_1 at the root node is "2". Such queries lead to only rectangular partitioning, as shown by graphical interpretation of a CART model later in this section.

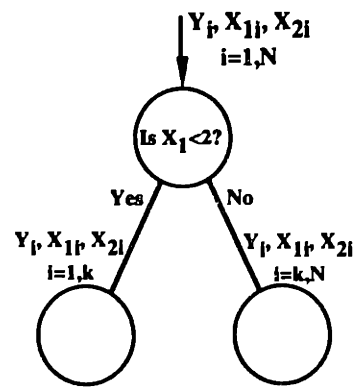

Figure 5. A typical query at split nodes of a CART tree

At the end of recursive partitioning, a regression model is developed. The model relates the output, $\hat{f}(x)$, to the various input variables used in the split nodes of the CART tree. These input variables are also called split variables. Equation 2 shows the mathematical representation of a typical CART model.

$$\hat{f}(x) = \sum_{m=1}^{M} a_m B_m(x)$$

Equation 2. A typical CART model

In Equation 2,

- M = total number of partitions created. M may or may not be pre-specified
- $\{a_m\}_1^M$ are the coefficients of expansion derived using least squares
- B_m's are defined below.

Equation 3 shows the basis functions for mth sub-region [16]

$$B_m(x) = \prod_{k=1}^{K_m} H[s_{km}(x_{(k,m)} - t_{km})]$$

Equation 3. General form of CART basis functions

In Equation 3,

- $H[\eta] = \begin{cases} 1 & \text{if } \eta >= 0, \\ 0 & \text{otherwise} \end{cases}$
- K_m = total number of non-terminal nodes to reach mth sub-region from the root node
- s_{km} = -1 or 1 for left child and right child respectively at kth non-terminal node
- $x_{(k,m)}$ = split variable at kth non-terminal node
- t_{km} = knot location at kth non-terminal node

Another representation for same basis function is:

$B_m(x) = I[x \in R_m]$, where

- I = Indicator function = $\begin{cases} 1 & \text{if argument is true} \\ 0 & \text{if argument is false} \end{cases}$

- $\{R_m\}_1^M$ are non-overlapping sub-regions *created* by CART using recursive partitioning on the learning data set.

The indicator function (I) results in a constant output value, a_i, within a sub-region, R_i. $\hat{f}(x)$ is a single order regression function with several discontinuities due to jumps at the boundaries of each R_i. CART can also be considered a piece-wise single-order parametric technique.

Figure 6 is a graphical representation of the CART tree shown in Figure 4. The x-axis in Figure 6 is X_1, and the y-axis is X_2. The box enclosed by the points A, B, C, and D represents the operating space of the system being modeled. The query at the root node of the CART tree in Figure 4 results in the line EF. The query at the right child of the root node results in the line GH. Note the rectangular partitions in Figure 6. This is because CART uses data only to create the regions. Any point in the operating space falls in only one partition. In each partition, CART develops a constant estimate of the output. This gives discontinuities to the estimated function at the boundaries of the partitions. In addition, the removal of a region, such as region II, would create a hole in the operating space. This is because other regions do not overlap with region II.

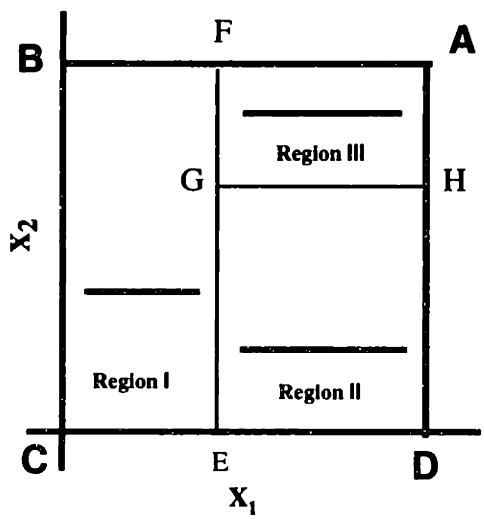

Figure 6. Graphical representation of CART tree

1.3.2.2.3 Multivariate Adaptive Regression Splines (MARS)

MARS is a modification of CART. It develops mathematical functions, $\hat{f}(x)$, that have (q-1) derivatives within R_m and a specified degree of continuity at the boundaries of R_m [23]. MARS differs from CART in two ways.

1. **Partitioning in MARS is non-recursive.** During partitioning, MARS can partition the current candidate node and any of its predecessors. A big advantage of this type of partitioning is in the creation of over-lapping regions in the space. The removal of a partition would not necessarily create a hole in the space, as illustrated graphically later in this section. Over-lapping regions give MARS the ability to develop additive models because a given value of the output can have contribution from more than one region. By contrast, CART partitions only the current node, but not its predecessors, and is unable to develop additive models. Since more than two partitions can happen at a node, MARS is no longer a binary tree. In addition, a data point can fall in more than one leaf node. As such it is difficult to even represent MARS as a tree.
2. **MARS uses different basis functions.** This results in polynomial functions within each small region. (CART gives only constant functions within each small region). A typical MARS model will also be mathematically represented by Equation 2. However, the basis functions for MARS are one-sided (or even two-sided) truncated power basis functions for representing qth order splines [14, 16, 19, 21, 22, 23, 36, 38, 63]. Equation 4 shows a typical basis function for a non-terminal node.

$$b_q(x\text{-}t) = (x - t)_+^q$$

Equation 4. Basis function for split node in MARS

In Equation 4,

- t is the knot location, and
- q is the order of the spline and the subscript indicates the positive part of the argument.

For $q > 0$, the spline approximation is continuous with (q-1) continuous derivatives. Equation 5 shows the general form of basis functions for MARS.

$$B_m^{(q)}(x) = \prod_{k=1}^{K_m} [s_{km}(x_{(k,m)} - t_{km})]_+^q$$

Equation 5. General form of MARS basis functions

In Equation 5,

- K_m = total number of non-terminal nodes to reach mth sub-region from the root node
- s_{km} = -1 or 1 for left child and right child respectively at kth non-terminal node
- $x_{(k,m)}$ = split variable at kth non-terminal node
- t_{km} = knot location at kth non-terminal node

Unlike Equation 4 for CART, Equation 5 develops polynomial functions. For q = 0, Equation 5 functionally reduces to the basis functions for CART. Thus for function development, CART is a special case of MARS.

Figure 7 shows MARS's graphical representation by modifying the CART tree of Figure 4. In Figure 7, region IV overlaps with regions I and II. The overlapping regions enable MARS to develop additive models [12, 29]. Instead of giving a constant output within each region, MARS generates polynomial functions. This gives continuous functions at the boundaries, and enables MARS to approximate the true system behavior more accurately within each region and in the operating space.

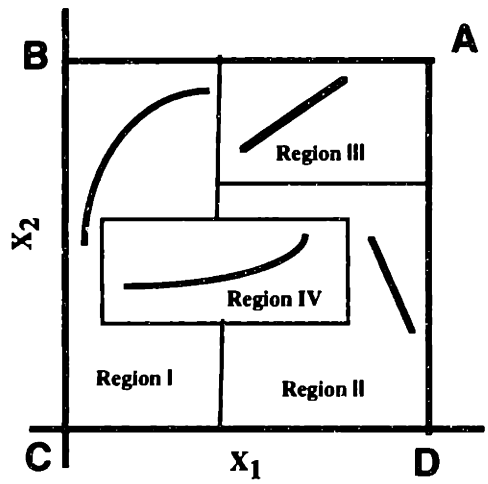

Figure 7. Graphical representation of MARS model

MARS model development is computationally intensive. Using a ranking methodology, Friedman [24] has reduced the MARS algorithm's computational effort by a factor of M (M = total number of final basis functions).

Several researchers have applied MARS to serially correlated data also. Lewis and Stevens [41, 42] have used MARS to develop a method for non-linear time series modeling called ASTAR (adaptive spline threshold autoregression). When applied to Icelanding river flow

data [42], ASTAR was more accurate, interpretable, and explicit in depicting the non-linearity in the river flow process than a complex model developed by the existing methodology of TAR (threshold autoregression) [78]. Applying MARS to do multivariate time series analysis is a research area in itself. This thesis research has used serially un-correlated data only.

MARS still has several limitations. Current applications of MARS use data only, and not any prior information, resulting in rectangular partitions like CART because of the nature of queries asked at the split nodes, as shown in Figure 7.

MARS has promise of incorporating prior information [23] (page 60), and this thesis research has identified five ways to incorporate prior information in MARS. Furthermore, this research has developed one of those ways, that of incorporating prior information as split variables, and has demonstrated its use in developing more accurate models.

1.3.2.2.4 Artificial neural nets (NN)

Neural networks (NNs) have recently become popular in modeling complex systems. They are often used in classification problems [18] and in recognition problems (speech and visual). The building block of NNs is a perceptron. It typically receives several weighted inputs, applies a defined parametric function on the sum, and then computes a single number as the output. Many perceptrons placed in layers result in a neural net. A training data set is required to determine the weights associated with each input of a perceptron in a NN. In this respect, they work analogously to non-parametric statistical classifiers. They can be viewed as high-order non-linear complex black boxes containing parametric models. Some authors prefer to call NNs a non-parametric way of modeling [17].

A taxonomy of NNs compiled by Lippmann [45] shows that NN's can result in only a pre-specified number of clusters for the purpose of classification. The number of clusters is determined by the number of nodes in the first layer in unsupervised nets, and by the number of nodes in the last layer in supervised nets. Unsupervised nets are those that do not have information concerning the correct class of data during training, where as supervised nets have this information. The use of multiple nodes in the input layer and in the output layer enables a NN to develop multiple input multiple output (MIMO) models.

However, conventional perceptron-layered NNs lack the ability to create more clusters on their own. They also lack the ability to develop interpretable models. Interpretability is important for process control and for process improvement. See Section 1.4.3 for a further discussion on model interpretability. NNs may also not converge in many situations.

De Veaux [17] and Psichogios [62] have compared MARS and NNs in system identification/modeling problems. They have shown that a MARS model is faster to develop than a NN model. Despite being multiple input single output (MISO) by nature, MARS models show superior performance over the MIMO NN models [17, 62]. Using MARS to develop MIMO models is another interesting area of research, and has been recommended as part of future work.

1.3.3 Hybrid models/Integrated models

A third category of modeling methods have attempted to combine the physics-based modeling approach and the data-driven modeling approach. These methods usually use one of the following two approaches:

1. first use the information about the system or physics-based understanding about the system, and then perform data analysis.
2. first perform data analysis, and then use physics-based understanding about the system to interpret the final results.

The first approach has been used more extensively than the second one. The following sections describe the two approaches.

1.3.3.1 Use engineering knowledge before data analysis

When engineering knowledge or system knowledge is used before data analysis, the goal of using the knowledge is to accomplish one (or more) of the following three activities:

1. **Determination of inputs**. This helps determine what inputs should be used in the final model development.
2. **Feature extraction from inputs**. This helps combine the inputs together as an expression, which may have a greater signal to noise ratio than the use of single inputs.
3. **Assumption about a parametric form**. This activity helps guide selection of parametric function relating the response to the inputs and features.

After accomplishing these activities, data analysis is performed to estimate the values of the coefficients in the assumed parametric function. The techniques used to estimate coefficient values aim to reduce a measure of the error. The following paragraphs provide examples of research in several areas that use engineering knowledge before data analysis. Each example accomplishes one or more of the three activities above before performing data analysis.

Physical knowledge and data have been used in many different fields of engineering, pure science and statistics. In thermodynamics, the state of a real gas is given by an equation with empirically determined constants [81] (page 421). However, the underlying combination of temperature, pressure and volume comes from prior knowledge about gasses. In Materials Science, there is a significant effort to develop phase diagrams from first principles. These phase diagrams are compared to those already determined using data only. To calculate the natural frequencies of beams and plates with several unusual loading conditions, the effect of inertia is first modeled by combining variables using prior knowledge from physics and vibration theory. Then, data analysis is performed to estimate parameter values associated with combined variables [4]. A similar strategy is followed for studying different regimes of non-laminar flow over solids of different shapes (in the area of Fluid Mechanics) [86] (page 325, 420), and for studying convective heat transfer over surfaces of different shapes (in the area of Heat Transfer) [31] (Chapter 6).

In the areas of machine design [72] (page 662 and on many other pages), design of heat engines [90] (page 467), combustion of fuels [58] (page 122) and materials testing, first principles (engineering knowledge) understanding is used in combining variables and then the parameter values in the different assumed relationships (additive, multiplicative and exponential parameters) are estimated by using data.

System knowledge has also been used in many statistical modeling methodologies. In DOE, knowledge about the system is used to decide on the choice of factors, the levels (2 for a factor with linear effect and 3 for quadratic), the resolution of the design, and which factors are allowed to confound with each other [3, 8, 57]. The selection of variables and their interaction in any parametric empirical model is always based on apriori information, if available [30, 56, 57].

Many researchers transform the original variables (by a linear or a non-linear combination) or create dimensionless numbers by using physical understanding of the system. Then, they use these dimensionless numbers and transformed variables in many different statistical modeling techniques. These techniques include DOE [44, 65, 89], CART [11] (page 131), MARS [23] (page 59, 60). etc. For a solder joint classification problem, Eppinger et. al [18] have extracted several engineering features from a solder joint data set, and then used them as inputs to a NN.

1.3.3.2 Use engineering knowledge after data analysis

Other researchers perform data analysis first, and then use physics-based knowledge or system knowledge to interpret the results. Examples of such researchers are Hu and Wu [33, 91]. They performed principal components analysis on autobody measurements. The principal components helped identify the planes of maximum variation. Using information about assembly equipment, Wu then identified the fixtures within those planes that introduced high variation.

In general, principal components and singular value decomposition (SVD) of data can help provide insight into the physical driving factors of a system. Physical interpretation of the parameters derived from DOE can also provide insight into the physics of the process and into the numerical estimates for the physical parameters of the system [9] (Chapter 12).

1.4 Motivation

1.4.1 A few observations

In many manufacturing companies, engineers typically understand the influence of one process step at a time on the EOL output. Statisticians term such influences as main effects of different process variables on the EOL output. Often, the understanding of main effects is qualitative. Some companies have developed statistical tools to understand main effects quantitatively. This understanding of main effects has improved product quality and yield in many manufacturing companies. The need for further improvements in yield and product quality now demands a greater understanding of the influence of process steps on the EOL output.

Current state-of-the-art modeling methods seem to be lacking in providing this understanding. When applied alone, neither physics-based modeling techniques nor data-driven modeling techniques are capable of modeling large, complex, non-linear systems such as a whole manufacturing line. The current techniques to combine the two approaches also fall short. The limitations of these techniques, and how they fail to account for the characteristics of a large system are outlined in the next section.

1.4.2 Summary of current limitations

Engineering models fail because they cannot relate variables that are not known to be connected by the laws of physics. A large manufacturing line has many such process variables. Moreover, purely physics-based models do not incorporate noise. Process noise is a common feature in manufacturing systems.

Parametric statistical methods make a big assumption about the system upfront, and usually develop global models. Global models fail to capture the desirable local deviations of a system, such as a manufacturing line. Non-parametric methods do capture local details but conventional clustering algorithms are not interpretable, and cannot be used easily for model development. Generalized additive models do not capture multiple variable interaction. In addition, most of these methods do not provide an effective mechanism to incorporate prior information.

The typical approach of combined models is to use engineering knowledge before performing data analysis or vice-versa. This approach comprises only a single jump from one domain to another. Such an approach will also be unable to model a large system. This is due to the lack of a systematic use of the engineering domain and the statistical domain more than once. In a large manufacturing line, little process physics is known about many process variables that are generally understood to influence the output variable. Clearly, physics-based information will fail to identify them as influential input variables. In addition, several unknown non-linear interactions may exist between several input variables. Unlike the typical approach of existing combined model development techniques to extract features, unknown interactions may not allow a simple way to extract features a priori. Due to multiple regions of operation of a large system, the assumption of a parametric relationship between the output and the input variables will be an invalid oversimplification. Thus, traditional ways of model development by combined modeling schemes fail for a large manufacturing line. Moreover, even if the final model appears interpretable, multi-collinearity between the input process variables confounds control decisions.

All this points toward the need for a method that

- can model a large system, such as a whole manufacturing line,
- combines physics-based understanding, empirical information and data, and
- captures true local deviations in the system

The desired characteristics in the resulting model are outlined and explained in the following section.

1.4.3 Desired model virtues: predictability, controllability and interpretability

The final model for a manufacturing line should have the following three properties:

1. Predictability
2. Controllability
3. Interpretability

1.4.3.1 Predictability

The model should be able to predict the output accurately from in-line measurements. In Figure 8, the solid line represents the actual output of the system. The dots represent the predicted value of the output from the model. Note the closeness of the dots to the solid line, except for the point "A" which is discussed later.

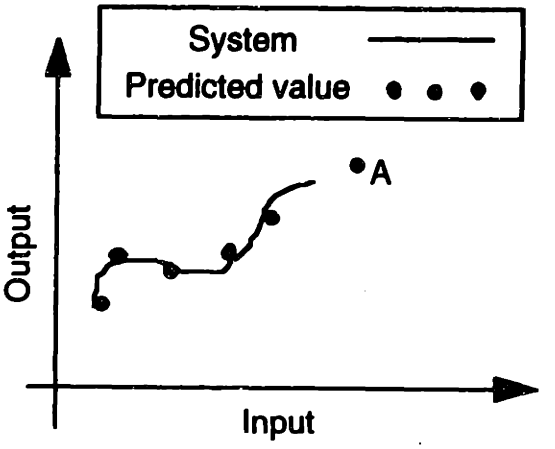

Figure 8. First desired virtue: Model predictability

A new region of operation is one from where no data were used for model development. An example of prediction in a new region is shown by the point "A" in Figure 8. Here, data are not collected. However, the desire from the system model is that the point "A" falls close to the predicted value by a simple extrapolation of the model.

Since purely data-driven models may lack the capability of predicting in a new region, it is important to incorporate physics-based knowledge. This will help in identifying better operating points for the manufacturing process. The term "better" could refer to more robust or higher yielding operating points.

1.4.3.2 Controllability

In Figure 9, assume that the input and the output variables were sampled at only two points. One could fit a straight line through them, shown as M_1 in Figure 9. With some

uncertainly, one can also fit a quadratic function through the sampled data. The quadratic line is shown as M_2 in Figure 9.

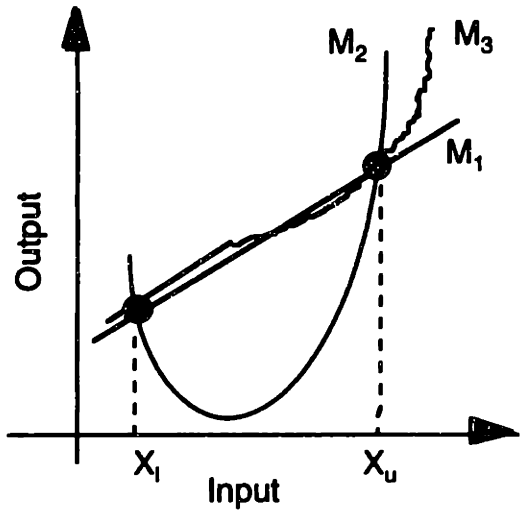

Figure 9. Second desired virtue: Process controllability

Note that both models, M_1 and M_2, predict the output accurately at both sample points. However, besides accurate prediction, we also want to know by how much the inputs should be changed to bring the output on target. If the current operating point is X_l, then M_1 suggests a completely different control action from M_2 in the neighborhood of X_l. This is because the sign of $\frac{\partial M_1}{\partial \text{Inputs}}_{X_l}$ is opposite to that of $\frac{\partial M_2}{\partial \text{Inputs}}_{X_l}$. Poor confidence in the first derivative of the output with respect to the inputs is a typical problem with purely data-driven models.

If physics-based knowledge about the system suggests that the output behaves linearly near X_l but quadratically near X_u, then we could force the model to fit a straight line near X_l and a quadratic line near X_u. This is shown as M_3 in Figure 9. Besides predicting the output accurately, note that M_3 also provides the correct sign for the first derivative of the output with respect to the inputs near X_l and X_u. As such, M_3 predicts accurately and provides for good process control.

An accurate first derivative is important for reasons of process control. Otherwise, engineers would not know how much, or even in which direction, to change input variables to bring the output on target. Purely data-driven models do not necessarily provide a good estimate of the first derivative. Physics-based models fulfill that requirement because they are causality-based models.

The obvious preferred approach is to combine data-driven modeling techniques with causality-based models. The combination would use the ability of data-driven techniques to model a large manufacturing line. It would also provide for good process control due to the causality-based connections derived from engineering models.

1.4.3.3 Interpretability

A large manufacturing line has many regions of operation. The output depends on a certain combination of process variables at one operating point. It could depend on a completely different combination of process variables at a different operating point. Interpretable models clearly show the different regions of operation of a process, and the combination of process variables in the different regions of operation.

In Figure 10, there are three regions of operation because the mathematical relationship of the output and the inputs is different in the three regions. t_{ox} and q_{ox} are thickness of gate oxide and charge on the gate respectively. To understand interpretability of models, they can be considered in this section as intermediate product measurements in a manufacturing line. Note that the combination of inputs in region I is the same as that in region II. Region III depends on a different combination of input variables than region I and region II.

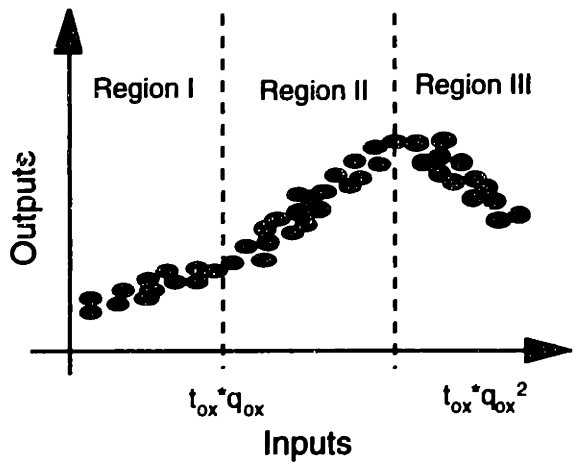

Figure 10. Third desired virtue: Interpretability

Such model interpretability helps identify the combination of process variables that produce good parts, and those that produce bad parts. In addition, it develops appropriate (and even different) mathematical relationships between the output and the appropriate input variables in the different regions of operation.

1.5 Specific problems addressed in this research

This thesis research has developed a new modeling methodology that combines engineering knowledge, statistical models and data to model a large manufacturing line. The objective of the methodology is to produce more accurate and interpretable models, see Figure 11.

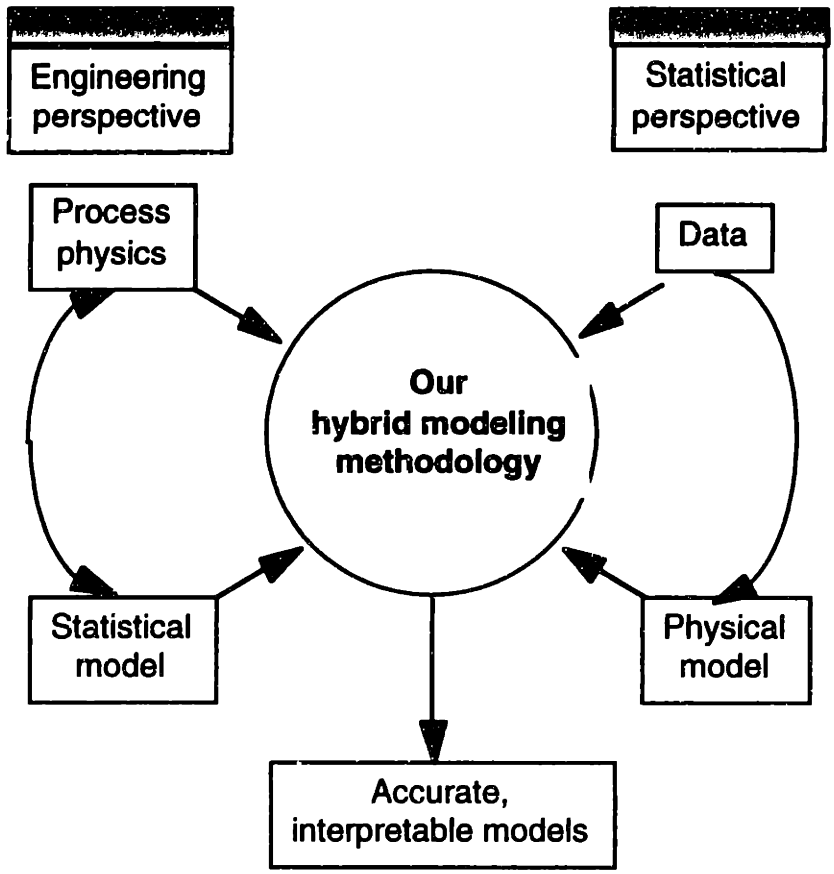

Figure 11. Our combined methodology: Objectives and inputs

This research has also attempted to answer the following two important questions:

1. Are combined models (developed using engineering knowledge, empirical models and data) better than engineering models or statistical models used alone?
2. Do local modeling methods characterize a manufacturing line better than global modeling methods?

1.6 Thesis overview

This thesis presents and discusses our hybrid methodology. It introduces an industrial problem at one of Intel Corporation's microprocessor manufacturing facilities, and analytically

applies our hybrid methodology to the industrial problem. It also applies alternative techniques to the industrial problem, and compares their results with the one obtained by using our hybrid methodology. This thesis also discusses issues with modeling and with the use of final models for process control and improvement.

- **Chapter 2** presents and discusses our proposed research methodology which combines engineering knowledge, statistical models and data. This methodology first helps identify influential process steps. It then helps develop a large-scale model relating the output to process variables from the influential process steps.
- **Chapter 3** introduces an industrial problem at one of Intel Corporation's high-volume micro-processor manufacturing facilities. It introduces the difficulties associated with modeling end-of-line channel length, and then analytically applies our hybrid methodology to the industrial problem.
- **Chapter 4** presents and compares our models to two alternative models. With the help of the models developed using our methodology, this chapter identifies several issues with process control and process improvements.
- **Chapter 5** summarizes the insights gained by this thesis research and makes recommendations for future investigation.

2. Research methodology: hybrid model development

2.1 Overview

Modeling a large system, such as a whole manufacturing line, is a challenging problem. This chapter aims to list and explain the steps in a new hybrid modeling methodology developed in this research. The hybrid methodology combines physics-based engineering models and data-driven statistical modeling methods.

The methodology can be broadly divided into two parts, each with a distinct function, as shown in Figure 12. The first part of the methodology identifies influential process steps for a given end-of-line (EOL) output. There are two outputs modeled in this research, channel length for n-channel transistors (L_{en}) and that for p-channel transistors (L_{ep}). The process parameters associated with those process steps are the potentially influential process parameters for the EOL output. The second part of the methodology then develops a mathematical model relating the EOL output to the potentially influential process parameters.

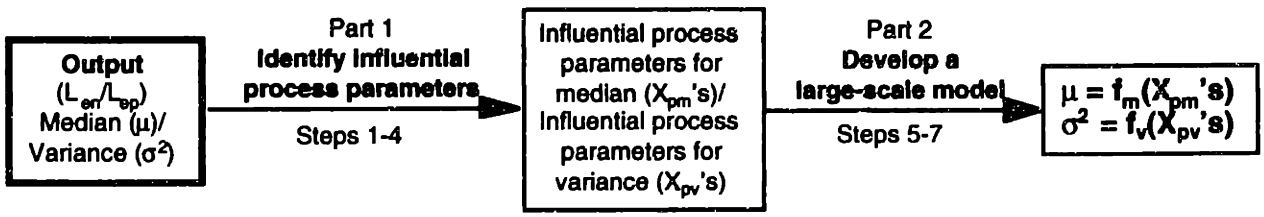

Figure 12. Two parts of the hybrid methodology

The methodology can be used to model the inter quartile range (or variance) and the median (or mean) value of the EOL output. Inter quartile range and median can be considered as a robust estimates of the variance and the mean, respectively. Both the median and the variance are important to model because

- the median and the variance would together characterize the EOL output better than only one of them would. If the EOL output is distributed normally, then by modeling its median and variance, we can fully characterize the EOL output. A caveat here is that data from manufacturing processes are not always distributed normally. However, median and variance will still be important parameters to model, even if they do not characterize a non-normal distribution fully.
- In a typical manufacturing line, a few process steps may influence only the median value of the EOL output. On the other hand, other process steps may influence only the variance of the EOL output. (Obviously, there could be a few process steps in a manufacturing line that influence both the median and the variance of the EOL output). To keep the EOL output on target, we need to understand the process steps that influence the median. Equally importantly, to improve process capability, we need to understand the process steps that influence the variance of the EOL output.

In Figure 12, some of the parameters identified as potentially influential for the median (X_{pm}'s) can be different than those for the variance (X_{pv}'s). Moreover, the mathematical function relating the median to X_{pm}'s can also be different from the one relating the variance to X_{pv}'s.

The current thesis research demonstrates the applicability of the hybrid methodology in modeling the median value of end-of-line channel length of metal oxide semiconductor (MOS) transistors. We believe that the methodology is also applicable in modeling the variance of EOL channel length, and have recommended that as part of future work.

While divided into two parts, our hybrid methodology consists of seven general steps. The first four steps fall in the first part of the methodology. They aim to identify potentially influential process parameters for a given EOL output. The last three steps fall in the second part of the methodology. They aim to develop an exact mathematical relationship between the output and the potentially influential process parameters. The methodology combines physics-based knowledge, statistical modeling approach and data in steps four, six and seven of the methodology.

The seven steps of the methodology are listed below:

Step 1: Choose EOL inspection data and identify model relationship

Step 2: Collect EOL inspection data

Step 3: Develop model and identify important EOL inspection variables

Step 4: Identify process steps influencing important EOL inspection variables and output

Step 5: Collect process data

Step 6: Determine piece-wise engineering and statistical models

Step 7: Develop large-scale model relating output to process data

Section 2.2 explains the seven steps in the context of the two functions (parts) of the hybrid methodology stated in Figure 12. Section 2.3 outlines and discusses the generalizable domains of the seven steps. Subsequently, Section 2.4 adds further perspective to the hybrid methodology by discussing its general applicability in different manufacturing environments. Section 2.4 also discusses a few special cases of missing information in different steps of the methodology. This chapter concludes with details of step seven of the methodology in Section 2.5. As part of step seven, we have customized multivariate adaptive regression splines (MARS) for our modeling purposes. Section 2.5 provides the details of how we have extended the use of MARS to combine physics-based models, empirical understanding and data to model an end-of-line (EOL) output.

2.2 Seven-step methodology for hybrid model development

2.2.1 Identification of influential process parameters

A typical manufacturing line has hundreds of process steps (and thousands of process parameters associated with those process steps). Figure 13 shows the domain of all end-of-line (EOL) inspection variables measured on a product made by such a manufacturing line.

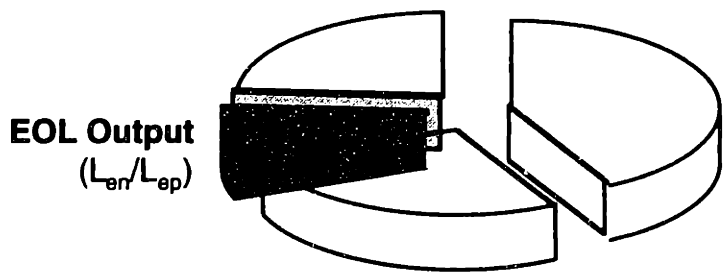

Figure 13. Domain of all end-of-line inspection variables

The dark pie wedge on the left in Figure 13 is the EOL output that we are interested in modeling as a function of influential process parameters. This thesis research has modeled EOL channel length (L_{en} and L_{ep}). Figure 14 shows the domain of all process steps in a manufacturing line. Only a few process steps in the manufacturing line influence the EOL output.

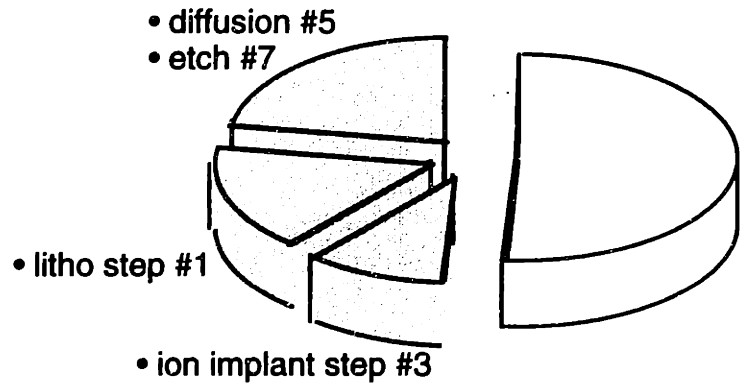

Figure 14. Domain of all process steps in a manufacturing line

The process steps diffusion #5, etch #7, litho step #1 and ion implant step #3 in Figure 14 represent the influential process steps for the EOL output (L_{en} and L_{ep}) in Figure 13. This section aims to explain the algorithm that helps identify the dark pie wedges (shown in Figure 14) that influence the output, shown as the dark pie wedge in Figure 13.

The main assumption of the algorithm is that the **EOL inspection variables are surrogate signatures of process steps** in the manufacturing line that made the product. Using this assumption, the algorithm exploits the oft-existing **multivariate relationship between EOL inspection variables.** It then uses several sources of information to determine potentially influential process steps for the EOL output.

The algorithm does not need process data to identify influential process steps because of the use of physics-based models, existing simulation tools and past experience of company personnel. As such, the algorithm can identify even those process steps as being influential for which no data are available. This is a strength of this algorithm.

The algorithm begins with the identification and selection of EOL inspection variables that may be related to the output. We use system knowledge, understanding of process physics and product physics (e.g., device-physics models) for this selection. The understanding from

physics can be in the form of existing formal models. It can also be in the form of qualitative reasoning developed by extending the underlying principles of physics.

To follow the steps of the methodology, consider the problem of modeling EOL channel length with only a limited set of variables EOL inspection variables and in-line process variables. The same example is considered later in Chapter 3 with the full range of relevant variables used in a real manufacturing set up. Assume that only channel length measurements, threshold voltages (V_t's), breakdown voltages (B_v's), electrical critical dimensions (ECD) for channel and for metal interconnects are available as EOL inspection variables. The EOL output is the channel length. Physics-based knowledge tells that ECD for the channel, threshold voltage, and breakdown voltage are important EOL inspection variables for channel length. However, ECD for metal interconnect is not an important variable. According to step one, ECD for channel and V_t's are selected for further use. ECD measurements for metal interconnects are discarded.

In addition, we identify several inherent model structures. These model structures could be between EOL inspection variables (Vt's and ECD for channel in the example) and the EOL output (channel length in the example), such as local or global relationships, univariate and multivariate relationships. If applicable, we can also identify relationships between several EOL inspection variables, such as nesting and hierarchical relationships. Identification of such model structures is important for future analysis when we develop a mathematical model (in step three of the methodology). The selection of relevant EOL inspection variables and the identification of underlying model structures mark the completion of step one of the methodology.

The purpose of step one is to reduce the number of EOL inspection variables to be used in the subsequent steps of the methodology. The rationale for this reduction stems from practical considerations of sample size and limited data handling capability of software packages, and is discussed in Section 2.4.2.

As part of step two of the methodology, we collect data for the EOL inspection variables identified in step one. We also collect data for the EOL output. For the example here, data are collected for V_t's, B_v's, ECD of metal interconnect and channel length. The end of data-collection marks the completion of step two.

The purpose of step three of the methodology is to further reduce the number of EOL inspection variables. (As such, the number of EOL inspection variables is first reduced by step one and then further by step three.) As part of step three of our methodology, we develop a statistical model relating the output to all EOL inspection variables for which we collected data in step two. We use the underlying relationships (local/global, univariate/multivariate, nested and

hierarchical) identified in step one to develop this statistical model. The modeling tool used here should be able to develop local models if several important local relationships were previously identified in step one.

Examples of modeling tools that develop local models are classification and regression trees (CART), multivariate adaptive regression splines (MARS). Examples of global modeling tools are simple regression, principal components regression (PCR), and partial least squares regression (PLS). Neural nets are hard to classify between local and global modeling tools.

In the example of channel length, a statistical model is developed that relates channel length to V_t's, B_v's, and ECD of the channel. Figure 15 depicts the development of the model as an arrow from the EOL output to the rest of EOL inspection variables in the domain of all EOL inspection variables. The objective of the statistical model is not to determine accurate model parameters. However, its objective is to provide a list of EOL inspection variables that are related to the EOL output in a statistically significant manner.

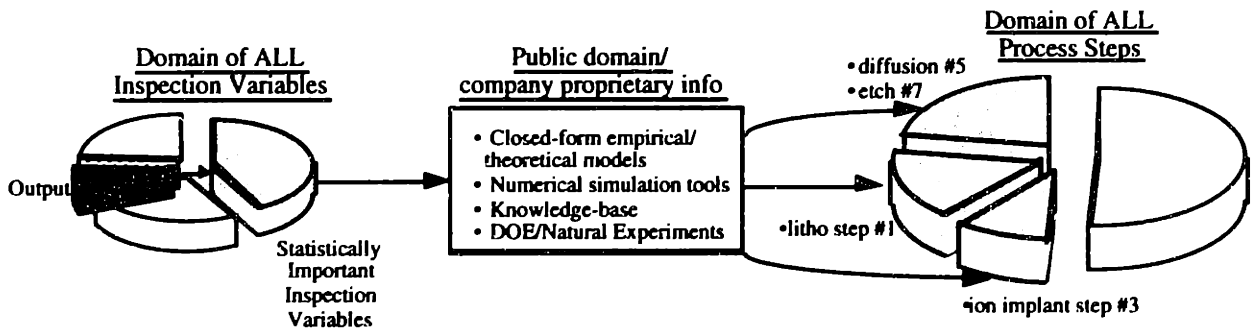

Figure 15. Algorithm to identify influential process steps

From the resulting statistical model, statistically significant EOL inspection variables are identified that can together predict the output within the desired level of accuracy. One can use standard statistical tests, such as t-test and F-test, to identify statistically significant EOL inspection variables. Alternatively, one can develop one's own algorithm for the same goal [52, 53, 54]. The identification of statistically significant EOL inspection variables marks the completion of step three of the methodology.

The statistically significant EOL inspection variables are called here secondary EOL inspection variables. In Figure 15, the secondary EOL inspection variables are represented by the shaded pie wedges on the right in the domain of all EOL inspection variables. For the channel-length example, assume that both V_t's and ECD of channel are identified as statistically

influential variables, but B_v's are not identified to be statistically important. V_t's and ECD of channel are called secondary EOL inspection variables.

The first three steps of the methodology have helped identify a list of EOL inspection variables that can predict (or model/explain) the EOL output within the desired level of accuracy. Now, the process steps that influence each of the secondary EOL inspection variables (and the EOL output) remain to be identified. Those process steps should be able to predict (or model/explain) the output with approximately the same accuracy as the secondary EOL inspection variables do. This is because the secondary EOL inspection variables are surrogate signatures of the process steps. The next step of the methodology (step four) helps identify process steps that influence each of the secondary EOL inspection variables and the EOL output.

As part of step four of our methodology, information from several sources is used, and is summarized in Figure 15. These sources are explained below.

1. **Theoretical and empirical models** found in text books and research literature can help identify influential process steps. For example, the threshold voltage (V_t) of an MOS transistor is related to its substrate doping concentration according to Equation 6 [80]. The parameter (γ) is a measure of the implant dose. Consequently, both well-implant steps (n-well and p-well) are important process steps that influence V_t. Therefore, they also influence EOL channel length of an MOS transistor.

$$V_t = \phi_{MS} - q_{ox}/C_{ox} + \phi_B + \gamma(\phi_B + V_{sb})^{0.5}$$

Equation 6. V_t as a function of substrate doping concentration

In Equation 6,

- ϕ_{MS} is the difference between the bulk potential and the gate potential
- q_{ox} is the charge on the gate per unit area
- C_{ox} is the capacitance of the gate oxide per unit area
- ϕ_B is the surface potential at the interface of the oxide and the substrate
- γ is the body-effect factor, and includes a measure of the implant dose in the substrate
- V_{sb} is the potential difference between the source and the substrate

2. **Numerical simulation tools**: These tools are typically developed or applied during product and process development. They relate different process steps to different product characteristics (including several secondary EOL inspection variables and the

EOL output), numerically. Examples of such simulation tools are Suprem, Pisces, Abaqus, etc. Although publicly available, different manufacturing companies usually customize these tools for their specific products and processes. Regardless, publicly available simulation tools can still provide useful information about influential process steps.

3. **Knowledge bases**: During product and process development (including the time when they both are in research phase), and during normal factory operation, several pieces of new information emerge, and are stored in knowledge bases. Often, existing formal models are incapable of explaining the new information. Thus, these data bases serve as independent sources to identify influential process steps.
4. **Experience of personnel** from factories and research and development sites is often an important source of information. Integration personnel from different sites can provide information from their accumulated experience over several generations of products and processes, and over several different product lines and process technologies.
5. **Natural experiments and design of experiments (DOE)**: One can also collect data for process variables either from existing data bases or by running a DOE. The data in data bases are typically collected during normal operation, and is also called happenstance data or data from natural experiments. By doing relevant statistical analysis on such data, one can identify influential process steps for secondary EOL inspection variables (and the EOL output). This source of information relies on process data, and was not used in this thesis research.

Step four of our methodology relies on several sources of information. Many of these sources may give similar results. We strongly recommend the use of as many sources as possible. This will result in a more complete set of influential process steps. The use of several sources of information also provides more confidence in process steps identified by more than one source. The influential process steps are shown as shaded or cross-hatched pie wedges in the domain of all process steps in Figure 15.

For the channel length example, knowledge bases (information source #3) and experience of personnel (information source #4) may show that spacer-etch step and the diffusion step that grows gate oxide are important steps for ECD of channel. Well implant steps were earlier identified to be important. Therefore, influential process steps for channel length are well implant and spacer etch steps and the diffusion step that grows gate oxide.

The identification of a list of process steps that influence each of the secondary EOL inspection variables (and the EOL output) marks the completion of step four of the methodology. The process parameters associated with these process steps are potentially influential process parameters for the EOL output.

To summarize, steps one through four of the hybrid methodology identify potentially influential process parameters for the EOL output. First, a statistical model relating the EOL output to the rest of relevant EOL inspection variables is developed. From that model, statistically significant EOL inspection variables, called secondary EOL inspection variables, are identified. Using five sources of information (listed earlier in this section), process steps that influence each of the secondary EOL inspection variables and the EOL output are identified. The process parameters associated with these process steps are potentially influential process parameters for the EOL output.

2.2.2 Development of a large scale model for end-of-line output

Steps one through four in Section 2.2.1 helped identify potentially influential process parameters for the end-of-line (EOL) output. By doing so, the number of candidate process parameters for the final model development (in step seven of the hybrid methodology) will typically reduce from a few thousand to a few tens. This section now aims to explain the algorithm that helps develop a large-scale model using steps five through seven of our hybrid methodology. The model relates the EOL output to the potentially influential process parameters, see Figure 12.

The final model (from step seven of the methodology) is called a large scale model because it is a model for an output at the end of a large manufacturing line as a function of process parameters from several different places in the manufacturing line.

Despite modeling a whole manufacturing line, the algorithm preserves local relationships between the output and different process parameters. For example, the output may depend on a certain set of process parameters in one operating region (such as the one in which short channel effects are predominant in an MOS transistor). However, the output may depend on a different set of process parameters in another operating region (such as the one in which long channel effects are predominant in an MOS transistor).

In addition, the algorithm combines physics-based understanding about the product and the manufacturing process steps, empirical understanding of process engineers about different process steps and data collected during normal operation. This combination gives a more

accurate model for the EOL output. Our algorithm is explained in the following paragraphs through steps five through seven of our hybrid methodology.

This section now picks up on our methodology at step five (steps one through four are discussed in Section 2.2.1). Step five is data collection for the process parameters that are potentially influential for the EOL output. These data are used later for the final model development in step seven of the hybrid methodology. In the channel length example, data are collected for etch rate of the spacers, temperature and pressure in the diffusion chamber that grows gate oxide, and the energy of the beam in the well implant steps.

In step six of the hybrid methodology, we identify pieces of information about the manufacturing line and the final product. We call these pieces prior information. Prior information can take several different forms. The following lines describe them.

1. **Formal physics-based models** from text books and research literature: These models could be represented as equations relating one variable (or several variables) to a set of variables.
2. **Formal empirical models**, such as the ones derived using DOE or RSM (response surface models).
3. **Qualitative understanding of engineers** and technicians: Often, experienced process engineers and technicians understand many nuances of manufacturing process behavior that are not captured by any formal models (physics-based or empirical). In manufacturing companies, this is often referred to as the "feel" for the process. Experienced engineers and technicians often use their "feel" for the process to improve the operation of their specific (set of) machines. Such knowledge is mostly qualitative by nature, and does not usually exist in a formal quantitative form. However, formalizing this qualitative knowledge may be important for several reasons. It can improve our understanding of all (critically important) machines in a manufacturing line, rather than that of just a few. This will help in providing better process control for critically important process steps. As such, the final product will have tighter tolerances. In addition, tolerance trade-off between different process steps in a manufacturing line can be better understood. Such qualitative understanding can be formalized as simple linear equations, and can be used in the final model development (in step seven of the methodology). Later in Section 3.5, we will present three examples of formalizing such qualitative understanding. Both those pieces of information proved useful in the final models presented in Chapter 4.

4. **Features**, such as the ratio of two variables provides more information about a product characteristic. However, the actual relationship between the product characteristic and the feature may not be known accurately. The features could be derived from physics-based understanding or empirical understanding of the process and product.

   For the example of channel length, three different models can be used

- Qualitative understanding of engineers can be used to determine a measure of implant dose from measurements on the energy of the beam in implanters.
- The measure of dose can be substituted in Equation 6 to estimate V_t, and
- Existing empirical models that connect spacer etch rate to ECD of channel.

The hybrid methodology treats each piece of prior information as a new variable in the final model development in step seven. A piece of prior information can be a physics-based model, an empirical model, a feature or a simple equation representing the qualitative understanding of process engineers. Appendix A illustrates the process of creating new variables through a numerical example. It is recommend that the user gather as much prior information as possible. The reason for doing this is explained in Section 5.1.3. The end of the collection of prior information marks the completion of step six of the hybrid methodology.

The last phase of the hybrid methodology is the development of a final large scale model which relates an EOL output to important process parameters. The model combines physics-based knowledge and empirical understanding about the process with data collected during normal operation. Step seven of the methodology develops this final model.

Before discussing the model development process, let us recapitulate the inputs for the final model. The inputs are

1. **data** for all potentially influential process parameters identified at the end of step four of our methodology, see Section 2.2.1. For the channel length example, they are beam energy in well implant steps, etch rate of spacers, and temperature and pressure in the diffusion chamber that grows gate oxide.
2. **numerical values for all the different pieces of prior information**. Prior information is collected in step six of the methodology, see previous paragraphs in this section. Data for influential process parameters are used to calculate (or estimate) numerical values for prior information. Since prior information is collected for influential process steps, and since data are also collected for influential process steps, the numerical values for prior information for different observations can be calculated using process data for different observations. In the channel-length example, numerical

values are determined for implant dose, V_t's and models connecting spacer etch rate to ECD of channel. Sometimes, a piece of prior information may contain a process variable in its formula (or model) for which no data are collected. One possible numerical value for this process variable is the nominal target value defined by the recipe of the process. However, it is highly recommend that one use an estimated value for this process variable, if it can be estimated from other data.

The output for the final model is obviously the EOL output.

Step seven of the methodology focuses on the final model development. For the inputs described before, many types of statistical tools can be used to develop the final model. Examples of these tools are CART, MARS, PLS, PCR, NN, etc. PCR and PLS are global modeling methods. On the other hand, CART and MARS are local modeling methods. MARS is chosen for the final model development in this thesis research by extending its use to implement the seventh step of the methodology. However, the methodology is general to allow the use of several different mathematical tools in step seven.

The final model may have fewer variables than the ones identified by the first four steps of the methodology in Section 2.2.1. This can be due to several reasons, and may have important implications in process monitoring and control. These reasons and implications are discussed later in Section 4.6.

2.3 Domains of generalizable seven steps of hybrid methodology

Our hybrid methodology combines the use of physics-based models of manufacturing processes and products, empirical knowledge about manufacturing processes and data. For clearer categorization, we put physics-based models in the domain of engineering models. We also put empirical knowledge in the domain of statistical models. Data are represented in their own independent domain called the "Data" domain. Figure 16 shows the three domains-engineering models, statistical models, and data-as three long horizontal lines.

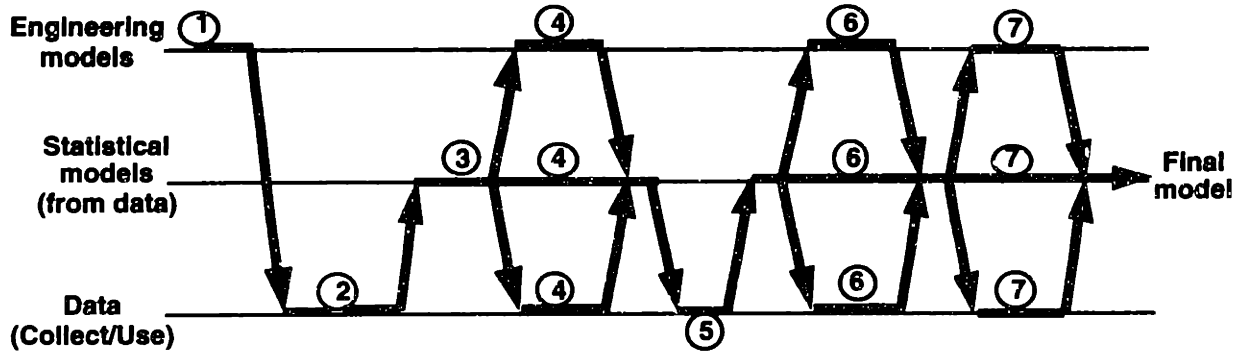

Figure 16. Hybrid methodology uses three domains systematically

The previous section, Section 2.2, described the steps in our hybrid methodology. This section now aims to depict our methodology's systematic use of the three domains. For details on the objectives and workings of the seven steps of the methodology, see Section 2.2.

In Figure 16, the numbers in circles on the three domains represent the steps in the hybrid methodology. These numbers run from one through seven. The number one being the first step of the methodology. It helps choose EOL inspection variables, and helps identify model relationships. The number seven being the last step of the methodology. It aims to develop a final large-scale model.

In Figure 16, the dark solid line connecting the circled numbers depicts progress in modeling an EOL output parameter through the use of the seven steps of the methodology. The presence of the dark solid line on a domain (shown as a long horizontal line) represents the use of information from that domain. For example, the dark solid line under the number one implies the use of engineering knowledge in step one. The presence of the dark solid line on the "Data" domain implies the use of or the collection of data. Note that a given step of the methodology can use information from more than one domain. As such, it would be represented on more than one horizontal line in Figure 16. Examples of the steps that fall in more than one domain are steps four, six and seven.

The inclined dark lines connecting two domains (shown as long horizontal lines) in Figure 16 represent jumps from one domain to another. Obviously, transitions to steps four, six and seven, which fall in more than one domain, are shown as dark inclined lines ending on several long horizontal lines.

Now, let us understand the hybrid methodology's systematic use of the three domains. Step one helps choose EOL inspection variables, and helps identify model relationships (univariate/multivariate, local/global, nested and hierarchical) through the use of engineering

information and system knowledge. Thus, it falls in the engineering domain, see Figure 16. The presence of the dark line under the number one depicts the use of the engineering domain.

The methodology moves to data-collection activity in step two. This is shown as the dark line under the number two in Figure 16. A dark inclined line from the "Engineering models" domain to the "Data" domain depicts the transition from step one to step two of the hybrid methodology.

Step three of the methodology aims to develop a statistical model relating EOL output to other EOL inspection variables. It also aims to identify secondary EOL inspection variables, see Section 2.2.1 for details. Step three falls in the domain of statistical models. A dark inclined line from the "Data" domain to the "Statistical models" domain depicts the transition from step two to step three of the methodology.

Step four of the methodology aims to identify influential process steps for each of the secondary EOL inspection variables and for the EOL output parameter. To do so, step four uses information from several sources. These sources are listed in Section 2.2.1. Some of those sources fall in the domain of engineering models, e.g., theoretical models. Some of them fall in the domain of statistical model, e.g., knowledge bases and experience of personnel. Yet, others fall in multiple domains. For example, natural experiments typically fall in the domains of statistical models and data. Thus, step four of the hybrid methodology uses information from all three domains. It is represented on all three domains in Figure 16.

At the end of step four, the methodology has identified influential process steps for the EOL output parameter. The process parameters associated with these process steps are potentially influential process parameters for the EOL output. Through the remaining three steps-five through seven-the methodology now aims to develop a large scale model relating the EOL output to the potentially influential process parameters.

After step four, the methodology transitions to the data-collection activity of step five. This activity is shown as a dark solid line under the number five in the "Data" domain.

Step six of the methodology aims to identify prior information about the manufacturing process and product. Different pieces of prior information typically explain different particular characteristics of the manufacturing process and product. Step six uses information from several sources to identify prior information. These sources are listed in Section 2.2.2. Some of those sources fall in the domain of engineering models, e.g., formal physics-based models. Some of them fall in the domain of statistical model, e.g., qualitative understanding of engineers and

technicians. Yet, others fall in multiple domains. For example, formal empirical models and features can typically fall in all three domains-engineering models, statistical models and data. Thus, step six of the hybrid methodology uses information from all three domains. It is represented on all three domains in Figure 16.

Step seven of the methodology then develops a large scale model which relates the EOL output to influential process parameters. Step seven combines piece-wise engineering models, statistical models and data. As such, it falls in all three domains.

Note that the hybrid methodology moved from step one through step seven. While a few steps fall in one domain only (e.g., steps one, two, three and five), other steps fall in multiple domains (e.g., steps four, six and seven). Multiple domains help gather more information. Greater amount of information helps increase model accuracy. We have shown later in Chapter 4 that the use of prior information (by including piece-wise engineering and empirical models) from step six resulted in more accurate final models than the ones developed using data only.

Steps four, six and seven of our methodology use information from all three domains. However, we recognize that information from all three domains may not always be available. Our hybrid methodology would still be applicable. For example, assume that piece-wise empirical models are unavailable in step six. The methodology would use piece-wise engineering models and data from step six to develop a final model in step seven. The final model may be less accurate than the one that also incorporated piece-wise empirical models. However, the methodology does not break-down if piece-wise empirical models are missing.

The next section, Section 2.4, provides greater discussion on the methodology's workings in the face of missing pieces of information. The section also comments on the generality of the hybrid methodology.

2.4 A few observations on the hybrid methodology

The previous sections explained our hybrid methodology in its most basic form. However, they have not discussed if the methodology is only applicable when the manufacturing process or the EOL output have certain characteristics or if the methodology is more generally applicable. In addition, the previous sections have also not discussed the applicability of our methodology in special situations of missing information. The following sections discuss these issues.

2.4.1 Generality of our hybrid methodology

This thesis research shows the applicability of our hybrid methodology in modeling a continuous EOL output parameter in a largely batch-processing semiconductor manufacturing process. (In microprocessor manufacturing, most process steps are batch processes. Only a few process steps are continuous processes). However, we believe that the methodology should be applicable to different types of EOL outputs (continuous and categorical) in a variety of manufacturing processes (batch processing and continuous processing). This is explained in the following paragraphs.

The seven steps of our hybrid methodology do not assume any characteristics about the EOL output or the manufacturing process, see Section 2.2. They aim to accomplish certain activities. These activities include the development of models, collection of information (quantitative knowledge and quantitative models) and data, and interpretation of the information gathered. Rather than assuming special characteristics about those activities, the methodology depicts each activity in a general way. For example, the methodology provides the flexibility to use any relevant model relationship (univariate/multi-variate, local/global, nested and hierarchical) in step one. The manufacturing process characteristics will determine the choice of the model relationship actually used in step one. Again, step three of the methodology provides the flexibility to develop any statistical model, and step seven provides the flexibility to use any of the several modeling tools.

None of the model development steps or the information gathering steps have characteristics that would bias them in favor of certain manufacturing processes (or EOL outputs) or bias them against others. The systematic use of information from the three domains-engineering models, statistical models and data-also assumes no prior knowledge about the type of output or about the characteristics of the manufacturing process. Note that the depiction of the hybrid methodology on the three domains is a general depiction without reference to any type of EOL output or manufacturing process.

As such, we believe that our hybrid methodology is general enough to model a variety of EOL outputs (continuous and categorical) in a variety of manufacturing environments (batch processing, continuous processing, or even mixed such as the one used in microprocessor manufacturing). In addition to modeling a whole manufacturing line, the methodology can also be used to model only a (set of) process step(s). A set of process steps is also called a cluster of tools. Section 5.1.1 presents an example of a continuous film-making process, and shows how

the steps of the methodology will be applied to the continuous process and to a largely batch-processing semiconductor-making process.

2.4.2 Why is step one important?

Several colleagues have wondered about the usefulness of step one of our hybrid methodology. (Step one of the methodology helps choose relevant EOL inspection variables, and helps identify model relationships between them. Step one discards EOL inspection variables that, based on our understanding of process physics and product physics, would not have any connection with the EOL output. Data are collected for the chosen EOL inspection variables in step two. Step three develops a statistical model using the model relationships identified in step one). A few colleagues suggest the use of all EOL inspection variables, thus by-passing step one. They suggest the development of a statistical model in step three using all EOL inspection variables, and then the identification of secondary EOL inspection variables from the statistical model. Such an approach can give erroneous results because of two practical considerations.

1. EOL inspection data are only a (preferably large) sample from a (much larger) population
2. The software for data analysis can handle only a limited set of data

These two considerations create serious limitations in a purely data-driven inference. For reasons of limited sample size and data-handling capability, a mere coincidence of numbers can make an EOL inspection variable an important one in modeling an EOL output. However, from an understanding of product design and manufacturing process, the EOL inspection variable may just not possibly have a connection with the EOL output. (The author confirmed the presence of such erroneous EOL inspection variables in modeling the EOL channel length in MOS transistors.) The use of such EOL inspection variables in subsequent steps of the methodology would complicate the modeling process, at best. More likely, it would introduce spurious process variables that may be hard to identify and eliminate. Spurious process variables confound control decisions, and impede progress in process improvement. Therefore, it is important to minimize (and possibly eliminate) erroneous conclusions about EOL inspection variables early in the modeling process.

For the channel-length example, ECD of metal interconnects may have been identified a statistically important inspection variable if it were not discarded earlier in step one. The geometry of an MOS transistor clearly shows that ECD of metal interconnects cannot influence channel length. If its statistical importance is identified in step three of the methodology, it would be purely due to a coincidence of numbers.

Given the limitations on sample size and data-handling capability, it is necessary to discard useless EOL inspection variables. Useless EOL inspection variables are those that cannot possibly have any connection with the EOL output, based on our understanding of process physics and product physics. By removing such useless EOL inspection variables, step one of the methodology reduces the chances of erroneous identification of secondary EOL inspection variables in step three of the methodology.

2.4.3 Other end-of-line (EOL) inspection variables missing

This section discusses the situation when data are collected for only one variable at the end of the manufacturing line. The variable for which data are collected is the EOL output. Since no other measurements are made on the product, there are no other EOL inspection variables. For example, assume that only the speed of the microprocessor is measured at the end of line inspection. No other electrical test characteristic is measured. Such situations are rare, and we have not come across one. However, this consideration is included here for the sake of completeness.

This section explains the workings of the methodology when the only variable for which data are collected is the EOL output parameter. Obviously, the methodology cannot identify secondary EOL inspection variables because no data exists for model development in step three. There are two options here:

1. The preferable option is to brainstorm to identify other product characteristics that may have relevant connection with the EOL output. For the microprocessor example in this section, such product characteristics are the threshold voltage, electrical measurement of poly-silicon width, etc. Now, assume that these newly identified EOL inspection variables are surrogate secondary EOL inspection variables. Then, apply step four of the methodology to identify influential process parameters for the surrogate secondary EOL inspection variables and for the EOL output. The rest of the methodology is then applicable in the usual way described in Section 2.2.
2. If there is no product characteristic with relevant connection with the EOL output, then we cannot even identify surrogate secondary EOL inspection variables. In this case, we skip the first three steps of the methodology. We apply step four of the methodology to identify influential process parameters for the EOL output. The rest of the methodology is then applicable in the usual way described in Section 2.2.

2.4.4 Only data available in step six

Assume that no prior information (in the form of piece-wise engineering and statistical models) are available in step six of our hybrid methodology. However, only data for potentially influential process parameters are available. Such situations arise during pilot runs in new product and process development, particularly when the experience from past products and processes cannot be extended to the new product and manufacturing process. This section aims to address the issue of how the methodology works when no prior information are available.

In the absence of prior information, step six of the methodology resides only in the "Data" domain. It is absent from the domains of engineering models and statistical models, see Figure 16. The methodology uses the available information in data to develop a final model in step seven of the methodology. The modeling tools used in step seven will have fewer input variables because additional variables from prior information are missing.

If prior information is missing, the methodology uses only limited information, whatever is available in data. The final model may not be as accurate as the one that also incorporates prior information. In Chapter 4, we have developed models for EOL outputs without using prior information. These models are less accurate than the ones that use prior information and data.

2.5 Integration of engineering models, empirical information and data through MARS

Multivariate adaptive regression splines (MARS) were chosen as part of step seven of the hybrid methodology. Currently, MARS uses data only. This research has extended the use of MARS to also incorporate prior information in the form of physics-based engineering models and empirical information about the manufacturing process and product.

This section aims to describe the extension of MARS to incorporate prior information with data, and develops on sections 1.3.2.2.2 and 1.3.2.2.3 which present the process of model development by MARS with the use of data only.

This thesis research has identified five places where MARS can use prior information. These five places are:

1. split variables in split nodes
2. definition of split criteria in split nodes
3. choice of basis functions

4. choice of the order of basis functions
5. definition of pruning criteria

We have explored and developed the first option of using prior information as split variables in split nodes. Other options are suggested as part of future work in Section 5.3.1.

Our approach treats each new piece of prior information as a new variable. Each engineering and statistical model results in a new variable. Consequently, there are as many new variables as the number of new pieces of prior information. Appendix A presents a numerical example to illustrate the process of creating new variables and of developing MARS models. The numerical values for the new variables (derived using prior information) are determined from those of the predictor variables. Predictor variables are the potentially influential process parameters identified using the first four steps of our methodology.

Each row of observations for the predictor variables creates a new row of observations for the new variables. The number of rows of predictor variables is the same as that of new variables. The matrix of numerical estimates of new variables is appended row-wise to the matrix of observations of predictor variables. The number of columns in the augmented data matrix is the sum of the number of columns in the old matrix of predictor variables and the number of pieces of prior information. The number of rows in the augmented data matrix is the same as that in the old matrix of predictor variables. The following example illustrates the process of creating new variables. Appendix A develops this example using numerical data.

Assume that the old matrix of predictor variables comprises two columns and fifty rows. Each column represents a predictor variable (or a potentially influential process variable identified using the first four steps of out methodology). The rows represent fifty observation on the two predictor variables (X_1 and X_2). Also assume an engineering model (through Equation 7, and a piece of prior information from empirical understanding (shown as Equation 8). X_1 and X_2 in Equation 7 and in Equation 8 are predictor variables. Y_1 and Y_2 represent product or process characteristics. Other symbols in Equation 7 and in Equation 8 are known coefficients, representing material properties or other physical constants.

$$Y_1 = a_1 \frac{X_1}{X_2} + a_2$$

Equation 7. Prior information as an engineering model

$$Y_2 = a_3X_1 + a_4X_2$$

Equation 8. Prior information as empirical information

To create a numerical values for Y_1, the values of X_1 and X_2 are substituted from the old matrix of predictor variables in Equation 7 to give fifty numerical values for Y_1. A similar substitution process creates fifty numerical values for Y_2. Y_1 and Y_2 are two new variables, one from each piece of prior information.

The numerical values of Y_1 and Y_2 are appended row-wise to the old matrix of predictor variables. The resulting augmented matrix has four columns and fifty rows. The columns represent original predictor variables X_1 and X_2, and the new variables Y_1 and Y_2. The rows represent different observations.

The new inputs to the MARS modeling tool (software) are a column of response data and the augmented data matrix, which holds old and new variables. The software for MARS model development was written by Hastie [28] in the statistical software package environment called Splus.

During classification (or partitioning) at any split node, MARS now has the option to split on a new variable (determined from prior information) or on an old predictor variable. Obviously, the split variable of choice is one that best satisfies the split criteria. This thesis research used reduction in variance as the split criteria. (Prior information can also be used to determine/choose an appropriate split criteria, described as the second option of the five options that identify the places where MARS can incorporate prior information).

MARS will continue to develop further splits, choosing at each split any of the variables from the augmented data matrix. MARS will result in several basis functions at the end of forward step-wise regression. These functions will assume one of the following forms (or higher order interaction terms):

- $[X_1 - X_{1cut}]_+$
- $[X_2 - X_{2cut}]_+$
- $[X_3 - X_{3cut}]_+[X_4 - X_{4cut}]_+$

X_1, X_2, X_3 and X_4 are split variables. X_{1cut}, X_{2cut}, X_{3cut} and X_{4cut} are their respective split values or knot locations. The split variables can be old predictor variables or new variables or a combination of both. The first two expressions represent main effects of X_1 and X_2. The last expression contains two variables X_3 and X_4, and represents a two-way interaction between

both variables. If there are higher order interactions, the MARS model will result in more basis functions with appropriate number of square bracket terms. Each square bracket term contains only one split variable. The terms in the square brackets are interpreted in the following way:

- $[x\text{-}x_{cut}]_+ = (x\text{-}x_{cut})$ for $x > x_{cut}$, 0 otherwise
- $[x\text{-}x_{cut}]_- = (x\text{-}x_{cut})$ for $x < x_{cut}$, 0 otherwise

The final MARS model is a linear combination of several functions generated earlier. A numerical coefficient multiplies with each basis function to modulate its contribution to the response. A typical MARS model is represented by Equation 9.

$$\text{Output} = a_0 + a_5[X_1 - X_{1cut}]_+ + a_6[X_2 - X_{2cut}]_-[Y_1 - Y_{1cut}]_+ + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

Equation 9. A typical MARS model

In Equation 9,

- X_1 and X_2 are old predictor variables with their respective know locations in the square brackets
- Y_1 is a new variable from prior information with Y_{1cut} being its knot location
- a_i's are constants
- ε is random noise

By substituting Equation 7 in Equation 9, we get Equation 10.

$$\text{Output} = a_0 + a_5[X_1 - X_{1cut}]_+ + a_6[X_2 - X_{2cut}]_-[a_1\frac{X_1}{X_2} - a_1\frac{X_{1cut1}}{X_{2cut2}}]_+ + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

Equation 10. MARS model: Old predictor variables substitute prior information

The variables in Equation 9 are the same as those in Equation 10, except X_{1cut1} and X_{2cut2} which help determine the knot location for X_1/X_2 after the substitution of Equation 7 in Equation 9 for Y_1.

Equation 10 only contains old predictor variables. However, note that it also contains prior information from Equation 8. Without the use of prior information, MARS would have perhaps tried to approximate the relationship provided by prior information by using simple univariate expressions. The approximation would be less accurate than the model developed using prior information. This thesis research later shows that the incorporation of prior

information results in more accurate models, see Chapter 4. Section 5.1.3 provides a graphical interpretation of a MARS model which uses data (or old predictor variables) only. It also compares that interpretation to that of a MARS model which uses data and prior information. Section 4.4 provides an intuitive understanding of the reasons why MARS models with prior information and data give better results than the ones developed using data (or old predictor variables) only.

This thesis research has used MARS as part of step seven of the methodology. (The reasons for choosing MARS are described in Section 1.3.2.2). However, our hybrid methodology is general to include and use several other modeling tools like NN, PLS, PCR, etc. In fact, the current thesis research has also developed models using PCR, and has compared them with the ones developed using MARS.

Before presenting the PCR and MARS models in Chapter 4, the next chapter, Chapter 3, describes an industrial application problem of modeling end-of-line (EOL) channel length of MOS transistors. Besides discussing general modeling difficulties, Chapter 3 also applies the seven step hybrid methodology to the problem of modeling channel length.

3. Semiconductor manufacturing application: Modeling end-of-line channel length of MOS transistors

This chapter introduces an industrial problem in semiconductor manufacturing. Specifically, the problem is that of modeling end-of-line (EOL) channel length in metal oxide semiconductor (MOS) transistors as a function of in-line process parameters. Channel length is an important factor in determining speed of a microprocessor, density of devices on a chip, and memory capacity of memory chips. Our methodology, described in Chapter 2, is applied to this industrial problem.

This chapter has six sections. Section 3.1 describes a semiconductor manufacturing process in general, and depicts a semiconductor manufacturing line. Section 3.2 explains channel length of MOS transistors, and depicts it graphically. The importance of channel length in product performance is described in Section 3.3. Section 3.4 reviews practical (and theoretical) difficulties in modeling channel length. The steps of our hybrid methodology are applied to the problem of modeling channel length in Section 3.5. This chapter concludes with a summary in Section 3.6.

3.1 The semiconductor manufacturing process

Semiconductor manufacturing is a multi-step process, consisting of a few hundred process steps, as summarized in Figure 17.

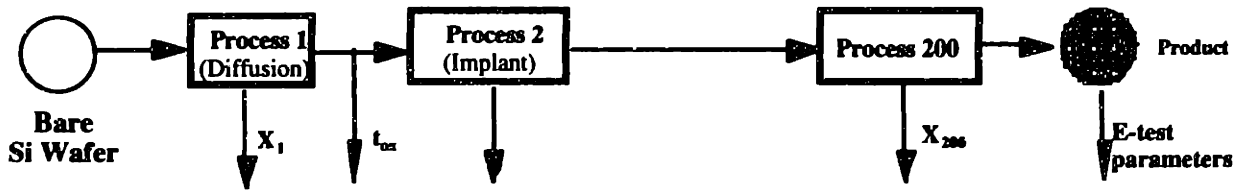

Figure 17. A semiconductor manufacturing line: A multi-step process

The starting raw material for a semiconductor manufacturing line is a bare silicon wafer. Four types of operations are performed several times on the wafer to produce final chips. These operations include etching, diffusion, lithography and thin film deposition. The operations result in several chips (or dice) on each wafer, as illustrated in Figure 18. Figure 18 also shows that several wafers travel together in a boat or in a lot. This marks the end of wafer processing [25, 88]. Wafer processing usually takes several weeks. Finished wafers are tested for about 100 electrical characteristics. Electrical tests or E-tests help determine if the several manufacturing process steps worked correctly. The chips are then sorted (or binned by doing probe tests), and sawed out of wafers. Acceptable chips are packaged and shipped to customers.

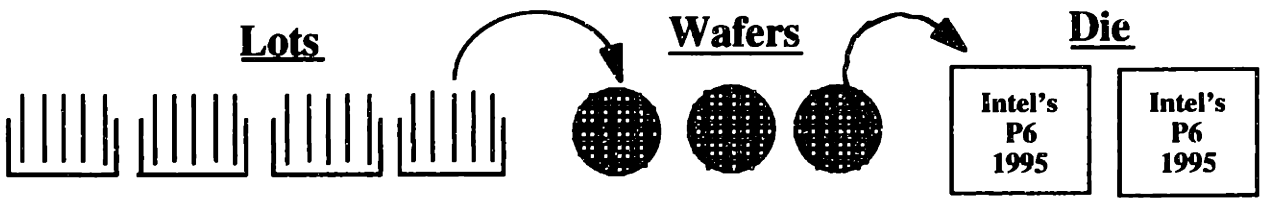

Figure 18. Chips nested in wafers nested in lots

During wafer processing, data are gathered for process variables at each process step, and for intermediate product measurements at the end of many process steps. However, data for some of those process variables and intermediate product measurements are not stored permanently. (Hence, these data are not available later for modeling). Data are also collected for E-tests at the end of wafer processing, and before binning.

This thesis research has focused on modeling two such E-test characteristics as a function of influential process variables and intermediate product measurements. These E-test characteristics are channel-length measurements for n-channel transistors (L_{en}) and for p-channel transistors (L_{ep}), see Figure 17. Section 3.2 explains channel length, and differentiates L_{en} from L_{ep}. Since L_{en} and L_{ep} are measured at the end of the line for wafer processing, they are called

end-of-line (EOL) channel length measurements. This thesis research has used data from an Intel Corporation wafer processing line, on which a family of microprocessors are fabricated.

3.2 What is end-of-line channel length?

Figure 19 shows a metal oxide semiconductor (MOS) transistor [80]. A typical microprocessor has millions of such transistors. (Intel's Pentium Pro™ has over 5.1 million transistors. Pentium Pro is the trademark of Intel Corporation.) The heavily doped source provides carriers (electrons or holes), which travel to the heavily doped drain through the more lightly doped silicon **channel region** under the oxide. The electrode above the oxide consists of heavily doped polycrystalline silicon, and is called the **gate**. (L_{drawn} is the length of the channel patterned at a lithography step, and is an important parameter influencing the effective channel length.)

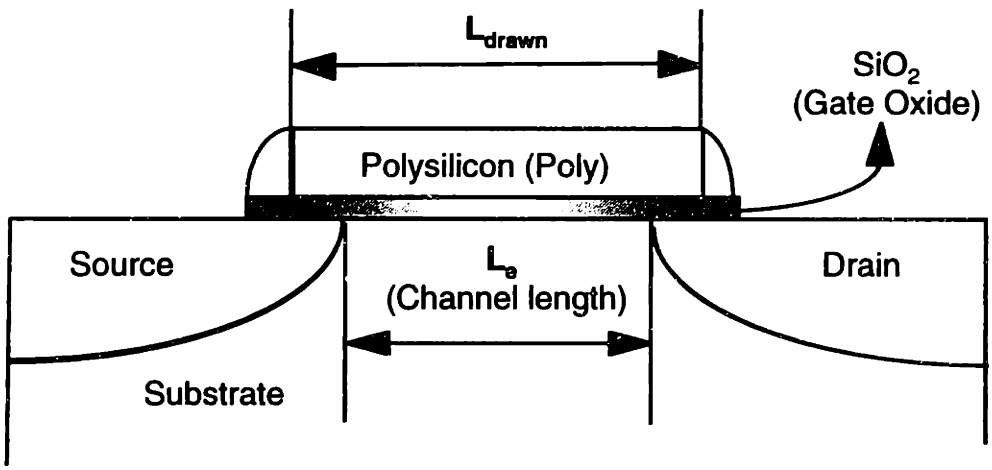

Figure 19. A metal oxide semiconductor (MOS) transistor

The channel region is non-conducting at room temperature with zero bias voltage on the gate terminal. When an appropriate voltage is applied at the gate, a large number of minority carriers move to the channel region (close to the oxide) from the rest of the substrate. When enough carriers exist to form a high conducting current path in the channel, **inversion** is said to occur and the device is **on**. A shorter channel will invert faster because it needs fewer total number of carriers from the substrate to achieve the critical carrier concentration (that is to say, smaller devices have smaller gate capacitances). The gate voltage needed to achieve inversion is called the **threshold voltage**. After inversion in the channel is achieved, an appropriate voltage difference between the source and the drain results in the free flow of current between these electrodes.

MOS transistors are of two types, n-channel transistors and p-channel transistors. The geometric representation of the two is the same. The difference between the two is in the type of dopants used in the fabrication of the source, drain, substrate and gate; the type of dopant determines whether electrons or holes will be the carriers in the channel. The substrate of an n-channel transistor is lightly doped with acceptor atoms such as boron, gallium, and indium, while the source/drain areas and the poly-silicon gate of an n-channel transistor are heavily doped with donor atoms such as phosphorous, arsenic, and antimony. Electrons act as carriers in the channel region of n-channel transistors. The substrate of a p-channel transistor, on the other hand, is lightly doped with donor atoms such as phosphorous, arsenic, and antimony; its source/drain areas and the poly-silicon gate are heavily doped with acceptor atoms such as boron, gallium, and indium. Holes act as carriers in the channel of p-channel transistors. The channel length of n-channel transistors is represented as L_{cn}, and that of p-channel transistors as L_{cp}.

For several reasons (including process variation, violation of the charge-sheet model, presence of edge effects, etc.), the important length parameter for the channel is not the geometric length of the poly gate (L_{drawn} in Figure 19). Rather, the key parameter is the **electrical** or **effective channel length**. In this thesis research, the term "channel length" always refers to effective channel length.

3.3 Why is end-of-line channel length important?

In Figure 19, a short distance between the source and the drain results in a short channel length. The transistor will turn on earlier due to faster channel inversion, and switching between high and low output (drain) voltages will be faster due to a shorter transit time of carriers in the channel. The resulting increase in microprocessor speed has been a major motivation for continual scaling of devices to smaller and smaller dimensions. In addition, smaller line widths result in higher packing density on a chip. Higher transistor counts have enabled microprocessors to handle more tasks and more complexity (such as multi-tasking) more efficiently [79]. Microprocessor capability and memory size have grown phenomenally over the past few decades, due in large part to the increase in device density according to Moore's law. (Moore's law states that the number of transistors on a given chip area double every eighteen months.)

However, if the channel is too short, carriers may move to the channel from the source, and may travel to the drain even before the threshold voltage is applied at the gate. This phenomenon is called transistor breakdown, which deteriorates the reliability of a

microprocessor or a memory chip. The careful design and control of channel length is thus critical to achieve both high speed and high reliability. Channel length is one of the most important determiners of product quality to Intel and to the semiconductor industry in general, and accurate models can have a tremendous payoff in manufacturing control.

3.4 Difficulties in modeling EOL channel length

Modeling end-of-line (EOL) channel length as a function of in-line process parameters is a challenging task. The difficulties associated with modeling EOL channel length are of broadly two types

1. **data-related** difficulties, which need a more practical approach to be addressed.
2. **model-development related** difficulties, which are more theoretical by nature.

3.4.1 Data-related difficulties

Wafer fabrication is a data rich environment characterized by data collection at multiple steps during processing, hierarchical sampling plans, multivariate data collected end of line, and both continuous and discrete response variables. Hierarchical sampling plans and nesting in variables create complex model structures between several EOL E-test variables. This complexity often makes even understanding the components of variance difficult.

It is important to know what the processing conditions were for a wafer, for which channel length measurements are being currently taken. This exact linking of in-line processing conditions with specific measurements at the end of the line is important to understand the influence of different processing conditions on the EOL channel length. Unfortunately, a clear linkage is often missing, increasing the challenge in modeling because these links have to be somehow established first, before any data-modeling activity is begun.

A further complication to unclear linkage is the presence of data on many different data bases. For several reasons (including convenience, feasibility, and/or cost), data from different process steps reside in many locations. Often, data exist only in dedicated station controllers, which are typically not connected to a central computer or do not transfer data to a central system. Moreover, station controllers have limited storage capacity, and data-acquisition systems typically collect data at high sampling rates (band width). Consequently, data stored in station controllers are usually for a few cycles of a process step. These data may not represent enough products, thus resulting in poor confidence in the parameters of the final model. In addition, station controllers for different machines (in different process steps of a manufacturing line) may have dissimilar data storage formats. This adds difficulty in setting up a single data

file for analyzing data from several process steps of a manufacturing line. The problem of poor data linkage and the presence of many disconnected data bases (which follow different storage formats) result in a lower level of monitor integration than desired.

Besides the issues with differing formats and many data bases, data contained in them do not lend themselves to straight forward statistical analysis because of the presence of outliers and missing data. Missing data can be due to machine error, sampling plans, operator error, data-base error, etc. Data from real systems also have outliers. Outliers can significantly influence several statistical analysis results. A typical challenge with the presence of outliers is to differentiate a genuine extreme data value from a false extreme data value. The genuine extreme data value is useful to understand system behavior. On the other hand, false extreme data values can give misleading results. Thus, a challenge of preliminary data analysis in the presence of outliers is to classify the outliers correctly.

Even if the outliers have been categorized correctly, it is important to analyze the genuine extreme values appropriately. If the system behaves in an extreme way locally, an extreme value should not affect our analysis about the system globally. The presence of outliers and missing values also need a rationale for choosing between the deletion of the whole row of observations (which has missing values or outliers), or the substitution of outliers and missing values by appropriate estimates [34]. Regardless, missing values and outliers add complexity to data handling, analysis, and interpretation of final results.

3.4.2 Model development related difficulties

Besides the practical data-related difficulties, several model-development issues also arise. The existence of several process variables and our lack of full understanding of their effect on final quality add complexity to the entire production process. Furthermore, even set points of process variables can vary in multiple runs. For example, the flow rate of a gas in diffusion step 3 can be different from that in step 5. The effect of process parameters on the final quality of the chip (clock speed, yield, etc.) is impossible to understand using only engineering models because such models do not readily represent process noise or subtle environmental and multi-parameter effects. Traditional parametric techniques fail to capture desirable local deviations in the relationships between process parameters and output quality. Using data, one can develop purely statistical models relatively easily. However, data-driven models do not incorporate causality considerations, which are necessary for process control and process improvement.

Sampling plans and the choice of variables on which data are not collected can significantly influence the accuracy of the final model. If no process variable data are collected for an influential process step, and no data are collected for intermediate product variables after the process step, then the final model may not be as accurate as it could be if the relevant data were collected and used in developing the model.

If numerical techniques of simulation are used for model development, then they could include causality connection. However, numerical techniques often fail to explicitly represent many transistor characteristics [80], such as the threshold voltage (V_t). Such characteristics include the voltage to bring the channel in depletion, in weak inversion, and in medium inversion, etc. An explicit representation of such characteristics is necessary later for accurate circuit simulation [80].

Such difficulties make modeling of EOL channel length more challenging. In this thesis research, a methodology has been developed to model channel length by incorporating well-known device-physics models (which bring causality connection), empirical information from process engineers, and data. The next section applies the different steps of our hybrid methodology to the problem of modeling EOL channel length.

3.5 Seven-step hybrid methodology applied to model EOL channel length

This thesis research has modeled EOL channel length at lot level, see Figure 18. This means that one representative number was used, the lot-level median value, as the measure of channel length for a lot. The numerical value of the representative number for L_{en} was typically different from that for L_{ep}. Modeling the variance of channel length is recommended as part of future work. The median value determined from channel-length measurements at multiple sites on every wafer in a lot was treated as the lot level median value for the lot. There were several reasons for modeling channel length at lot level. These reasons include:

1. availability of data for process variables and most intermediate product measurements at lot-level only. These data were not available at wafer level or at die level.
2. the desire to determine the highest level of data aggregation (or representation), that provided the desirable signal. In high volume manufacturing a reduction in time for inspection and measurement is important for high throughput. Data aggregated at a high level (such as at lot-level) need less time for inspection than data aggregated at lower levels (such as at wafer-level or at die-level), and is preferable in high-volume manufacturing. In addition, the higher the level of data aggregation, the less the total

data there is to analyze for a given number of products (or of process cycles). This can help in obtaining inference faster, and in implementing real-time control. Low levels of data aggregation usually result in large volume of data. Besides needing a large upfront capital investment in data-acquisition equipment, a large amount of data takes more analysis time, needs more storage space, and minimizes opportunities for real-time control.

Thus, a high level of data aggregation is preferred over lower levels of aggregation.

This section is divided into two parts, based on the two parts (functions) of our hybrid methodology, as illustrated by Figure 12. Section 3.5.1 discusses the first part. It focuses on the identification of influential process parameters for EOL channel length. In this Section, steps one through four of our hybrid methodology are applied to the problem of modeling channel length. Section 3.5.2 discusses the second part of the methodology. It focuses on the development of a large-scale model relating EOL channel length to influential process parameters (identified in the previous Section). In this Section, steps five through seven of our hybrid methodology are applied to the problem of modeling channel length.

In sections 3.5.1 and 3.5.2, several tables and figures report intermediate results for both, L_{en} and L_{ep}, together. To be sure, the steps of the methodology should be, and were, applied separately to model L_{en} and L_{ep}. Since L_{en} and L_{ep} are very similar, we have chosen to present the intermediate results for the two together. Where ever possible, the tables and figures clarify important differences.

3.5.1 Identification of influential process parameters

Step 1: Choose relevant EOL E-tests and identify model relationship

Table 1shows a list of all EOL E-tests. Table 1 includes tests that have direct connection with channel length (e.g., threshold voltages and critical dimension of poly lines) and also those that cannot possibly have a connection with channel length (e.g., critical dimensions for metal inter-connects). Repeating names represent the same E-test, but at different bias conditions or measurements on different test structures.

sheet resistance of wide n+ diffusion	junction leakage of p+ area	junction leakage of p+ edge
scaled resistance of un-nested n+ diffusion lines	drain current of n channel in saturation regime	n+ spiking junction breakdown voltage
sheet resistance of wide p+ diffusion	drain current of p channel in saturation regime	p+ spiking junction breakdown voltage
scaled resistance of un-nested p+ diffusion	drain current of n channel in saturation regime	n+ spiking junction leakage
ecd of un-nested n+ diffusion lines	drain current of p channel in saturation regime	n+ spiking junction leakage
ecd of un-nested p+ diffusion lines	drain current of n channel in saturation regime	threshold voltage of n channel
sheet resistance of n well	drain current of of p channel in saturation regime	threshold voltage of p channel
n well leakage current	drain current of n channel in saturation regime	threshold voltage of p channel
salicide resistance of n+ source/drain	drain current of p channel in saturation regime	threshold voltage of n channel
sheet resistance of n+ implanted wide poly	drain current of n channel in saturation regime	threshold voltage of p channel
sheet resistance of n+ implanted narrow un-nested	drain current of p channel in saturation regime	threshold voltage of n channel
resistance of n+ implanted poly	transconductance of n channel	threshold voltage of p channel
resistance of n+ implanted poly on fox	transconductance of p channel	threshold voltage of n channel
resistance of n+ implanted poly per p+ diffusion	transconductance of n channel	threshold voltage of n channel
ecd of n+ implanted narrow un-nested poly	transconductance of p channel	threshold voltage of p channel
sheet resistance of wide m1 on fox	transconductance of n channel	threshold voltage of n channel
scaled resistance of narrow m1 on fox	transconductance of of p channel	threshold voltage of p channel
scaled resistance of narrow m1 on poly on fox	transconductance of n channel	threshold voltage with back biasing
ecd of narrow m1 on fox	transconductance of p channel	threshold voltage with back biasing
ecd of narrow m1 on poly on fox nested high	transconductance of n channel	drain current of n channel
sheet resistance of wide m2 on fox	transconductance of p channel	drain current of p channel
scaled resistance of narrow m2 lines isolated on fox	transconductance of n channel	drain current of n channel
scaled resistance of narrow m2 lines nested on fox	transconductance of n channel	drain current of p channel
ecd of narrow m2 lines isolated on fox	transconductance of p channel	drain current of n channel
ecd of narrow m2 lines nested on fox	punchthrough of n channel	drain current of of p channel
sheet resistance of wide m3 on fox	punchthrough of p channel	drain current of n channel
scaled resistance of narrow m3 lines isolated on fox	punchthrough of n channel	drain current of p channel
scaled resistance of narrow m3 lines nested on fox	punchthrough of p channel	drain current of n channel
ecd of narrow m3 lines isolated on fox	punchthrough of n channel	drain current of p channel
ecd of narrow m3 lines nested on fox	punchthrough of of p channel	drain current of p channel
m1-m1 combs/topography shorting structure	punchthrough of n channel	external resistance of n channel
m2-m2 combs/topography shorting structure	punchthrough of p channel	corrected vt of n channel
m3-m3 combs/topography shorting structure	punchthrough of n channel	mobility degradation factor of n channel
m1-m2-m3 capacitor shorting structure	punchthrough of p channel	uocoxze of n channel
resistance of n+ contact chain	punchthrough of n channel	effective electrical length of p channel
resistance of p+ contact chain	punchthrough of p channel	external resistance of p channel
resistance of poly contact chain	well current of n channel	corrected vt of p channel
resistance of via1 chain	well current of p channel	mobility degradation factor of p channel
resistance of via2 chain	well current of n channel	effective electrical length of n channel
accumulation capacitance of area oxide over p well	well current of of p channel	external resistance of narrow source/drain
inversion capacitance of area oxide over p well	effective electrical length of n channel	corrected vt of n channel narrow source/drain
accumulation capacitance of area oxide over n well	external resistance of n channel	mobility degradation factor of n channel
inversion capacitance of area oxide over n well	corrected vt of n channel	uocoxze n channel narrow source/drain
accumulation capacitance of area oxide over n well	mobility degradation factor of n channel	process bias for channel of nmos
thickness of area oxide over p well	uocoxze of n channel	process bias for channel of pmos
thickness of area oxide over n well	effective electrical length of p channel	le difference between n and p channel
thickness of area oxide over n well	external resistance of p channel	effective width of n channel
breakdown voltage of area oxide over p well	corrected vt of p channel	effective width of p channel
breakdown voltage of area oxide over n well	mobility degradation factor of p channel	process bias for width of nmos
capacitance of n+ area	uocoxze of p channel	process bias for width of pmos
capacitance of n+ edge	effective electrical length of n channel	n+ to n+ punchthrough voltage
capacitance of p+ area	external resistance of n channel	p+ to p+ punchthrough voltage
capacitance of p+ edge	corrected vt of n channel	n+ to n well punchthrough voltage
junction breakdown voltage of n+ area	mobility degradation factor of n channel	n+ to p well punchthrough voltage
junction breakdown voltage of n+ edge	uocoxze of n channel	threshold voltage of n channel field transistor
junction breakdown voltage of p+ area	effective electrical length of p channel	threshold voltage of n channel field transistor
junction breakdown voltage of p+ edge	corrected vt of p channel	n+ contact to gate breakdown voltage
junction leakage of n+ area	mobility degradation factor of p channel	p+ contact to gate breakdown voltage
junction leakage of n+ edge	uocoxze of p channel	n+ junction breakdown voltage
drain current of n channel with gate structure in saturation regime	effective electrical length of n channel	p+ junction breakdown voltage

Table 1. List of all EOL E-tests

From the list of all E-tests, only a few E-tests were selected that could possibly have a connection with EOL channel length (L_{en} and L_{ep}), and are compiled in Table 2. This selection was guided mainly by physics-based models and understanding (and very little by past experience of company personnel).

sheet resistance of wide n+ diffusion	well current of p channel	drain current of n channel
sheet resistance of wide p+ diffusion	well current of n channel	drain current of p channel
n well leakage current	well current of of p channel	drain current of n channel
ecd of n+ implanted narrow un-nested poly	effective electrical length of n channel	drain current of p channel
breakdown voltage of area oxide over p well	external resistance of n channel	external resistance of n channel
breakdown voltage of area oxide over n well	uocoxze of n channel	uocoxze of n channel
capacitance of n+ edge	effective electrical length of p channel	effective electrical length of p channel
junction breakdown voltage of p+ edge	external resistance of p channel	external resistance of p channel
drain current of n channel with gate structure in saturation regime	uocoxze of p channel	effective electrical length of n channel
drain current of n channel in saturation regime	effective electrical length of n channel	external resistance of narrow source/drain
drain current of p channel in saturation regime	external resistance of n channel	uocoxze n channel narrow source/drain
drain current of n channel in saturation regime	uocoxze of n channel	process bias for channel of nmos
drain current of p channel in saturation regime	effective electrical length of p channel	process bias for channel of pmos
transconductance of n channel	uocoxze of p channel	le difference between n and p channel
transconductance of p channel	effective electrical length of n channel	effective width of n channel
punchthrough of n channel	junction leakage of p+ edge	effective width of p channel
punchthrough of p channel	threshold voltage of n channel	n+ to n+ punchthrough voltage
punchthrough of n channel	threshold voltage of p channel	p+ to p+ punchthrough voltage
punchthrough of p channel	threshold voltage of p channel	n+ to n well punchthrough voltage
punchthrough of n channel	threshold voltage of n channel	n+ to p well punchthrough voltage
punchthrough of p channel	threshold voltage of p channel	n+ junction breakdown voltage
well current of n channel	threshold voltage of n channel	p+ junction breakdown voltage

Table 2. EOL E-tests selected by step one

Since L_{en} and L_{ep} are modeled at lot-level, no inherent model structure was used. For modeling at wafer-level, the nesting structure of wafers being nested in lots should be used. Here, the median value of L_{en} (and L_{ep}) is modeled as a function of the median values of other E-tests selected from Table 1.

Step 2: Collect EOL inspection data

Data were collected for the chosen EOL E-tests. 385 lot level median values were collected for each E-test. These lots were processed over two quarters (or six months). Each row in the resulting data file represented E-test values for a single lot.

Step 3: Develop model & identify secondary EOL E-tests

Several statistical methods exist in literature to develop a model for the purpose of identifying statistically important predictor variables [40, 52, 53, 54]. Here, two different models were developed for L_{en}, and for L_{ep}. They were principal components regression (PCR) models and classification and regression trees (CART) models [11, 49]. These models related L_{en} (and L_{ep}) to other EOL E-tests parameters, chosen from Table 1. From each of those models, we identified a list of significant EOL E-tests that predicted EOL L_{en} (and L_{ep}) accurately [40, 68, 69, 70, 71]. These significant EOL E-tests are also called here secondary EOL E-tests. For both L_{en} and L_{ep}, the secondary E-test variables for the PCR model matched very well with those from the CART model. Table 3 shows the secondary E-test variables for L_{en} and L_{ep}. Repeating names represent the same E-test, but at different bias conditions or measurements on different test structures.

Secondary EOL E-tests for L_{en} & L_{ep}
ecd of n+ implanted narrow un-nested poly
well current of n channel
well current of p channel
drain current of p channel in saturation regime
drain current of p channel in saturation regime
breakdown voltage of area oxide over p well
breakdown voltage of area oxide over n well
threshold voltage of p channel
threshold voltage of p channel
drain current of n channel in saturation regime
drain current of n channel in saturation regime
punchthrough of p channel
punchthrough of p channel
punchthrough of p channel
ecd of un-nested n+ diffusion lines
punchthrough of n channel
punchthrough of n channel
punchthrough of n channel
junction leakage of n+ edge
junction leakage of p+ edge
le difference between n and p channel
threshold voltage of n channel
threshold voltage of n channel
threshold voltage of n channel
n well leakage current
external resistance of n channel
effective width of n channel
effective width of p channel

Table 3. Secondary E-test variables for L_{en} and L_{ep}

The secondary E-tests in Table 3 are of the following types:

- ECD's (electrical critical dimensions), that are related to the geometry of the poly gate.
- Threshold voltages, drain currents, breakdown voltages of the area oxide related to the gate oxide thickness that determine the degree of control of gate voltage on the channel.
- Well currents, punchthrough voltages, leakage currents and external resistance that are related strongly to the channel and source/drain doping.

Step 4: Identify process steps influencing secondary EOL E-tests and EOL channel length

The use of the following sources of information helped identify influential process steps for secondary EOL E-tests and for EOL channel length:

1. **Theoretical and empirical models** from process physics and device physics to determine what process steps influence secondary EOL E-tests, L_{en} and L_{ep}. An example of a relevant theoretical model for threshold voltage (V_t) is given by Equation 6 on page 47. In addition, we use Intel Corporation's process flow handbook, E-Test handbook and FMEA for more information. (FMEA stands for Failure Modes and Effects Analysis. This document lists possible types of faults in the final product, their possible causes, severity to customers, how often they estimated to occur and how often the faults are estimated to be caught.) Other examples of physics-based models can be found in [6, 7, 25, 32, 35, 40, 43, 44, 64, 65, 73, 74, 88].
2. **Numerical simulation tools** such as Intel Corporation's numerical simulation tools that build on process simulators such as Suprem and device simulators such as Pisces. Through process and device simulation, these tools identify influential process parameters for a given EOL E-test parameter at different operating conditions.
3. **Knowledge bases:** Intel Corporation had two different sets of knowledge bases for the current technology. These knowledge bases contained information on what process variables were identified as the causes of problems in the past. These knowledge bases helped identify influential process steps for L_{en} and L_{ep} and for secondary E-test variables.
4. **Experience of personnel**: Given list of secondary EOL E-tests (and L_{en} and L_{ep}), integration engineers at the manufacturing site were asked to identify process steps that most influence each of the secondary EOL E-tests, L_{en} and L_{ep}.

Design of experiments (DOE) or **natural experiments** were not used in the first four steps of the methodology to demonstrate that influential process steps can be identified even without the use of process data.

The influential process steps obtained from the different sources mentioned above are reported in Table 4. The process parameters associated with the process steps are potentially influential process parameters for L_{en} and L_{ep}. There were sixty nine such process parameters.

Activates dopants implanted in poly & S/D	Gate anneal	Plasma etch of plugs in via1
Anneals wafers in hydrogen ambient	Grows buffer oxide	Plasma etch of plugs in via2
Defines active and isolation regions	Grows edge oxide	Plasma etch of via1
Defines polysilicon gates & interconnects	Grows field oxide	Plasma etch sputter 1
Densifies BPSG layer	Grows gate oxide	Plasma etch sputter 2
Deposit base material for Salicide formation	Grows oxide to protect gate edges	Plasma etch sputter 3
Deposit high quality thermal oxide	Grows sacrificial oxide	Plasma etch through BPSG for contacts
Deposit interlevel dielectric between devices & metal 1	N+ S/D implant	Plasma etch to open bond pads
Deposit spacer	NTip implant	Preclean to remove remaining buffer oxide
Deposits a conformal polysilicon film	Nwell implant	Precleans gate
Deposits dielectric between metal 1 & 2	P+ S/D implant	PTip implant
Deposits hermetic conformal passivation layer	Plasma etch for deposition of ILD	Pwell implant
Drives n-well dopant into the epi-layer	Plasma etch for metal 1 interconnects	Remove silicon nitride from active regions
Drives p-well dopant into epi layer	Plasma etch for metal 2 interconnects	Reoxidizes top surface of poly & S/D
Etches away silicon nitride & buffer oxide	Plasma etch for metal 3 interconnects	Salicide formation
Etches polysilicon gates & interconnects	Plasma etch for via2	Sputter
Etches spacer	Plasma etch of plugs connecting poly or S/D to metal 1	Start material condition

Table 4. Influential process steps for L_{en} and L_{ep}

The next Section focuses on the final model development, which relates L_{en} (or L_{ep}) to potentially influential process parameters.

3.5.2 Development of large-scale model combining engineering and statistical information, and data

This Section applies steps five through seven of our hybrid methodology to the problem of modeling EOL channel length.

Step 5: Collect data for in-line process variables and intermediate product variables

Data were collected for process parameters associated with the influential process steps reported in Table 4. There were a total of sixty nine such parameters. 242 lot level median values were collected for each influential process parameter and for EOL L_{en} and L_{ep}. Each row in the data file represented measurements for the same lot.

Step 6: Determine engineering models and empirical information

Several physics-based models were collected. They are reported as six equations and one expression (feature) in Table 5 [80]. The six equations explain the behavior of an MOS transistor under different operating conditions. For example, Equation 1 in Table 5 is a general equation for threshold voltage for long channel devices. Equations 2 through 4 in Table 5 model short channel effects. Equation 5 models the mobility of carrier. Carrier mobility is important because it explains how fast carriers move from the source to the drain. Equation 6 models miller-capacitance. The last expression is a feature that is important in the calculation of channel

length. (Other sources of such piece-wise models are [6, 7, 25, 32, 35, 40, 43, 44, 64, 65, 73, 74, 85, 88].)

Equation or Expression number	Name	Equation/Expression
Equation 1	Threshold voltage (V_t)	$V_t = \phi_{MS} - q_{ox}/C_{ox} + \phi_B + \gamma(\phi_B + V_{sb})^{0.5}$
Equation 2	δV_t due to pinch off	$\delta V_{tp} = 2(\varepsilon_s/\varepsilon_{ox})(t_{ox}/L)(\phi_B + V_{sb})$
Equation 3	δV_t due to narrow channel	$\delta V_{tc} = \pi(\varepsilon_s/\varepsilon_{ox})(t_{ox}/W)(\phi_B + V_{sb})$
Equation 4	Change in channel length due to pinch off	$\delta L = \text{sqrt}(2(\varepsilon_s/(qN_A)))(\text{sqrt}(\phi_D + (V_{ds}-V_{ds})) - \text{sqrt}(\phi_D))$
Equation 5	Effective carrier mobility	$\mu_{eff} = (\text{Bulk mobility}/2)/(1 + (0.025C_{ox}/2\varepsilon_s)(V_{gs} - V_t + 2\gamma\,\text{sqrt}(\phi_B + V_{sb}) - 0.5(1 - \delta)V_{ds}))$
Equation 6	Overlap/Miller Capacitance	$C_{over} = \varepsilon_{ox}\pi/r$
Expression 1	MuCZe	$\mu C_{ox} W$

Table 5. Physics-based models

In the equations and the expression of Table 5,

- ϕ_{MS} is the difference between the bulk potential and the gate potential
- q_{ox} is the charge on the gate per unit area
- C_{ox} is the capacitance of the gate oxide per unit area
- ϕ_B is the surface potential at the interface of the oxide and the substrate
- γ is the body-effect factor, and includes a measure of the implant dose in the substrate
- V_{sb} is the potential difference between the source and the substrate
- ε_s is the permittivity of single crystal silicon
- ε_{ox} is the permittivity of the gate oxide
- t_{ox} is the thickness of the gate oxide
- L is the nominal length of channel
- W is the nominal width of channel
- q is the magnitude of electronic charge

- N_A is the substrate dose
- V_{DS} is the potential difference across the source and the drain
- V_{DS}' is the potential difference across the source and the pinch-off point in the channel
- V_{GS} is the potential difference across the gate and the source
- $\mu = \mu_{eff}$ = effective mobility of the carriers
- d is an approximation constant
- r is the depth of LDD implant

$$\phi_D = \frac{\varepsilon_s E_l^2}{2qN_A}$$

- E_l is the electric field that results in velocity saturation of electrons in the channel

The right side of equations 1 through 6 and the expression in Table 5 consist of

- in-line process parameters,
- in-line intermediate product variables,
- material properties.

Numerical values of relevant material properties were determined from standard tables. Most in-line process parameters and in-line intermediate product variables were part of the list of the sixty nine process variables for which data were collected, see step five of the hybrid methodology. However, measurements for three variables were still missing. These variables are accumulated charge on the gate (q_{ox}), implant dose in the source/drain areas (S/D dose), and implant dose in the tip areas (Tip dose), also called the lightly doped drain (LDD) areas.

The values for the three parameters (q_{ox}, S/D dose and Tip dose) were estimated using empirical knowledge of process engineers. The information from engineers was qualitative by nature. However, it was translated into a quantitative piece of information. Table 6 shows the process of estimating numerical values for q_{ox}, S/D dose and Tip dose.

Equation or Expression number	Name	Equation/Expression
Equation 1	gate charge (q_{ox})	q_{ox} = f(equipment data from plasma process steps)
Equation 2	Source/Drain dose (S/D dose)	S/D dose = g(equipment data from S/D implant steps)
Equation 3	LDD dose (Tip dose)	LDD dose = h(equipment data from LDD implant steps)

Table 6. Quantitative depiction of empirical information for q_{ox} and implant dose

In the equations of Table 6, the functions "f", "g" and "h" are determined using the experience of process engineers. Examples of such functions include natural log, square root, normalization about median, a combination of them, etc.

At the end of step six, seven models from device physics and three models from the qualitative empirical understanding of process engineers have been identified. The qualitative information was translated into mathematically usable quantitative equations. Then, numerical estimates of the seven device-physics models and of the three empirical models were obtained for each observation of data.

Step 7: Develop large-scale model relating output to process data

The inputs for the final model were the numerical estimates of seven physics-based models, three empirical models, and data for sixty nine in-line variables. Response values (L_{en} and L_{ep}) were also given to the MARS modeling tool for model development [28] in the Splus statistical computing environment on windows and UNIX. (Helpful references for using Splus are [13, 50, 82].)

Response	In-line # 1	In-line # 2	.	In-line # 69	Physics model #1	.	Physics model #7	Empirical model #1	.	Empirical model #3
23	23	456	.	12.5	54.6	.	67.8	834	.	8
23.1	43	435	.	13.2	56.2	.	65.7	867.9	.	8.3
43	44.6	453	.	13.4	55.6	.	65	864	.	7.9
34	43.5	467	.	11.5	58.9	.	66.9	789	.	7.7
35.6	87.9	453	.	14.2	51	.	68.9	813	.	7.1
23.56	12.6	546	.	11.6	54.6	.	61.7	888	.	7.65
23.3	23.8	654	.	12.5	61	.	67	857	.	7.5

Table 7. Representative data file for MARS modeling tool

Table 7 shows a representative data file (with data values transformed for confidentiality reason) given to the MARS tool for model development. The first column is the response variable (L_{en}/L_{ep}). The next sixty nine columns are the in-line process variables and in-line intermediate product measurements. The subsequent seven columns are the seven physics-based engineering models. The last three columns are the empirical models. Appendix A illustrates the process of creating new variables and of developing MARS models with the help of a numerical example.

The MARS software was given this data file and a number telling MARS to develop models with a maximum of six-variable interaction. MARS discovered that a maximum of two-way interaction gave most accurate models for L_{en} (and L_{ep}), as summarized in Equation 11.

$$L_e = a_0 + a_1[\text{Etch}_{\text{rate}} - a_2]_+ + a_3[\text{ECDL}_e - a_4]_+ + a_5[\text{S/D dose} - a_6]_- +$$
$$a_7[t_{ox} - a_8]_-[\text{ECDL}_e - a_4]_+ + a_9[\text{Tip dose} - a_{10}]_-[\delta V_{tL} - a_{11}]_+ + a_{12}[\text{Etch}_{\text{rate}} - a_2]_+[\text{Tip dose} - a_{10}]_+ +$$
$$a_{13}[\text{Etch}_{\text{rate}} - a_2]_-[L_{\text{drawn}} - a_{14}]_+ + a_{15}[\text{Poly}_{\text{etch rate}} - a_{16}]_-[\text{ECDL}_e - a_4]_+ + a_{17}[\text{Etch}_{\text{rate}} - a_2]_-[V_t - a_{18}]_+ + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

Equation 11. MARS model for L_{en}/L_{ep}

In Equation 11,

1. $\text{Etch}_{\text{rate}}$ is the etch rate of the process step just before BPSG deposition
2. ECDL_e is the electrical critical dimension of the poly gate measured just after the gate is patterned
3. S/D dose is the dose in the source/drain areas
4. t_{ox} is the thickness of the gate oxide
5. Tip dose is the dose of the LDD areas
6. δV_{tL} is the change in threshold voltage due to pinch off
7. L_{drawn} is the geometric length of the channel measured when the gate is patterned
8. $\text{Poly}_{\text{etch rate}}$ is the etch rate of the poly gate
9. V_t is the threshold voltage

Equation 11 has eighteen constants (numerical values are not disclosed for confidentiality reasons), and effectively includes parameters from nine process steps. It has three main-effects terms, and six two-way interaction terms. Chapter 4, on results and discussions, lists and these results and compares them to two alternatives.

3.6 Summary

The focus of this chapter was to show the steps of our hybrid methodology, as they would be applied to an industrial problem. The chapter began by presenting background information on semiconductor manufacturing. It then introduced and motivated an industrial problem of modeling EOL channel length of MOS transistors. The chapter outlined and discussed the difficulties associated with modeling EOL channel length. Then, the different steps of our hybrid methodology were applied to the problem of modeling channel length of MOS transistor. The next chapter discusses and presents the MARS models as well as a few alternative ways of modeling channel length.

4. Results and discussion

This chapter provides the results of applying the methodology to the problem of modeling end-of-line (EOL) channel length of MOS transistors in microprocessor manufacturing. The chapter is divided into six main sections. Section 4.1 gives the final MARS model for EOL channel length. This model incorporates device-physics models and empirical knowledge about implant steps and plasma-etch steps. The model relates channel length to influential process steps identified in Section 3.5.1. Section 4.2 lists two alternate models for EOL channel length, the MARS model with data only and the PCR model with data and device physics. Using residuals plots, Section 4.3 compares the MARS model derived using device-physics models, empirical knowledge about process steps and data to the two alternatives. Section 4.4 provides intuitive understanding for why the MARS model developed using prior information (in the form of device-physics models and empirical information) gives better results than the one developed using data only. Section 4.5, then identifies a good process operating region from the MARS model which uses device-physics models, empirical knowledge about process steps and data. The good process operating region is one that gives parts close to target. Section 4.5 also discusses different ways of using the region. The last section, Section 4.6 discusses some important implications of the variables present in the final MARS model on strategies for process control.

4.1 MARS model for EOL channel length with device-physics models, empirical knowledge and data

Separate models were developed for end-of-line (EOL) channel length for p-channel (L_{ep}) and n-channel (L_{en}) MOS transistors. These models relate respective channel lengths to process parameters from influential process steps. 150 lot level-observations were used to develop the

models, which were then tested on an independent set of 92 lot-level observations. Both training and test observations were spread over two quarters (or six months) at Intel. Since the order of lots varies from operation to operation, there does not exist a unique chronological order of lots.

The framework of multivariate adaptive regression splines (MARS) was used to develop models for channel length [23]. These models used information from device-physics knowledge (see Table 5), empirical understanding about implant steps and plasma-etch steps (see Table 6) and data. Appendix A shows an example of developing a MARS model for a set of random observations. The example illustrates the inputs and outputs of the software that generates MARS models, and the method of interpreting the outputs to compile the final model equation. The software used here was developed by Hastie [28].

MARS models for L_{en} and L_{ep} (with and without the use of physics-based models and empirical knowledge) were developed in the following way. Of the 150 observations, 149 were first used to develop a MARS model, which was then used to predict the remaining observation. Then, the remaining observation was used for model development and a different observation was kept aside for prediction. These steps were repeated to develop 150 MARS models, each of which was tested on the remaining 150th observation. Such a process of model development is called generalized internal cross validation. The model which resulted in minimum root mean square prediction error was chosen to be the final MARS model. This model was then used on a different set of 92 observations for external cross validation.

Equation 12 shows the MARS model for L_{en}. (For proprietary reasons, the numerical values of coefficients are not disclosed). The model for L_{ep} has the same structure only with analogous process parameters pertinent to p-channel transistors, and with different constants. In Equation 12, the term inside a pair of square brackets is a univariate basis function. It contributes to the model under two circumstances:

1. if the numerical value of the term and the subscript next to the square brackets are positive, and
2. if the numerical value of the term and the subscript next to the square brackets are negative.

Device-physics models are represented through terms V_t (threshold voltage) and δV_t (change in threshold voltage due to pinch off) in Equation 12.

$$L_e = a_0 + a_1[Etch_{rate} - a_2]_+ + a_3[ECDL_e - a_4]_+ + a_5[S/D\ dose - a_6]_- +$$
$$a_7[t_{ox} - a_8]_-[ECDL_e - a_4]_+ + a_9[Tip\ dose - a_{10}]_-[\delta V_{tL} - a_{11}]_+ + a_{12}[Etch_{rate} - a_2]_+[Tip\ dose - a_{10}]_+ +$$
$$a_{13}[Etch_{rate} - a_2]_-[L_{drawn} - a_{14}]_+ + a_{15}[Poly_{etch\ rate} - a_{16}]_-[ECDL_e - a_4]_+ + a_{17}[Etch_{rate} - a_2]_-[V_t - a_{18}]_+ + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

Equation 12. MARS model for L_{en} and L_{ep} using device-physics models, empirical information and data

Equation 12 has nine predictor variables, two of which are from physics-based models, another two are from empirical information, and the rest are process variables. These nine predictor variables appear below:

1. $Etch_{rate}$ is the etch rate of the process step just before BPSG deposition, a process variable
2. $ECDL_e$ is the electrical critical dimension of the poly gate measured just after the gate is patterned, a process variable
3. S/D dose is the dose in the source/drain areas, from empirical information
4. t_{ox} is the thickness of the gate oxide, a process variable
5. Tip dose is the dose of the LDD areas, from empirical information
6. δV_{tL} is the change in threshold voltage due to pinch off, from a physics-based model
7. L_{drawn} is the geometric length of the channel measured when the gate is patterned, a process variable
8. $Poly_{etch\ rate}$ is the etch rate of the poly gate, a process variable
9. V_t is the threshold voltage, from a physics-based model

Equation 12 has eighteen constants. It has three main-effects terms, and six two-way interaction terms. Company personnel in semiconductor manufacturing typically examine only L_{drawn} in the event of any problems with channel length. This model helps identify a manageable set of additional process variables to examine those can influence channel length.

This paragraph defines a metric called "**model capability ratio**" used here to assess the goodness of the models. Model capability ratio is defined in a manner analogous to process capability or C_{pk}, and is the ratio of the acceptable error (in L_{en} or L_{ep}) and the model standard error (for L_{en} or L_{ep}). For the model shown in Equation 12, this ratio is 2.96 for L_{en} and 2.167 for L_{ep}, and is a measure of the acceptable error in terms of the number of standard deviations of the model error. The higher this ratio is, the better the model is.

Model capability ratio has an interesting graphical interpretation. Consider the area under a standard normal probability distribution curve between the points (zero - model capability ratio) and (zero + model capability ratio) on the x-axis, as shown in Figure 20 for L_{en} and for L_{ep}.

This area is defined here as **model prediction capability** or C_{pred}. C_{pred} for L_{en} was 0.9969 on the test set of 92 lot level observations. A C_{pred} value of 0.9969 (or 99.69%) means that the model would predict only one in 322 L_{en} observations beyond the acceptable limit. The higher the model prediction capability (or C_{pred}) is, the more the model has identified the systematic part of the signal.

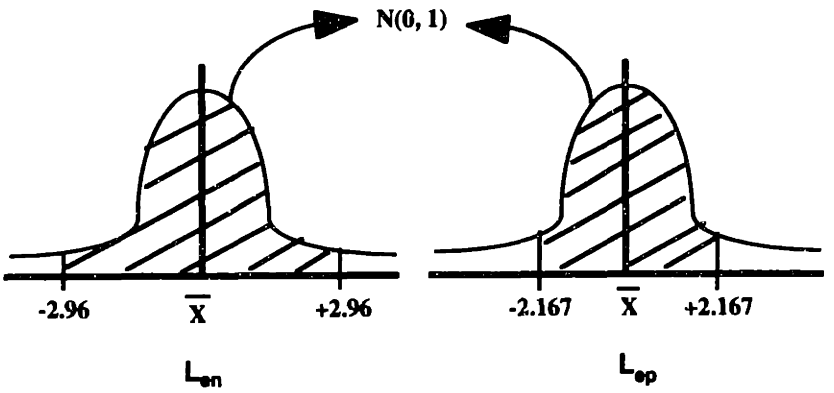

Figure 20. Model prediction capability for L_{en} and L_{ep} models of Equation 12

Figure 21 shows the L_{en} residuals for the test set of 92 observations. The variance does not appear to be uniformly spread for different observations. Unfortunately, due to the manner in which lots are processed at different operations, there is not a clean, clear order in which the residuals can be analyzed. As such, a specific reason for the variation in the spread of the residuals could not be determined.

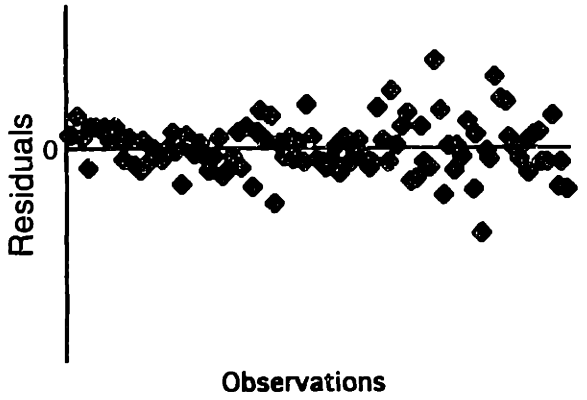

Figure 21. L_{en} residuals for MARS model with device-physics models, empirical information and data

Residuals analysis (normal cumulative probability plot and the plot of residuals vs. fitted values) of L_{en} showed that the residuals were distributed approximately normally. This is true for residuals of L_{en} and L_{ep} from other models as well.

Model capability ratio for L_{ep} is 2.167 and the Model prediction capability is 0.969 (or 96.9%), tested on a cross-validation data set of 92 lot level observations. A C_{pred} of 0.969 means that one in 32 predictions will fall beyond the acceptable error bounds. Figure 22 shows the L_{ep} residuals. For reasons mentioned earlier about the varying order of processing of lots from operation to operation, a specific reason for the variance in the spread of the residuals could not be determined for different observations in Figure 22. (This is true for the later residuals plots also.)

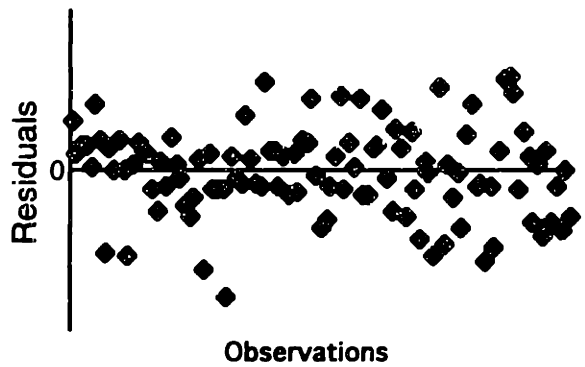

Figure 22. L_{ep} residuals for MARS model with device-physics models, empirical information and data

4.2 Alternative models

Two reasonable alternatives to compare with the above MARS models are:

1. MARS models using data only, and
2. modified principal components regression (PCR) models using data, device-physics models and empirical information. Traditional applications of the PCR method have used data only.

This section presents the alternative models.

4.2.1 MARS models with data only

Equation 13 and Equation 14 show MARS models for L_{en} and L_{ep}, respectively, derived using data only. In Equation 13 and Equation 14, most predictor variables are the same as in Equation 12. The ones that are different are listed below:

- $t_{sal\text{-}base}$ is the thickness of the base material for salicide formation
- $t_{spacer\text{-}dep}$ is the thickness of the spacers
- $Link_{E0}$ is a machine-related variable which measures energy during UV exposure in lithography of the gate
- a_i's are constants different from those in Equation 12

Unlike Equation 12, device-physics models are missing in Equation 13 and in Equation 14. Also, the machine parameters estimating gate charge are not used here, but are used in Equation 12 through the use of empirical information (Equation 1 in Table 6), which becomes part of V_t and δV_t (due to pinch off) in Equation 12. Oxide thickness (t_{ox}) does not appear in Equation 13, but is a part of the models in Equation 12 through a main effect term and as part of V_t and δV_t (due to pinch off). As before, the models in Equation 13 and in Equation 14 were developed using 150 lot level observations, and were then tested on a completely different set of 92 lot level observations

$$L_{en} = a_0 + a_1[Etch_{rate} - a_2]_+ + a_3[Etch_{rate} - a_2]_- + a_4[ECDL_{en} - a_5]_- + a_6[Tip\ dose - a_7]_+ + a_8[t_{sal\text{-}base} - a_9]_- a_{10}[t_{spacer\text{-}dep} - a_{11}]_- + a_{12}[Etch_{rate} - a_2]_-[Link_{E0} - a_{13}]_+ + a_{14}[Etch_{rate} - a_2]_-[Link_{E0} - a_{13}]_- + a_{15}[t_{sal\text{-}base} - a_9]_+ a_{16}[Tip\ dose - a_7]_+ + a_{17}[Tip\ dose - a_7]_-[t_{sal\text{-}base} - a_9]_+ + a_{18}[t_{sal\text{-}base} - a_9]_-[Link_{E0} - a_{13}]_+ + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

Equation 13. MARS model for L_{en} using data only

Figure 23 shows the L_{en} residuals. The x and y axes scales in Figure 23 are the same as those in Figure 21. The model capability ratio for L_{en} model with data only is 2.287.

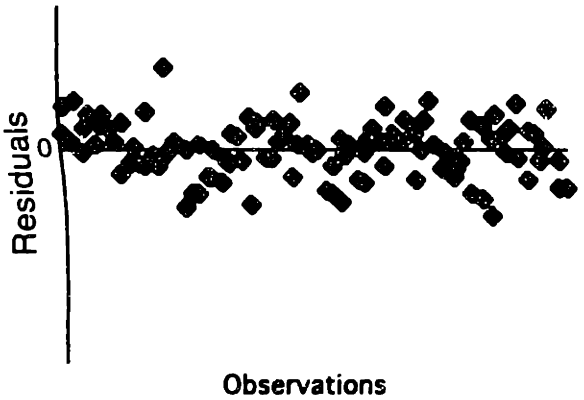

Figure 23. L_{en} residuals for MARS model with data only

$$L_{ep} = a_0 + a_1[Etch_{rate} - a_2]_+ + a_3[ECDL_{ep} - a_4]_+ + a_5[ECDL_{ep} - a_4]_- + a_6[Tip\ dose - a_7]_- + a_8[t_{ox} - a_9]_-[ECDL_{ep} - a_{10}]_+ + a_{11}[Tip\ dose - a_7]_-[Etch_{rate} - a_2]_+ + a_{12}[Etch_{rate} - a_2]_-[Tip\ dose - a_7]_- + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

Equation 14. MARS model for L_{ep} using data only

The corresponding model fit results for L_{ep} are shown in Figure 24. They are all based on the predictions of the model on the cross-validation data set of 92 lot level observations. The residuals plot has the same horizontal and vertical scales as those in Figure 22. The model capability ratio for L_{ep} model using data only is 1.439.

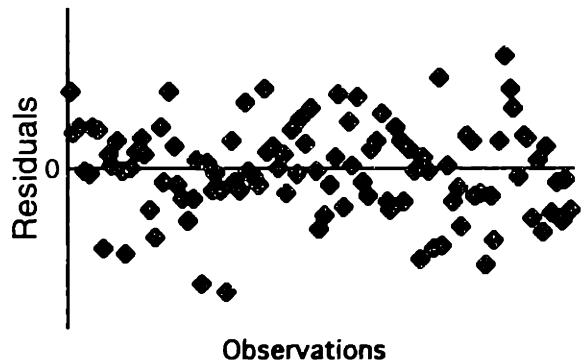

Figure 24. L_{ep} residuals for MARS model with data only

4.2.2 Modified PCR models with device-physics knowledge, empirical information and data

Traditional applications of principal components regression (PCR) use data only [51]. The modified version of (PCR) developed here uses device-physics models, empirical

information and data. The covariance matrix was determined for a data matrix (shown in Table 7), which had device-physics models and empirical information represented as separate additional columns in the original data matrix. Principal components were determined for the covariance matrix. All principal components were chosen for regression [84]. The output, L_{en} and L_{ep}, was regressed (one at a time) over the principal components. The t-scores of the regression coefficients were examined. If the absolute value of a t-score was less than two, the corresponding principal component was eliminated from the regression, and a new regression equation was developed using the remaining principal components. This process was repeated until the absolute values of all t-scores were above two. The resulting PCR model had only statistically significant principal components. There was one modified PCR model for L_{en} and another one for L_{ep}. They both turned out to have similar structure with different constants. Later, these models are compared to the ones derived using the methodology developed by this research reported in Chapter 2. By contrast, conventional PCR selects only a few principal components (that explain a given amount of variation, say 80% or 90%), and then develops a regular regression between the output and the selected principal components.

Equation 15 shows the structure of modified principal components regression (PCR) models for L_{en} (and L_{ep}). The model for L_{ep} has the same structure only with analogous process parameters pertinent to p-channel transistors, and with different constants. These models are derived using device-physics models, empirical knowledge about process steps and data. "p_i's" in Equation 15 represent principal components of the covariance matrix. The predictor variables used here are the same as those used by the hybrid methodology developed in this thesis research. As before, the model in Equation 15 was developed using 150 lot level observations, and then tested on a completely different set of 92 lot level observations.

$$\text{Output} = a_0 + a_1 p_6 + a_2 p_7 + a_3 p_{13} + a_4 p_{19} + a_5 p_{34} + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

Equation 15. Modified PCR model for L_{en} and L_{ep} using device-physics models, empirical information and data

In Equation 15,

- p_i's represent different principal components
- a_i's are model constants
- ε is random normal error term with mean zero and variance σ^2

Model capability ratio for L_{en} is 0.955 giving a model prediction capability (C_{pred}) of 0.6162 (or 61.62%), which means that one in three predictions will fall beyond acceptable

limits. Figure 25 shows the corresponding L_{en} residuals. The horizontal and vertical scales in Figure 25 are the same as those in Figure 21 and in Figure 23.

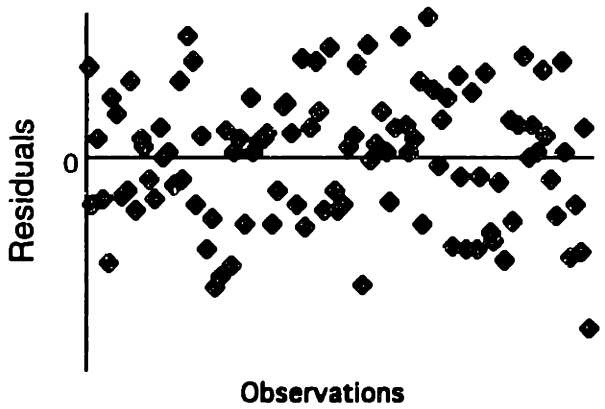

Figure 25. L_{en} residuals for modified PCR model with device-physics models, empirical information and data

Model capability ratio for L_{ep} is 0.769 giving a model prediction capability (C_{pred}) of 0.5162 (or 51.62%), which means that almost every other prediction will fall beyond acceptable limits. The corresponding model fit results for L_{ep} are shown in Figure 26 on the cross-validation data set of 92 lot level observations. The horizontal and vertical scales in Figure 26 are the same as those in Figure 24 and in Figure 22.

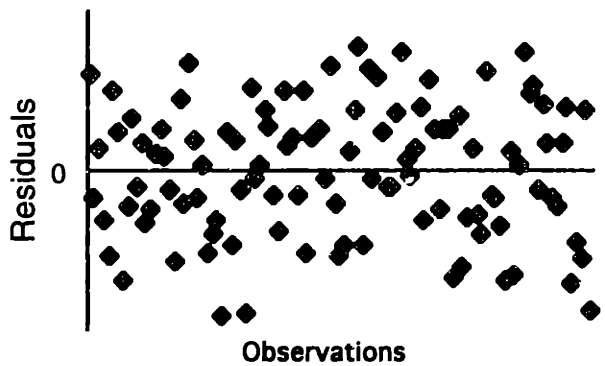

Figure 26. L_{ep} residuals for modified PCR model with device-physics models, empirical information and data

4.3 Comparison between models from different methods

Figure 27 compares the residuals plots of L_{en} for the three models. While actual numbers are not reported on the Y-axis for confidentiality reasons, the vertical scales are the same for the three plots. The first plot is the residuals plot for the modified PCR model (with device-physics models, empirical information and data). The second plot is that of the MARS model with data only, and the third plot is that of the MARS model with device-physics models, empirical information and data.

Figure 27 shows that the residuals of the MARS model with device-physics models, empirical knowledge and data are much tighter around zero than that of the MARS model with data only. The standard error of the MARS model with device-physics models, empirical information and data is more than 22% less than that of the MARS model with data only. This percentage is calculated through the use of Equation 16.

$$\% \text{ improvement} = \frac{SE_{\text{MARS with data only}} - SE_{\text{MARS with prior information and data}}}{SE_{\text{MARS with data}}}$$

Equation 16. Percentage improvement calculation for MARS models

In addition, the residuals of the MARS model with data only are much tighter around zero than that of the modified PCR model with device-physics models, empirical information and data. The standard error of the MARS model with data only is more than 58% less than that of the modified PCR model with device-physics models, empirical information and data (calculated using Equation 16).

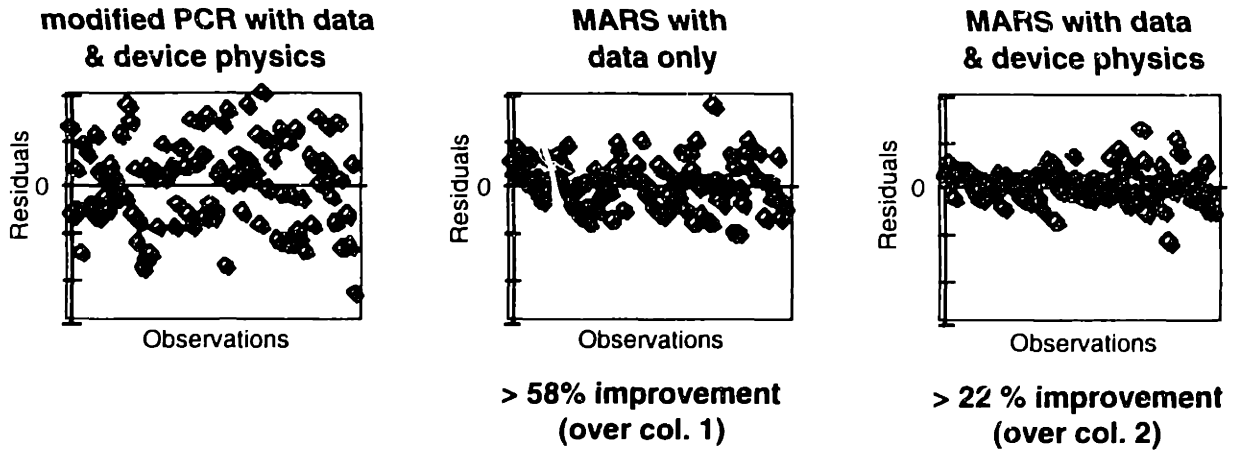

Figure 27. Residuals plots of L_{en} for the three models

Figure 28 compares the corresponding residuals plots of L_{ep} for the three models. The vertical scales are the same for the three plots.

Again, Figure 28 shows that the residuals of the MARS model with device-physics models, empirical knowledge and data are much tighter around zero than that of the MARS model with data only. The standard error of the MARS model with device-physics models, empirical information and data is more than 33% less than that of the MARS model with data only. The percent improvement here is calculated using Equation 17.

$$\% \text{ improvement} = \frac{SE_{\text{modified PCR with prior information and data}} - SE_{\text{MARS with data only}}}{SE_{\text{modified PCR with prior information and data}}}$$

Equation 17. Percentage improvement of MARS over modified PCR models

The residuals of the MARS model with data only are much tighter around zero than that of the modified PCR model with device-physics models, empirical information and data. The standard error of the MARS model with data only is more than 46% less than that of the modified PCR model with device-physics models, empirical information and data (calculated using Equation 17). Table 8 compares the improvements in standard error achieved by current models for L_{en} and L_{ep}.

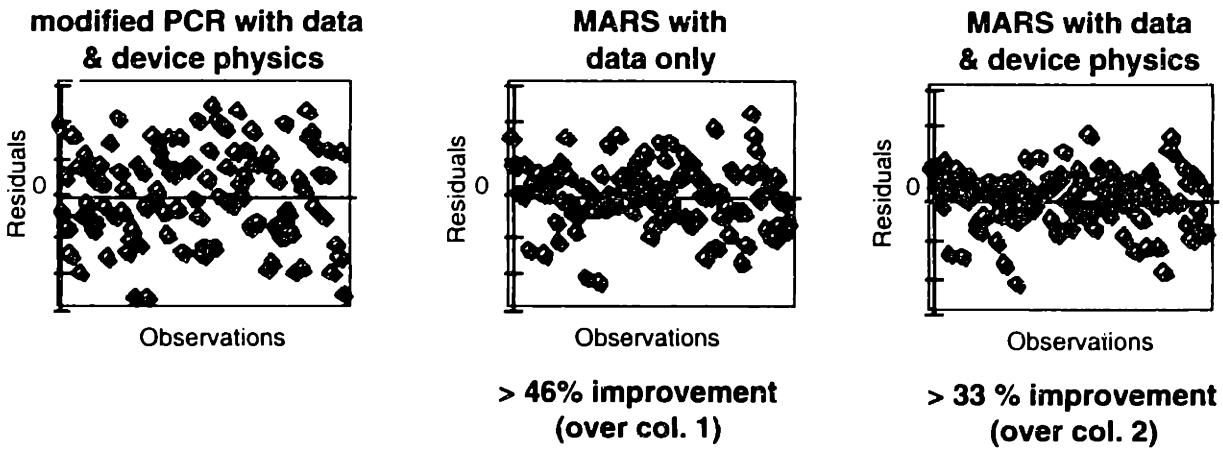

Figure 28. Residuals plots of L_{ep} for the three models

	Percent improvement in standard error of current MARS model using data only over modified PCR	Worst case percent improvement in standard error of MARS model using data only over modified PCR	Percent improvement in standard error of current MARS model using data and prior information over current MARS model using data only	Worst case percent improvement in standard error of MARS models using data and prior information over MARS models using data only	Percent improvement in standard error of current MARS model using data and prior information over modified PCR	Worst case percent improvement in standard error of MARS model using data and prior information over modified PCR
Len	58.27	56.27	22.80	14.94	67.78	66.20
Lep	46.54	43.31	33.60	24.55	64.50	62.10

Table 8. Percent improvement in standard error for current models and for the worst case

An obvious question here is if the improvements reported above are significant. This is answered by comparing the standard error of the MARS models reported here to the inherent variation in those models. If the inherent variation is small, the improvements are significant. Otherwise, the improved model reported here is a fortuitously selected one. A measure of the inherent model variation was had from the generalized cross validation standard error (GCV_{SE}). During model development, 150 MARS models were developed using only 149 observations at a time. Prediction error was computed for the remaining observations for each of the 149 MARS models. The standard error calculated from the 150 prediction errors is the generalized cross validation standard error (GCV_{SE}) representing inherent model variation.

The ratio of GCV_{SE} and standard error of the different MARS models are 0.0159 (L_{en} using data only), 0.0163 (L_{en} using prior information and data), 0.02 (L_{ep} using data only), and 0.0226 (L_{ep} using prior information and data). The ratios show that the inherent model

variability is only a couple of percentage points of the model standard error used for comparing different types of model for L_{en} and L_{ep}.

Table 8 reports the worst case comparison between different types of models for L_{en} and L_{ep} with respect to the improvement in standard errors. The worst case comparison between MARS models (using data only) and modified PCR models is when the standard error of the MARS models is higher than their current value by three times the GCV_{SE}. Specifically, the improvements in L_{en} and L_{ep} are 56.28% and 43.31%, respectively, for the worst case comparison. The worst case comparison between MARS models using data only and MARS models that use data and prior information is when the standard error of MARS models using data only is lower than their current value by three times their GCV_{SE} and that of MARS model using data and prior information is higher than their current value by three times their GCV_{SE}. Here, the worst case improvements are 14.9% and 24.54% for L_{en} and L_{ep} respectively. Similar numbers comparing MARS models using data and prior information to modified PCR models show improvements of 66.2% and 62.1%, respectively, for L_{en} and L_{ep}. The comparison numbers for the worst case confirm that the improvements in standard errors of the MARS models (with and without the use of prior information) are still quite a bit higher than that of the modified PCR models, and the improvements in standard errors of MARS models that use prior information are still higher than that of MARS models that use data only.

Graphically, the comparison between MARS models that use prior information with those that use data only provides another perspective to the system being modeled. The use of data only in MARS results in hyper-rectangular regions (or partitions) in the process-operation space. Within each of the hyper-rectangular regions, MARS fits a function that obeys some continuity criteria at the boundaries of the partition. By using physics-based models and empirical information, MARS splits the process-operation space into non-hyper-rectangular regions. The flexibility in the shape of the partitions can help divide the process-operating space more accurately. As such, the final model would characterize the system more accurately.

We can now compare the above results to the objective of this research outlined in Section 1.5. Both Figure 27 and Figure 28 show that our methodology developed more accurate models for L_{en} and L_{ep} than the two alternative models. This is clear if we compare the plots in column three in the two figures to those in columns one and two. In addition, note that the plots in column three give improved standard errors over those in column two (more than 22% improvement for L_{en} and more than 33% improvement for L_{ep}). This shows that, in this case, combined models (which use device-physics models and empirical information with data) for the current application problem are more accurate than statistical models (which use data only) alone.

Also, the plots in column three gave improved standard errors compared to those in column one. This shows that, in this case, the local modeling method (MARS) characterized a large manufacturing operation better than a global modeling method (modified PCR).

To be sure, the comparison between MARS models (with and without the use of prior information) and conventional PCR models comes with a caveat. Conventional PCR modeling method does not use the response to determine principal components used in the final regression. It chooses only a few that explain a specified variation in the input variables instead. As such, conventional PCR modeling method may identify principal components that may not have enough signal to model the response, even though the original inputs may be rich in such a signal. This limitation of conventional PCR is overcome here by using all principal components, and then using a t-test to identify the statistically significant ones. On the other hand, MARS modeling method uses response information at every stage, in developing the partitions and in developing the regression.

PCR was chosen as a comparison alternative in this research because (conventional) PCR is a popular multivariate statistical modeling method that is used extensively in manufacturing companies (including those in the semiconductor industry). A comparison of MARS models with partial least squares (PLS) regression would also be appropriate and interesting, and is recommended as part of future work.

4.4 Why does MARS with prior information give improved models?

Section 4.3 demonstrated that MARS models developed using prior information (in the form of device-physics models and empirical information) predict end-of-line (EOL) channel length better than (an improved version of conventional) PCR and MARS with data only. This section aims to provide intuitive understanding about why MARS models with prior information give better results.

Figure 29 shows three scatter plots. The first plot is between the output, end-of-line channel length, and the thickness of the oxide (t_{ox}). The second plot is between end-of-line channel length and the charge on the gate (q_{ox}). The third plot is between end-of-line channel length and an implant machine parameter called thermal wave (TW). For simplicity, the output is shown as L_e. L_e could represent channel length of n-channel transistors (L_{en}) and/or of p-channel transistors (L_{ep}).

All three scatter plots look like "fuzzy balls", implying lack of correlation (or signal) between the output (L_e) and t_{ox}, q_{ox} or TW. Figure 29 indicates that L_e cannot be modeled from t_{ox}, q_{ox} or TW.

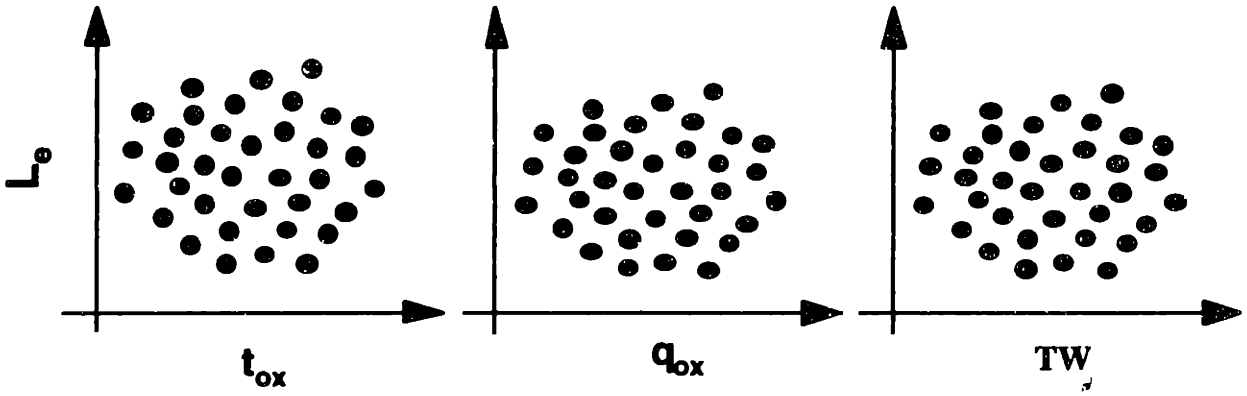

Figure 29. Scatter plots between channel length Vs oxide thickness, gate charge and TW

Figure 30 shows the scatter plot between the output (L_e) and the device-physics model in the form of $q_{ox}*t_{ox}$. The correlation (or signal) between the output and $t_{ox}*q_{ox}$ improves in part of the scatter plot. The scatter plot in Figure 30 can be divided into three regions. Region I and region II show some signal between L_e and $q_{ox}*t_{ox}$. In these regions, L_e depends on $q_{ox}*t_{ox}$ as a straight line. The slope and intercept of the straight line in region I can be different from that in region II. Region III of the scatter plot in Figure 30 is still a fuzzy ball, implying lack of correlation (or signal) between L_e and $q_{ox}*t_{ox}$.

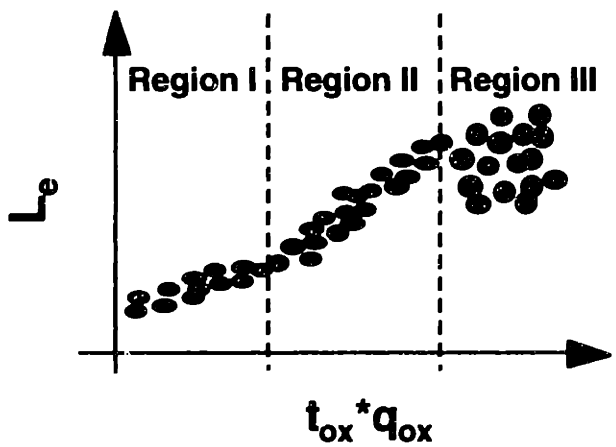

Figure 30. Scatter plot between channel length and device-physics model ($t_{ox}*q_{ox}$)

However, different prior-information (e.g., empirical information about source drain implant process step such as S/D dose value calculated from many TW parameters) in region III in Figure 30 can replace the fuzzy ball by a meaningful signal. Figure 31 shows a scatter plot between L_e and $q_{ox}*t_{ox}$ in regions I and II, and between L_e and S/D dose in region III.

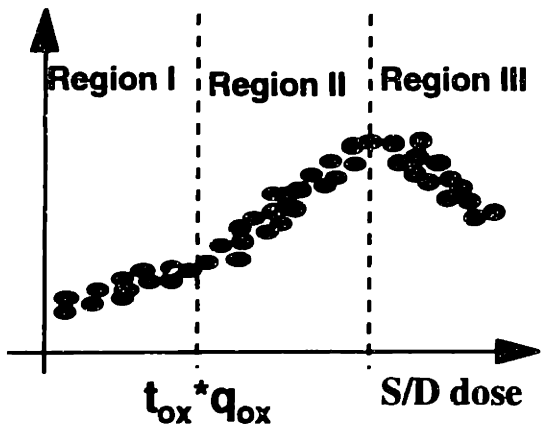

Figure 31. Scatter plot between channel length and device-physics model ($t_{ox}*q_{ox}$) in regions I and II, and S/D dose in region III

Figure 31 shows that L_e can now be modeled. It is a linear function of $q_{ox}*t_{ox}$ in regions I and II. In addition, it is a linear function of S/D dose in region III. The identification of $q_{ox}*t_{ox}$ and S/D dose as predictor variables in regions I, II and III respectively helped generate the signal for L_e.

How does MARS help in generating and maximizing the signal for L_e? MARS exploits the fact that prior information can help generate clear signals for L_e, compare Figure 29 to Figure 31. In addition, MARS exploits the fact that some pieces of prior information can generate a clearer signal for L_e in a region than what other pieces of prior information can do. For example, S/D dose helped generate a clearer signal for L_e in region III than what $q_{ox}*t_{ox}$ did, compare region III in Figure 30 with region III in Figure 31. MARS accomplishes two tasks simultaneously. It determines the different regions (e.g., the three regions in Figure 31), and selects different pieces of prior information to maximize the signal to noise ratio. It defaults to the use of data if data provides a better signal than any of the available prior information.

A possible situation where available prior information may not be useful is when the data are collected from one region of operation, but the available physics-based models and empirical information are valid in a completely different region of operation. For example, consider that the MARS software is provided with models for short channel effects only and data that are

collected near operating points where long channel effects are predominant. In this case, prior information from device-physics models would not be used; MARS will try to model long channel effects by using data only.

4.5 Identification of good process operating region

A distinct advantage of MARS models is to identify regions of process operation that result in good parts, and those that result in bad parts. If the MARS model includes physics based understanding, then the process operating region identified from it should be based on causality relationship of the physical model. On the other hand, a region identified from a MARS model using data only could result from a mere coincidence of numbers. As such, the use of physics-based models can help determine good process operating regions more reliably. Here, we now present an example of a process-operating region which results in good parts.

Figure 32 shows a simplified 3D region which resulted in most parts in the test data that fell within a small window of target +/- 0.5σ. The 3D region is made up of the basis function in the MARS model that these parts invoked. In the figure, the x-axis is a measure of gate charge (q_{ox}) accumulated during processing. The y-axis is the thickness of the gate oxide (t_{ox}). The z-axis is etch rate of the process step just before BPSG deposition. As long as the three parameters along the three axes fall within the solid during processing, there would be good parts at the end of the manufacturing line. However, if a point falls outside of the solid during processing, there would be a bad part at the end of the line. According to Figure 32, the interaction between t_{ox} and q_{ox} is important in determining the final quality. But, the etch rate before BPSG deposition remains an independent axis. That the etch rate is independent, makes intuitive sense. Also, that there are bounds on etch rate again makes intuitive sense. According to a process engineer at Intel, a very high etch rate will etch away too much silicon from the source/drain (S/D) area, and push the top surface of the S/D below the lower edge of the gate oxide. As such, the S/D will have more difficulty transferring minority carriers to and from the channel region under inversion. This will result in a dysfunctional transistor.

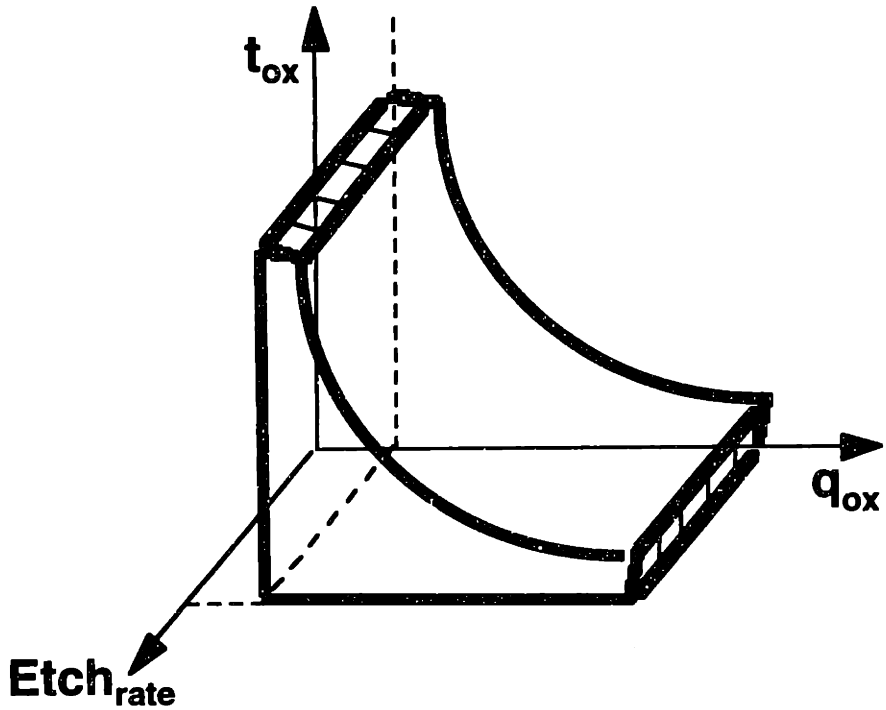

Figure 32. A 3D sketch of a good process operating region

The following discussion investigates the interaction between gate oxide thickness (t_{ox}) and gate charge (q_{ox}) accumulated during processing. Figure 33 shows a 2D cross-section of the 3D solid in Figure 32. The shaded region is a good process operating region for t_{ox} and q_{ox}. Figure 33 makes intuitive sense. According to Figure 33, there is a particular relationship between t_{ox} and q_{ox} which defines the bounds of good operation. If the process runs beyond these bounds, it would produce bad parts at the end of the line. (Mathematically, the region below the curved line is represented by $t_{ox}*q_{ox}<$ constant). The following paragraph explains the shaded region qualitatively.

In the shaded region, the gate still has control over the channel. Outside of the shaded region, the gate loses that control. In Figure 33, line 1 shows that for a given oxide thickness, there is a critical value of the gate charge (q_{cr}), beyond which the gate loses control on the channel. Similarly, according to line 2, there is a critical value of oxide thickness (t_{cr}) for a given amount of gate charge. For oxide thickness less than the critical oxide thickness, we still have control on the channel. For oxide thickness more than that the critical oxide thickness, the gate loses control on the channel[1]. q_{cr} and t_{cr} are determined from the mathematical relationship for

[1] The author wishes to thank Intel Corporation's Ann Nelson for assisting in developing the arguments for critical gate charge and oxide thickness through the use of lines 1 and 2.

the shaded region (t_{ox}*q_{ox}< constant). For a given value of one of the variables (t_{ox} or q_{ox}), the critical value for the other can be determined from this relationship.

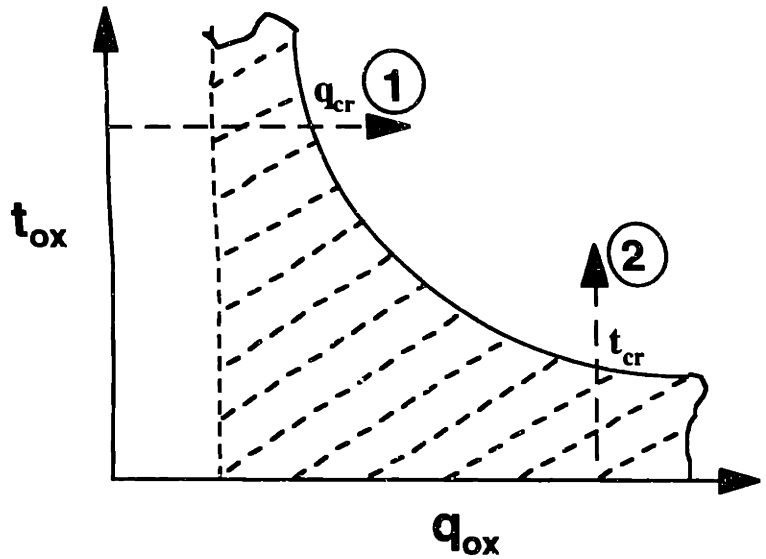

Figure 33. A 2D cross-section of a good process operating region

Engineers and researchers in the semiconductor industry know that t_{ox} and q_{ox} interact with each other. They know that for a given q_{ox}, there is a critical oxide thickness beyond which the gate will tend to lose control over the channel, and vice-versa. The methodology developed in this research has helped formalize that intuitive understanding mathematically. It has helped determine the shape of the interaction between t_{ox} and q_{ox} (from physics-based models and empirical information), and helped identify the exact parameters of that interaction (based on measurement data).

4.5.1 Practical uses of the good process operating region

Section 4.5 discussed the identification of a region in the process operating space that results in good parts. In doing so, our methodology formalized an intuitive understanding mathematically. This Section discusses a key practical question "How should we best use the good region of Figure 33?".

All points in the good operating region of Figure 33 do not result in equally good parts. Some points result in better parts than others. If we want to produce parts within a small window of target +/- 0.5σ, then we have to solve Equation 12 for L_{en} equal to target - 0.5σ and

again for L_{cn} equal to target + 0.5σ. (For L_{cp}, Equation 12 will be solve with different values of the constants in it). This will give us a subregion within the region shown in Figure 33. This subregion is shown in Figure 34. The next paragraph explains how we use the subregion.

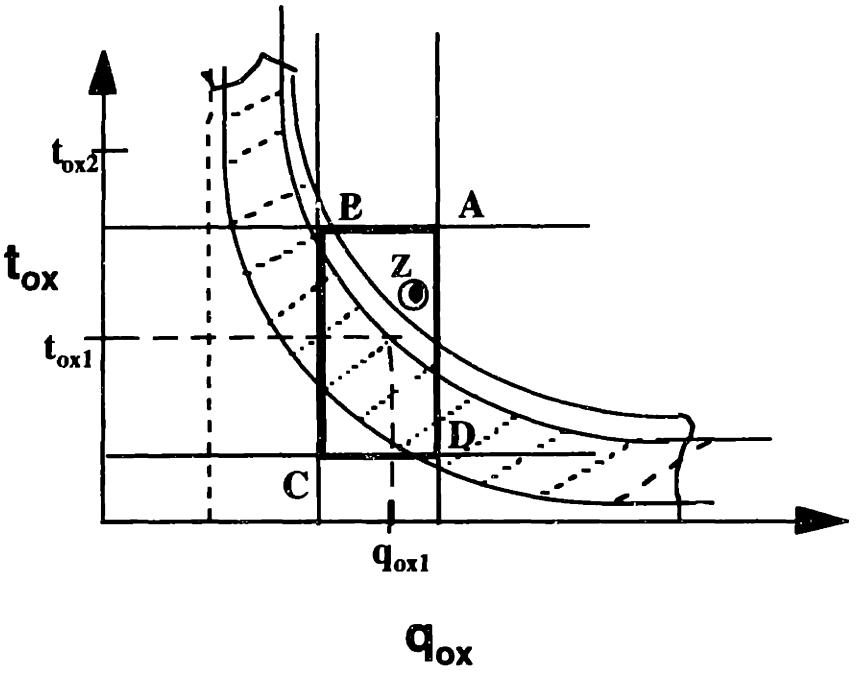

Figure 34. A subregion resulting in output within target +/- 0.5σ

Assume current specification limits (spec limits) for t_{ox} and q_{ox} as shown in Figure 34. The intersection of the two spec limits is shown as the box ABCD. The curved lines represent constant L_{cn} (or L_{cp}) lines, such that the lower curved line is target - 0.5σ and the upper curved line is target + 0.5σ. For reasons discussed in Section 1.2, points like the circled point Z in Figure 34 are avoidable. These points can be avoided by **tightening the spec limits** on t_{ox} and q_{ox} such that the box ABCD lies completely within the subregion marked by the lower and the upper curved lines. Figure 35 shows the new box, EFGH, resulting from tightened spec limits on t_{ox} and q_{ox}.

An obvious result of tightening spec limits (to generate univariate control charts) is the reduction in the operating space. Box EFGH is smaller than box ABCD. For a high dimensional problem, this operating space may become impractically small, so much so that a feasible set of univariate spec limits may not even exist. Instead of using the multivariate model resulting from the current research methodology for generating univariate control charts, the

model should be used to develop appropriate multivariate control charts and control strategies. Additional control options resulting from an accurate multivariate model are discussed further.

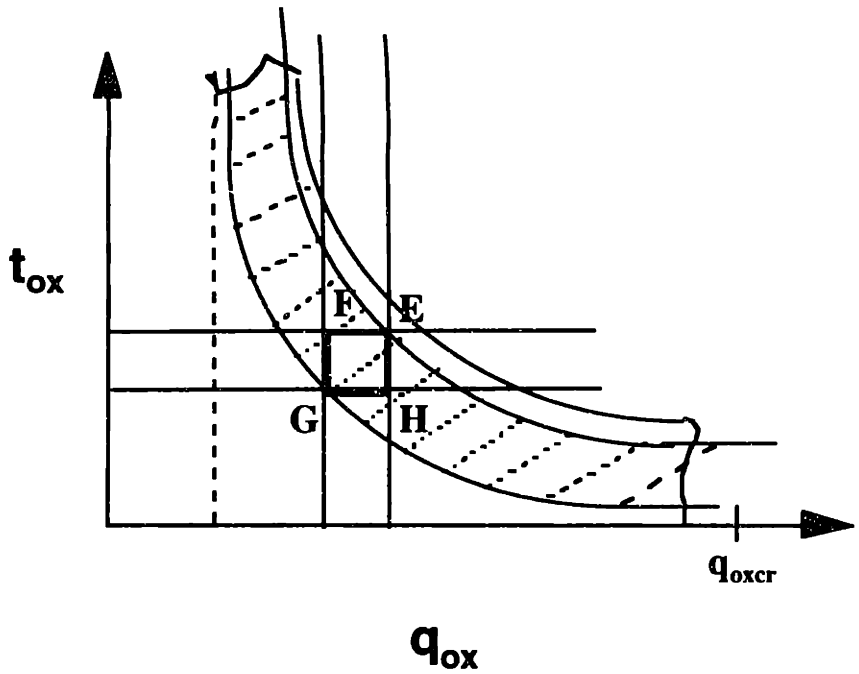

Figure 35. Multiple uses of subregion from Figure 34

Determine more **robust set-points** for t_{ox} and q_{ox}. If the process parameter influencing q_{ox} is a harder parameter to control than the one influencing t_{ox}, then the set point for q_{ox} should be moved to a new position q_{oxr} shown in Figure 35. This makes the output L_{en} (or L_{ep}) more robust to disturbances in q_{ox}. This step should be understood in some additional perspective. By changing the set point for q_{ox}, L_{en} (or L_{ep}) has become more robust to disturbance in q_{ox} for the same target value of L_{en} (or L_{ep}). However, the new set point for q_{ox} may affect some other EOL product characteristic adversely. To avoid such undesired effects, each critical EOL product characteristic should be modeled as a function of important in-line parameters, as discussed in Section 5.3.1. Those models will show if a critical EOL product characteristic is affected adversely by the new set point for q_{ox}.

The subregion in Figure 34 can also be used in determining appropriate **feed-forward control** action. The act of changing down-stream process parameters in a manufacturing line to off-set upstream disturbances is called feed-forward control action. Suppose that gate oxide is grown and measured chronologically before the process steps that result in gate-charge accumulation. Assume that the spec limits for t_{ox} and q_{ox} are shown as the box ABCD in Figure 34, and that t_{ox} was measured to be t_{ox1}. From Figure 34, it is clear that parts will be out-of-spec

for $q_{ox} > q_{oxl}$. To avoid out-of-spec parts, the process step resulting in gate-charge accumulation should be run such that q_{ox} is within its spec limits but less than q_{oxl}.

Feed-forward control action needs good data acquisition, integrated monitoring and control. Besides being expensive, it is also contrary to the idea of robust manufacturing processes. Robust processes are necessary in high-volume manufacturing for minimum variation in final products. Consequently, feed-forward control is not a recommended technique for process control. As such, the process should be run to result in oxide thickness different from t_{oxl}. If the oxide thickness on a part is equal to t_{oxl} during processing, **feed-back control** action should be taken so that future parts do not have oxide thickness like t_{oxl}.

Lastly, the region in Figure 34 can also be used to decide whether a part should be rejected early in the process if the variation introduced in it is already very high. If the oxide thickness on a part is t_{ox2}, see Figure 34, it may be more cost-effective to **reject the part early** than to add value to a potentially bad final product or to take a feed-forward control action to offset the variation in oxide thickness.

We do not recommend one way of using the good process operating region over another. The optimal method for keeping products on target will be decided by feasibility (robust setting may not always be possible due to the working limits of an actuator) and cost considerations (tightening certain process specifications may require costly equipment for control). However, a practically meaningful analysis of available options is possible only if regions such as the one shown in Figure 33 are identified and understood with the help of an accurate **multivariate model**.

4.6 Process monitoring and control implications of variables in MARS model

Step four of our hybrid methodology had identified sixty nine potentially influential process parameters for EOL channel length. These parameters are from the fifty one process steps listed in Table 4. However, the final MARS models contain only nine unique process parameters, as shown by Equation 12. Does that mean that only nine out of sixty nine process parameters are important to monitor and control EOL channel length? Should the rest of the sixty parameters not be monitored or used for process control? This section aims to address these issues.

The nine process parameters in Equation 12 are certainly important parameters to predict EOL channel length. This is confirmed through cross validation of the two equations (or

models) on independent data sets of 92 observations. The model prediction capabilities of the two models were 96.9% and 99.6% respectively using a completely independent data set from the one used to develop the two models. This is a respectably high prediction accuracy for L_{en} and L_{ep} from in-line process parameters.

However, good prediction from a statistical model does not necessarily imply good process control, see sections 1.3 and 1.4.3. This is because statistical models are "correlation-based" models that provide good prediction accuracy. On the other hand, causality based models are necessarily for a precise process control. MARS is a statistical modeling tool. (So are NN, PCR, PLS, CART, etc.) As such, MARS alone does not guarantee precise process control.

The use of prior information in MARS can bring in a degree of causality in MARS models. Physics-based engineering models are derived from causality considerations. The presence of these models in the final MARS models helps develop causality, and improves confidence in process control.

Process control (and monitoring) strategies developed using process variables that are part of the piece-wise engineering models in MARS are likely to provide good process control. In addition, often intuitive or physics-based reasons support the presence of many other process variables in the final MARS model. Such reasons add confidence in process control techniques developed using those process variables, even if they were not part of a formal engineering model. Objectively speaking, the accuracy of and the confidence in strategies for process control through other variables (not supported by physics-based reasons or piece-wise engineering models) is hard to predict. Their use in process control can only be checked through controlled experiments.

Process variables that do not enter the final MARS models, e.g., the remaining sixty one variables, are not necessarily unimportant process variables. They do not enter the final MARS model for one of the following two reasons:

1. they have no signal (or information) for EOL channel length
2. their effects are taken up by another process variable that entered the final MARS model (main effect) or by a set of such process variables (interactions). This is a common situation with MARS (and with other statistical modeling techniques) when multicollinearity exists in the input variables

In the first case, if the process variable neglected in the final MARS model has no information for EOL channel length, then it is not an important process variable for channel

length. In the second case, the neglected process variables are unimportant for channel length only if the process variables present in the final MARS models are the only drivers for (or are causal to) the EOL output. The neglected process variables are merely correlated to the process variables in the MARS models

On the other hand, a MARS model will be poor for process control if the actual drivers for the EOL output have been neglected in favor of non-drivers, but correlated, process variables. Unfortunately, there is no clear method of differentiating between driver and non-driver process variables, except by using physics-based causal information (and models). Besides providing a high prediction accuracy, the process variables supported by physics-based information are also good candidates for process control. The need for good process control is a major motivation to include physics-based models and information at several steps in this thesis research on developing a hybrid modeling methodology.

5. Conclusions

The goal of this research was to develop a new modeling methodology combining physics based modeling methods and statistical modeling methods. An important question that the research aimed to address is whether combined models (models that combine physics-based models and data-driven statistical models) are better than physics-based models or statistical models used alone. In addition, the research also endeavored to show that the local modeling method, MARS, characterizes the large manufacturing system modeled in this research better than the global modeling method, PCR.

Besides, a notable practical accomplishment of the work reported here is the use of state-of-the-art multivariate statistical methods on industrial data [76]. Even with the availability of more complete data bases, the lack of which prohibited sophisticated multivariate analysis in the past, current industrial data-analysis practice has not moved beyond univariate analysis or simple multivariate analysis like conventional PCR. This research work has applied a few sophisticated multivariate statistical methods (like CART, MARS and a modified PCR) to industrial data.

This chapter summarizes the main lessons learned from our effort. It is divided into three parts. Section 5.1 lists the contributions of this research, while Section 5.2 discusses the limitation of the current research. Finally, Section 5.3 recommends a few areas for future investigation.

5.1 Thesis contributions

This thesis research makes contributions in the following areas:

1. development of a new methodology that develops hybrid models (the hybrid methodology comprising seven steps)
2. methods to identify influential process steps for a given end-of-line (EOL) output
3. extension of MARS to combine physics-based models and empirical information with data (the seventh step of the hybrid methodology)
4. identification of good process-operating regions
5. increased understanding about the influence of process parameters on EOL channel length

The following sub-sections describe each contribution.

5.1.1 Generalizable hybrid model development methodology

This thesis has developed a new comprehensive methodology to model an EOL output parameter as a function of important in-line process parameters. Prior methods were limited in modeling large systems primarily because they failed to use the domains of engineering and statistical models effectively, see sections 1.3 and 1.4. Their single jump from engineering domain to statistical domain, and vice-versa, prevented them from borrowing information effectively.

The methodology proposed in this research uses information from three domains. These are the engineering (or physics-based) domain, the statistical modeling domain and the data domain. By using the three domains methodically, this research takes a step further in modeling large systems, such as a whole manufacturing line. When applied to microprocessor manufacturing, the methodology has given very encouraging results.

We believe that researchers in the area of system modeling know of our hybrid methodology in bits and pieces. We also acknowledge that engineers in manufacturing companies have used bits and pieces of the methodology for decades. However, those bits and pieces have given only limited, but respectable, results. By putting those bits and pieces together and adding important missing ones, the methodology proposed in this research provides a formalized, disciplined approach to the process of modeling.

The methodology, comprising seven steps, does not assume any particular characteristic about the EOL output or the manufacturing process. We believe that the methodology is applicable to several EOL output parameters (continuous and categorical) and manufacturing processes (continuous and batch). As a testimony to the generality of the applicability of the methodology, the next section applies the steps of the methodology to a continuous

manufacturing process that makes photographic film. For comparison purposes, the applicability of the methodology to the largely batch-processing semiconductor-manufacturing process is also reviewed.

5.1.1.1 Hybrid methodology applied to a continuous manufacturing process

This section discusses the applicability of the hybrid methodology to a continuous manufacturing process which makes photographic film. The raw material is a plastic which is first melted and put on a rotating extruder wheel. The extruder wheel lays the melted plastic flat as it cools on the rotating wheel. The film is then stretched in the long direction and in the cross direction. A heating and cooling cycle releases the stress in the film generated during stretching. A few types of chemical bases are applied on the film to improve its strength. At the end of the line, a few measurements are taken on the film.

Table 9 shows the seven steps of our methodology as they would be applied to the continuous manufacturing process of making the film. The output is the thickness of the film, and the inputs are the several process variables (such as temperature, clamp force during stretching, air-flow rate on the extruder wheel, etc.) and any intermediate product measurements (such as opacity of the intermediate film).

Table 9 also shows the seven steps of our hybrid methodology as they were applied to the largely batch-manufacturing process at Intel Corporation to model end-of-line channel length.

Step Number	7 steps of hybrid methodology	Modeling EOL channel length on Intel's batch-processing manufacturing operation	Modeling EOL film thickness in a continuous manufacturing process
Step 1	Choose EOL inspection variables	channel length, breakdown voltages, threshold voltages, saturation currents	film thickness, film opacity, film refractive index
Step 2	Collect data for EOL inspection variables	Collect data for channel length, breakdown voltages, threshold voltages, saturation currents	Collect data for film thickness, film opacity, film refractive index
Step 3	Develop model relating output to other EOL inspection variables and identify statistically significant inspection variables	channel length = f(threshold voltages, breakdown voltages, saturation currents). Identify statistically important inspection variables. Assume they are threshold voltages & breakdown voltages	thickness = f(opacity, refractive index). Assume that opacity and refractive index are statistically significant
Step 4	Identify influential process steps (and associated process variables) influencing statistically significant EOL inspection variables and the EOL output	Influential process steps those that deposit poly, etch poly lines, substrate implant, etc.	Influential process steps are: longitudinal and cross-stretch operations, annealing operation, and extruder-wheel station.
Step 5	Collect process data	Collect process data for process variables in the process steps identified in step 4	Collect data for process variables identified as influential in step 4
Step 6	Collect piece-wise engineering and empirical models	Collect device-physics models and empirical models, described in section 3.5.2.	Collect models, e.g., thickness = f(extruder-wheel rpm, air-flow rate on wheel); thickness = f(cross-stretch & longitudinal force); thickness = f(anneal temp & time); Collect constitutive models showing influence of temp & chemical composition of polymers
Step 7	Develop a large-scale model	Develop a large-scale model using the data for process variables collected in step 5 and for engineering and empirical models in step 6	Develop a large-scale model (MARS, CART, NN, etc.) by using data for process variables collected in step 5 and for engineering and empirical models in step 6

Table 9. Seven steps of methodology applied to model film thickness in a continuous manufacturing process

5.1.2 Identification of influential process steps without using process data

This investigation has helped us develop a new way to identify influential process variables for a given EOL output, even if process data are unavailable. The procedure to do it is developed in the first four steps of our methodology, and relies on exploiting the multivariate relationship among EOL inspection variables and the output, and then on the use of physics-based models, statistical modeling on data, and empirical qualitative information from past experience.

Often, data from critical process steps are unavailable. The reasons for data unavailability include technological limitations and cost. Traditional statistical methods would not help if process data are unavailable. DOE would not even be applicable. In the case example of Chapter 3, no data are collected for several critical process steps, e.g., well implants. While most statistical methods would have failed to identify them as influential, our methodology managed to identify them as important process steps for channel length, see Table 4. Once clearly identified as being important, these process parameters can then be monitored appropriately for process control.

Our methodology incorporates techniques of using in-line process data, empirical models and physics-based models. But, it is not critically dependent on any one of them. Of course, the methodology will fail if none of them are available. However, it will still work if any information (empirical models/physics-based models) or data are available. Like most other methods, the efficacy of our methodology would most likely improve with the availability of more information and data.

The ability of the methodology developed in this research to use traditional statistical methods and models, but not be critically dependent on any one of them is an important advantage in modeling real production processes.

5.1.3 Extension of MARS to combine engineering & statistical models and data

Multivariate adaptive regression splines or MARS has traditionally used only data for model development. This investigation identified five places where MARS can use prior information about the system. Prior information can be in the form of empirical models or physics-based models. The five places where MARS can use prior information are:

1. through split variables in split nodes

2. definition of split criteria in split nodes
3. choice of basis functions
4. choice of the order of basis functions
5. definition of pruning criteria

This research extended the use of MARS by exploring and developing the first option. Currently, the use of only data in MARS results in rectangular regions in hyper-space (defined by several predictor variables), unless interaction terms are explicitly used as new variables in the input data file to the MARS modeling software. In the absence of prior information, identification of a reasonable set of interaction terms from an unmanageably large set of possibilities is almost impossible. As such, explicit use of interaction terms is not common. By using prior information (in the form of empirical models and physics-based models) as split variables, MARS develops non-rectangular regions in hyper-space.

The use of physics-based models within the non-rectangular regions can result in a more accurate model relating the output to the inputs over the use of conventional MARS. This is because physics-based models can bring in functions like exponential, ratios, log, etc. that conventional MARS will find difficult to represent if the functions are not explicitly supplied as additional data columns in the input data file to the modeling software.

The development of non-rectangular regions helps characterize a large dynamic system more accurately and efficiently. This is particularly useful when an output depends on a particular combination (or functional form) of inputs at some operating point. But, it depends on a very different combination (or functional form) of inputs at another operating point. This behavior is commonly observed in production lines.

In general, more pieces of prior information would result in a more efficient characterization or modeling of production lines. With more piece-wise models, MARS will have a greater choice in finding split variables. There will be a greater possibility of finding a split variable from prior information that minimizes total variance. In the absence of prior information, the reduction in variance would be accomplished by using several other predictor variables instead. However, the reduction in variance may still be less than the reduction in variance when prior information is used.

In addition, in the absence of an appropriate piece of prior information, MARS selects substitute pieces of prior information (and predictor variables) based on "correlation" characteristics rather than on causality considerations. (The use of input/output correlation is an advantage of MARS over methods like conventional PCR that do not exploit the input/output

correlation). This complicates strategies for process monitoring and for process control, as discussed in Section 4.6. As such, the identification of as many pieces of prior information as possible in step six of the hybrid methodology is highly recommended. Step six aims to identify piece-wise physics-based models and empirical information, as discussed in Section 2.2.2. The use of as many pieces of prior information as possible in the final model development step, step seven, of the hybrid methodology is also highly recommended.

5.1.4 Identification of good process-operating region to improve quality

When the critical process variables (identified using the first four steps of our methodology, see Chapter 2) in a production line fall within a particular range of values, the EOL output is within specification (or in-spec). Otherwise, the EOL output goes out-of-spec.

This research helps identify the range of values of the critical variables that results in products in-spec. We call this the good process operating region, see Figure 33 for an example of a good process operating region. This research also helps us identify the range of values of the critical parameters that results in out-of-spec products. We call this the bad process operating region.

The identification of good process regions can help in several aspects of technology development and process control, see Section 4.5.1. Here, we will discuss one such aspect. Through the example of oxide thickness (t_{ox}) and gate charge (q_{ox}) in Figure 34, we also demonstrate that the range of values of one critical variable could be dependent on the range of values of another. This dependence has significant implications in setting up spec-limits for different process variables. The current practice of setting up spec-limits assumes independence between different process variables. This assumption is valid only when there is no interaction between those variables. In graphical terms, those variables only give rectangular regions of operation in hyper-space. For example, the spec-limits of etch rate are independent of those of t_{ox} (or q_{ox}) in Figure 32. Also note that the cross-section of the 3D solid in Figure 32 along the etch-rate/t_{ox} plane or along the etch-rate/q_{ox} plane is a rectangle. However, the spec-limits of q_{ox} are very much dependent on those of t_{ox}. The cross-section of the 3D solid in Figure 32 along the q_{ox}/t_{ox} plane is shown in Figure 33. That cross-section is not a rectangle.

The idea of dependent spec-limits for different process variables is not new [46, 77]. However, this research presents a way to identify good regions of operation. These regions can be more effectively used to identify clearer dependence between the spec-limits for different process variables.

An important question often asked in manufacturing companies is "How should we run our production line to produce quality parts?". By helping identify good and bad process operating regions, this research directly aims to address that question.

5.1.5 Increased understanding of EOL channel length

Conventional understanding about channel length in the semiconductor manufacturing industry is to examine L_{drawn} if there are problems with the measurement of channel length. L_{drawn} is the geometric length of channel measured just after poly gate is patterned. In existing research, there are several physical characteristics of a transistor understood to influence channel length (including L_{drawn} and oxide thickness). However, a mathematical relationship depicting the influence is still missing, making it difficult for manufacturing companies to exploit the information in additional variables beyond what can be had through univariate analysis.

This thesis research has developed models for channel length. Besides L_{drawn}, the models identify many more variables that can together predict channel length within the desired accuracy. The models also show the influence of these variables mathematically through main effects and two way interaction terms in Equation 12. An improved understanding about channel length through an accurate multivariate model will help examine variables other than L_{drawn} in manufacturing, and help determine more efficient, cost effective ways to achieve smaller channel-length variation in high volume production.

5.2 Limitations of current research

The strengths of our methodology have helped discover and develop improved modeling techniques. Simultaneously, they have helped identify several areas of limitation. In these areas, parts of the methodology are either inapplicable or they rely on assumptions about the system to be modeled. These areas of limitation are:

1. data-rich environment
2. model single output at a time
3. class of models made possible by MARS

The following paragraphs discuss each of those limitations. Section 5.3 focuses on possible techniques to avoid these limitations, and discusses other areas for further investigation.

5.2.1 Data-rich environment

Several steps in the methodology rely on the availability of a large amount of data. Examples of these steps are:

- step three, which develops a statistical model relating EOL output to the rest of EOL inspection variables
- step seven, which develops a large-scale model relating EOL output to influential process parameters. Moreover, MARS modeling technique needs more data than simple parametric regression methods to model a system.

Not all manufacturing environments are data-rich. Most data-poor environments suffer from a combination of the following two characteristics:

1. no data are available for a few variables
2. infrequent data samples are available for several process variables.

In our methodology, if EOL inspection data are unavailable, step three of the methodology will be unimplementable. If process variable data are unavailable, step seven of the methodology will be unimplementable. In addition, sparsely available data may compromise the confidence in the final results.

There may be ways to work around limited availability or unavailability of data in some situations. However, the methodology in its current form needs additional development to handle systems that lack enough data.

5.2.2 Model single (univariate) output

For most products, there is more than one characteristic important to a customer. Examples of these characteristics are functional accuracy, product reliability and product's appearance. The median and the variance of the same output variable often become a two-pronged response for understanding robustness issues. This thesis research has developed a new way to model "one" such characteristic as a function of process parameters. Using the model, we have discussed several control options to keep the characteristic close to target. These control options often involve changing processing conditions. While the new processing conditions promise to bring the characteristic, which we have modeled, close to target, they may also affect other characteristics, which we have not modeled.

A straight-forward approach is to model all important product characteristics as separate functions of process variables. Besides being tedious, this approach does not exploit the

dependence (or correlation) between different characteristics of a product. As such, it fails to provide an understanding of the trade-off between different characteristics of a product when the processing conditions are changed. A few modifications in our current methodology are discussed in Section 5.3. These modifications address the issue of modeling several outputs (or characteristics of a product) simultaneously as a function of process variables.

5.2.3 Class of models made possible by MARS

MARS is a very versatile modeling tool. It can develop generalized additive models. It also identifies important interactions between predictor variables. However, in one variable it cannot develop higher-order models than the order of the spline used for model development. This is because MARS cannot develop models showing interaction of one variable with itself. Note that the interaction of a variable with itself will give higher order models in that variable. Section 5.3 discusses ways to overcome this limitation.

5.3 Recommendations for future work

This thesis research has developed and verified some novel ideas. It has also resulted in some useful and promising results. However, there remain some unanswered questions and unverified ideas that emerged in the course of this research. They need further investigation. The recommended future work falls into the following categories:

1. develop improved models
   - include other classes of models in MARS
   - incorporate prior-information at other places in MARS
   - develop models for multiple outputs
   - develop models for categorical outputs
   - develop weighted regression models
2. develop confidence limits on model structure
3. develop models for variance
4. update models
5. evaluate non-proven competing opinions about processes.

5.3.1 Develop improved models

Models for manufacturing processes can be improved in several ways. The following paragraphs describe those ways.

Include other classes of models in MARS

MARS is currently incapable of developing quadratic (or higher order) models in one variable, see Section 5.2.3, if the order of the spline is only one. However, it can develop quadratic (or higher order) models if

- the higher order part appears as an expression in an engineering model or in an empirical model
- the order of the basis functions in MARS is chosen as two or higher

The engineering models and empirical models used in this research did not include quadratic terms. The basis functions were also just univariate basis functions.

Incorporate prior information at other places in MARS

This research has identified five places where MARS can incorporate prior information. However, it has explored and developed only one such option. The remaining four, listed below remain unexplored.

1. definition of split criteria in split nodes
2. choice of basis functions
3. choice of the order of basis functions
4. definition of pruning criteria

Further development of these options may give further improvements in the models for L_{en} and L_{ep}. Developing these options will also be a respectable contribution to the research literature in modeling.

Modeling multiple outputs

Currently, MARS models one output (or response) only. This places limitations on understanding trade-off when more than one product characteristic is important. This research has modeled only one output at a time (L_{en} or L_{ep}).

Multiple outputs can be modeled by developing a single statistic comprising all outputs. Such a statistic can have a varying amount of contribution from the outputs. Mathematical techniques used to develop such a statistic are principal components analysis, factor analysis, etc. These are linear techniques only. Alternatively, system knowledge can also be used to develop a statistic which may be a non-linear function of the outputs.

Include categorical outputs

The current thesis research modeled two continuous outputs (L_{en} and L_{ep}). It did not explore model development with categorical outputs. Categorical outputs are important in many industrial problems, particularly in modeling yield (in semiconductor manufacturing), defect rate (in a continuous photographic film-making process), etc. MARS can model categorical output too. It will be interesting to see if the issues in modeling a categorical output will be any different from those in modeling a continuous output, or if the results will be as remarkable as that of modeling L_{en} and L_{ep}.

Developing weighted regression models

We assumed that all predictor variables and observations are equally important for model development. In reality, some predictor variables are more reliable than others. This could be due to different sensor errors, different amount of stochastic noise associated with different predictor variables, dissimilar sampling frequency, etc. In addition, some observations are more reliable than others. It will be an interesting process to develop a way to attach different weights to different predictor variables and to different observations simultaneously. The use of such weights can help in performing sensitivity analysis [2].

5.3.2 Develop confidence limits on model structure

Using the methodology developed in this research, a model was developed for L_{en} and another for L_{ep}. While the variance on the predicted output was low (determined by the model standard error and by the generalized cross validation standard error), the confidence limits on the model structure have not yet been established. The establishment of such limits will help understand more clearly how well the current models for L_{en} and L_{ep} would be applicable to new data and process technologies.

5.3.3 Develop models for variance of the output

This thesis research has modeled the median value of L_{en}/L_{ep} as a function of influential in-lines. The variance of L_{en}/L_{ep} still remains unexplored. A model for variance is important to understand signal to noise ratio, and to identify robust process settings. With a better understanding about variance, the tradeoff between median and variance can be better understood. For example, precise control of the median may be at the expense of a high variation.

There are some interesting issues with modeling the variance. In general, physics-based models are for the median (or mean) value. They are rarely found for the variance. The understanding about variance is often qualitative by nature except in some cases where the understanding is developed by using numerical simulation tools. Quantitative results of numerical simulation tools are also often interpreted qualitatively. These issues make the model development of variance a difficult task by traditional physics-based modeling methods and by traditional statistical modeling methods. As such, these issues also render the variance a very interesting output to model by the methodology developed in this research.

Broadly speaking, the seven steps of the methodology will be applied in much the same way to model the variance as they would be applied to model the median (or the mean). A few interesting differences will be in the implementation of step three and step six, where the methodology relies on physics-based knowledge, empirical information and data. Empirical information will be used much more to model the variance than how much it was used to model the median (or the mean). Often, the variance depends on the median (or the mean). This is again verified empirically. The use of physics-based models in step six of the methodology can help test the dependence of variance on the median (or the mean).

5.3.4 Update models

Due to deliberate changes in a manufacturing process, and due to process maturity after the process is developed and transferred to high-volume manufacturing, the characteristics of the manufacturing line change. The model for the output has to account for such changes. Two types of modifications are needed in a model to account for such changes. They are:

1. **updating parameters** (or coefficients) of the model with new estimates. Several estimation techniques can be used to determine new parameter values. These techniques include, but are not limited to, bayesian techniques, time series analysis, etc.
2. **updating the structure of the model**. This involves ripping apart the old model and developing a completely new model. This type of update may be needed when an otherwise non-influential process parameter may become an important one after a change in the manufacturing process. This update may also be needed after a sufficiently long time to allow the process to mature.

An important consideration with both types of updates is the determination of the frequency of updates. In fact, updates for the two types can be done hierarchically. The model coefficients can be updated more frequently than the structure of the model.

5.3.5 Evaluate competing non-proven opinions about processes

When a manufacturing process step is not well understood, a set of process engineers may believe that the process step behaves in one way (say linearly). However, another set of engineers (or process developers) may believe that the process step behaves in a different way (say quadratically). It is important to determine the different operating conditions under which one opinion is more pertinent than another, and by how much one opinion is more pertinent that the other.

The use of prior information (engineering and empirical models) and data can help determine the different regions of operation in which one opinion is more important than another. Bayesian and Dempster-Shafer theory [66] can also be used to attach probability to the different opinions in different regions of operation. In addition, those probabilities can be updated periodically with the use of more data.

References

[1] M. R. Anderberg, *Cluster Analysis for Applications*: Academic Press, 1973.

[2] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression diagnostics: identifying influential data and sources of collinearity*: New York: Wiley, 1980.

[3] S. Bisgaard, "The use of Statistics to Improve Manufacturing Systems," Center for Quality and Productivity Improvement, University of Wisconsin, 610 Walnut Street, Madison, WI 53705 73, 1991.

[4] R. D. Blevins, *Formulas for natural frequency and mode shape*: Robert E. Krieger Publishing Company, 1984.

[5] G. Boothroyd, *Fundamentals of Metal Machining and Machine Tools*: Scripta Book Company, 1975.

[6] E. D. Boskin and C. Spanos, "A method for modeling the manufacturability of IC designs," presented at Proceedings of the IEEE International Conference on Microelectronic Test Structures, 1993.

[7] E. D. Boskin and C. J. Spanos, "IC performance prediction from electrical test measurements," presented at IEEE/SEMI International Semiconductor Manufacturing Science Symposium, 1992.

[8] G. Box, "Understanding Exponential Smoothing-A Simple Way to Forecast Sales and Inventory and Feedback Control by Manual Adjustment and Bounded Adjustment Charts," Center for Quality and Productivity Improvement, University of Wisconsin, 610 Walnut Street, Madison, WI 53705 71, 1991.

[9] G. E. P. Box and N. R. Draper, *Empirical model-building and response surfaces*: John Wiley & Sons, 1986.

[10] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control.* San Francisco: Holden-Day, 1976.

[11] L. Brieman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*: Wadsworth International Group, 1984.

[12] A. Buja, T. Hastie, and R. Tibshirani, "Linear smoothers and additive models," *The Annals of Statistics*, vol. 17, pp. 453-555, 1989.

[13] J. M. Chambers and T. J. Hastie, *Statistical Models in S*: Wadsworth & Brooks/Cole Advanced Books & Software, 1992.

[14] C. Chao and L. S. Milor, "Performance modeling using adaptive regression splines," *IEEE transactions on semiconductor manufacturing*, vol. 8, pp. 239-251, 1995.

[15] W. S. Cleaveland, *Visualizing Data*: Hobart Press, 1993.

[16] C. de Boor, *A Practical Guide to Splines*, vol. 27: Springer-Verlag, 1978.

[17] R. D. De Veaux, D. C. Psichogios, and L. H. Ungar, "A comparison of two non-parametric estimation schemes: MARS and neural networks," *Computers & Chemical Engineering*, vol. 17, pp. 819-837, 1993.

[18] S. D. Eppinger, C. D. Huber, and V. H. Pham, "A methodology for manufacturing process signature analysis," *Working Paper #3631-93-MSA. MIT, 50 Memorial Drive, Cambridge, MA 02139*, pp. 36, 1993.

[19] I. D. Faux and M. J. Pratt, *Computational Geometry for Design and Manufacture*: Ellis Harwood Limited, 1979.

[20] B. Flury, *Common principal components and related multivariate models*. New York: Wiley, 1988.

[21] J. H. Friedman, "Adaptive Spline Networks," Laboratory for computational statistics, Department of Statistics, Stanford University, Technical report 107, March 1991.

[22] J. H. Friedman, "Estimating functions of mixed ordinal and categorical variables using adaptive splines," Laboratory for computational statistics, Department of Statistics, Stanford University, Technical report 108, June 1991.

[23] J. H. Friedman, "Multivariate Adaptive Regression Splines (with discussion)," *The Annals of Statistics*, vol. 19, pp. 1-141, 1991.

[24] J. H. Friedman, "Fast MARS," Laboratory for computational statistics, Department of Statistics, Stanford University, Technical Report 110, May 1993.

[25] P. E. Gise and R. Blanchard, *Semiconductor and Integrated Circuit Fabrication Techniques*: Reston Publishing Company, Inc., 1979.

[26] R. D. Gordon, "How can industrial data be used to achieve continuous improvement?," : (Statistician, Microcomputer Component Group) Intel Corporation. Personal communication, 1994.

[27] D. E. Hardt, "Real-Time Process Control: Limits to Progress," *Submitted to ASME Manufacturing Reviews*, 1990.

[28] T. Hastie, "Multivariate adaptive regression splines," : Computer program, Stanford University, 1995.

[29] T. Hastie and R. Tibshirani, "Generalized additive models (with discussion)," *Statistical Science*, vol. 1, pp. 297-318, 1986.

[30] R. V. Hogg and J. Ledolter, *Engineering Statistics*: Macmillan Publishing Company, 1987.

[31] J. P. Holman, *Heat transfer*, 5th ed: McGraw-Hill Book Company, 1981.

[32] F. Hsu, R. S. Muller, C. Hu, and P. Ko, "A simple punchthrough model for short-channel MOSFET's," *IEEE transactions of electron devices*, vol. ED-30, pp. 1354-1359, 1983.

[33] S. Hu, Y. B. Chen, and S. M. Wu, "Multi-output modal parameter identification by vector time-series modeling," *Design Engineering Division, ASME*, vol. 18, pp. 259-265, 1989.

[34] P. J. Kempthorne and M. Vyas, "Risk Measurement in Global Financial Markets with Asynchronous, Partially Missing Price Data," Massachusetts Institute of Technology April 1994.

[35] J. K. Kibarian, D. A. Hanson, and K. W. Michaels, "Using semiconductor physics to predict and analyze device distribution and yield," PDF Solutions, 425 North Craig, Suite 501, Pittsburgh, PA 15213 1995.

[36] C. C. Kontoes, D. Rokos, G. G. Wilkinson, and J. Megier, "The use of expert system and supervised relaxation techniques to improve spot image classification using spatial context," *IEEE Transactions*, pp. 1855-1858, 1991.

[37] P. A. Lachenbruch, *Discriminant Analysis*: Hafner Press, 1975.

[38] J. D. Layne, "Adaptive spline networks for estimating camera control parameters in robot-vision system," presented at Third workshop on neural networks: Academic/Industrial/NASA/Defense, 1992.

[39] S. Leang and C. Spanos, "Application of feed-forward control to a lithography stepper," presented at IEEE/SEMI International Semiconductor Manufacturing Science Symposium, 1992.

[40] S. F. Lee and C. J. Spanos, "Prediction of Wafer State After Plasma Processing Using Real-Time Tool Data," *IEEE Transactions on Semiconductor Manufacturing*, vol. 8, pp. 252-261, 1995.

[41] P. A. W. Lewis and J. G. Stevens, "Smoothing time series for input and output analysis in system simulation experiments," presented at Proceedings of the 1990 Winter Simulation Conference, 1990.

[42] P. A. W. Lewis and J. G. Stevens, "Nonlinear modeling of time series using Multivariate Adaptive Regression Splines (MARS)," *Journal of the American Statistical Association*, vol. 86, pp. 864-877, 1991.

[43] K. Lin and C. Spanos, "Manufacturing: an application for LPCVD," *IEEE transactions on semiconductor manufacturing*, vol. 3, pp. 216-229, 1990.

[44] K. Lin and C. J. Spanos, "Statistical equipment modeling for VLSI: Application for LPCVD," *IEEE Transactions on Semiconductor Manufacturing*, vol. 3, pp. 216-229, 1990.

[45] R. Lippmann, "An Introduction to Computing with Neural Nets," in *IEEE ASSP Magazine*, 1987, pp. 4-22.

[46] T. J. Lorenzen, "Setting realistic process specification limits-A case study," General Motors Research Laboratories, Warren, MI 48090-9057, Research Publication GMR-6389, August 15, 1988.

[47] J. M. Lucas, "Achieving a robust process using response surface methodology," Du Pont Company, Wilmington, Delaware 19898 1992.

[48] J. F. MacGregor and T. Kourti, "Statistical process control of multivariate processes," *Control Engineering Practice*, vol. 3, pp. 403-414, 1995.

[49] Splus Manual, "Splus for windows manual. Statistical analysis in Splus. Chapter 6," . Seattle: Mathsoft, Inc., 1993.

[50] A. Marazzi, *Algorithms, Routines, and S functions for Robust Statistics*: Wadsworth & Brooks/Cole Statistics/Probability Series, 1992.

[51] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*: Academic Press, 1992.

[52] G. P. McCabe, "Principal variables," *Technometrics*, vol. 26, pp. 137-143, 1984.

[53] G. P. McCabe, "Prediction of principal components by variable subsets," Purdue University 86-19, June 1986 1986.

[54] G. P. McCabe Jr., "Computations for variable selection in discriminant analysis," *Technometrics*, vol. 17, pp. 103-109, 1975.

[55] L. F. M. Meeting, "LFM review meeting at Massachusetts Institute of Technology," , 1993.

[56] D. C. Montgomery, *Introduction to Statistical Quality Control*, 2nd ed: John Wiley & Sons, 1990.

[57] D. C. Montgomery, *Design and Analysis of Experiments*, 3rd ed: John Wiley & Sons, 1991.

[58] F. T. Morse, *Power Plant Engineering*: East-West Press, 1953.

[59] R. H. Myers, *Classical and Modern Regression with Applications*: Prindle, Weber & Schmidt (PWS), 1986.

[60] J. B. Neuhardt, "Effects of correlated sub-samples in statistical process control," *IIE Transactions*, vol. 19, pp. 208-214, 1987.

[61] K. Ogata, *Modern Control Engineering*: Prentice-Hall of India Private Limited, 1989.

[62] D. C. Psichogios, R. D. De Veaux, and L. H. Ungar, "Non-parametric system identification: A comparison of MARS and neural networks," presented at Proceedings of the 1992 American Control Conference, Chicago, 1992.

[63] C. H. Reinsch, "Smoothing by spline functions," *Numerische Mathematik*, vol. 10, pp. 177-183, 1967.

[64] E. Sachs, R. Guo, S. Ha, and A. Hu, "Process control system for VLSI fabrication," *IEEE Transactions on semiconductor manufacturing*, vol. 4, pp. 134-144, 1991.

[65] E. Sachs, G. H. Prueger, and R. Guerrieri, "An equipment model for polysilicon LPCVD," *IEEE Transactions on Semiconductor Manufacturing*, vol. 5, pp. 3-13, 1992.

[66] G. Shafer, *A mathematical theory of evidence*: Princeton University Press, 1976.

[67] S. Sharifzadeh, J. R. Koehler, A. B. Owen, and J. D. Shott, "Using simulators to model transmitted variability in IC manufacturing," *IEEE transactions on semiconductor manufacturing*, vol. 2, pp. 82-93, 1989.

[68] V. Sharma, "Determining E-Tests Influencing Wafer-Level Median Values of Len and Lep," Intel Corporation Tuesday, September 6 1995.

[69] V. Sharma, "Identifying E-Test Parameters Statistically Related to Wafer-level Var(Le)," Intel Corporation Friday, August 18 1995.

[70] V. Sharma, "Methodology to extract important process steps influencing end-of-line (EOL) output parameters," Intel Corporation Sunday, November 26 1995.

[71] V. Sharma, "Understanding Variation in Channel-Length as a Function of Variation in all Other Short Package E-Test Parameters," Intel Corporation Sunday, July 30 1995.

[72] J. E. Shigley and L. D. Mitchell, *Mechanical Engineering Design*, 4th ed: McGraw-Hill Company, 1985.

[73] C. J. Spanos, H. Guo, A. Miller, and J. Levine-Parrill, "Real-time statistical process control using tool data," *IEEE transactions on semiconductor manufacturing*, vol. 5, pp. 308-318, 1992.

[74] C. J. Spanos, S. Leang, and S. Lee, "A Control & Diagnostics Scheme for Semiconductor Manufacturing," presented at American Control Conference, San Francisco, 1993.

[75] K. Stoddard, P. Crouch, M. Kozicki, and K. Tsakalis, "Application of feed-forward and adaptive feedback control to semiconductor device manufacturing," presented at Proceedings of the American Control conference, Baltimore, Maryland, 1984.

[76] W. R. Sype, "State of multivariate statistics application in semiconductor industry," : Personal communication, 1996.

[77] N. Takezawa, "An improved method for establishing the process-wise quality standard," *Rep. Stat. Appl. Res., JUSE*, vol. 27, pp. 63/1-12/74, 1980.

[78] H. Tong, B. Thanoon, and G. Gudmundsson, "Threshold Time Series Modeling of Two Icelanding Riverflow Systems," *Water Resources Bulletin*, vol. 21, pp. 651-660, 1985.

[79] N. Tredennick, "Technology and Business: Forces driving microprocessor evolution," *Proceedings of the IEEE*, vol. 83, pp. 1641-1652, 1995.

[80] Y. Tsividis, *Operation and modeling of the MOS transistor*: McGraw-Hill Book Company, 1987.

[81] G. J. Van Wylen, *Thermodynamics*: John Wiley & Sons, Inc., 1959.

[82] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S-Plus*: Springer-Verlag, 1994.

[83] W. W. S. Wei, *Time Series Analysis: Univariate and Multivariate Methods*: Addison-Wesley, 1990.

[84] R. E. Welsch, "How can principal components regression use the response and all inputs for a more accurate model?," : Personal communication, 1996.

[85] D. A. White. In-situ wafer uniformity estimation using principal component analysis and function approximation methods. MS, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1995.

[86] F. M. White, *Fluid Mechanics*, 2nd ed: McGraw-Hill Publishing Company, 1986.

[87] S. Wold, A. Ruhe, H. Wold, and W. J. Dunn Hi, "The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses," *SIAM Journal of Sci. Stat. Comput.*, vol. 5, pp. 735-743, 1984.

[88] S. Wolf and R. N. Tauber, *Silicon Processing for the VLSI Era*, vol. 1: Lattice Press, 1986.

[89] K. S. Wong. Sequential Optimization Using Grouped Variables. M.S., Massachusetts Institute of Technology, 1990.

[90] D. A. Wrangham, *The theory and practice of heat engines*, 2nd ed: The English Language Book Society and Cambridge University Press, 1962.

[91] S. M. Wu and S. Hu, "Impact of 100% in-process measurement on statistical process control (SPC) in automobile assembly," *Monitoring and control of manufacturing processes, ASME, Production Engineering Division*, vol. 44, pp. 443-448, 1989.

[92] R. E. York. Distributed gauging methodologies for variation reduction in the automotive body shop. L.F.M. Thesis, Massachusetts Institute of Technology, 1995.

Appendix A

This appendix aims to show an example of the inputs to and the outputs from the software that develops MARS models [28]. The objective of the appendix is not to develop an accurate MARS model. However, the appendix demonstrates an input file to the software that develops MARS model, and interprets the output of the software. The example chosen here is only a fictitious one. (Friedman has discussed MARS through many examples [23].)

Figure 36 shows a manufacturing line with two process steps connected serially. The two process variables are X_1 and X_2, and the output is Y. Equation 18 shows an assumed physics-based model relating X_1 and X_2. Equation 19 shows an assumed empirical relationship between X_1 and X_2. An assumed model relating Y to X_1 and X_2 is shown by Equation 20, and is used solely for the purpose of creating data for Y. A random normal error structure is assumed on the output (Y), as shown by Equation 20. This section attempts to develop a MARS model relating Y to X_1 and X_2.

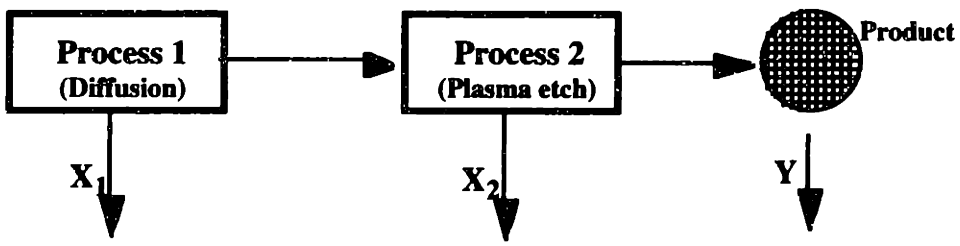

Figure 36. A manufacturing line with two process steps

$$M_1 = \frac{X_1}{X_2} + 2$$

Equation 18. Physics-based model relating X_1 and X_2

$$M_2 = 3X_1 + 4X_2$$

Equation 19. Empirical model relating X_1 and X_2

$$Y = 2\frac{X_1}{X_2} + 5X_1 - 3X_2 + \varepsilon, \quad \varepsilon \sim N(0, 0.01)$$

Equation 20. Assumed model relating Y to X_1 and X_2

Data were generated for X_1 and X_2 by creating fifty random numbers for X_1 and by creating another fifty for X_2, as shown in Table 10.

Output	Process variable	Process variable	Engineering model	Empirical model
(Y)	(X_1)	(X_2)	(M_1)	(M_2)
4.48	0.01	-1.56	1.99	-6.20
-4.06	-0.04	1.27	1.97	4.97
-11.07	-1.02	1.54	1.34	3.13
-3.53	-0.13	0.80	1.84	2.82
1.09	-0.36	-0.59	2.62	-3.42
-3.04	-0.03	0.89	1.96	3.45
-0.80	-1.86	-2.36	2.80	-15.08
-1.71	0.34	1.27	2.27	6.09
3.28	0.00	-1.11	2.00	-4.43
8.62	1.21	0.56	4.14	5.87
-0.96	-0.02	0.25	1.92	0.92
-12.68	-1.01	0.29	-1.47	-1.87
3.54	0.92	0.99	2.93	6.70
-34.56	-1.38	0.10	-11.66	-3.74
-6.33	-0.47	0.30	0.45	-0.19
-40.02	-0.80	0.04	-15.93	-2.23
1.27	0.90	1.44	2.63	8.45
2.45	-1.16	-2.46	2.47	-13.30
-2.08	0.10	0.91	2.12	3.94
0.37	0.23	0.51	2.45	2.73
10.90	2.40	-1.12	-0.14	2.71
-0.84	0.08	0.51	2.16	2.30
-2.06	-0.02	0.58	1.96	2.25
-2.46	0.75	2.29	2.33	11.40
28.60	-1.11	-0.07	19.02	-3.58
-3.45	-2.23	-0.90	4.48	-10.28
7.60	1.23	0.69	3.78	6.43
6.00	1.56	-0.76	-0.05	1.63
1.40	-0.52	-1.02	2.51	-5.66
-1.51	0.42	1.35	2.31	6.65
-6.13	-0.31	1.36	1.77	4.52
-3.50	0.56	2.25	2.25	10.67
13.09	2.69	-1.27	-0.12	3.00
15.94	1.09	0.20	7.51	4.07
-6.88	0.10	2.48	2.04	10.23
1.36	-0.92	-1.60	2.57	-9.17
-15.30	-1.76	1.35	0.70	0.12
2.30	0.30	-0.58	1.48	-1.42
1.01	-0.52	-0.49	3.08	-3.52
7.60	1.47	0.96	3.52	8.28
2.33	0.45	-0.56	1.19	-0.89
-1.33	0.41	1.32	2.31	6.52
4.04	0.54	-0.87	1.39	-1.89
5.52	0.08	-1.70	1.96	-6.57
5.41	0.32	-1.42	1.77	-4.72
-12.58	-1.35	0.80	0.30	-0.87
8.12	-2.42	-0.25	11.74	-8.26
3.23	0.34	-0.83	1.58	-2.28
16.64	2.46	0.75	5.29	10.39
17.18	2.99	1.10	4.72	13.36

Table 10. Inputs to MARS software

The second and the third columns in Table 10 show the data for X_1 and X_2 respectively. Data for X_1, X_2 and Equation 18 were used to generate data for M_1. Data for X_1, X_2 and

Equation 19 were used to generate data for M_2. The fourth and the fifth columns in Table 10 show the data for M_1 and M_2 respectively. Data for X_1, X_2 and Equation 20 were used to generate data for Y. The first column in Table 10 shows the data for Y.

The software used for MARS model development was in the Splus environment. The command used to develop MARS model is shown by Equation 21.

```
app.mars <- mars(x=app.in, y=app.out, degree=2)
```

Equation 21. Command to develop MARS model

In Equation 21

- app.mars contains the output of the routine "mars" that develops MARS models.
- mars is the routine that develops MARS models. (There are other arguments than the ones outlined here that a user can give to the mars routine. However, the other arguments are not described here because mars uses default values in the absence of user-supplied ones. The default values of those arguments were used here.)
- x is the data matrix that contains the predictor variables. In the example considered here, x consists of the last four columns of data in Table 10, and is called app.in.
- y is a vector of outputs. In the example, y consists of the first column of data in Table 10, and is called app.out.
- degree is the maximum degree of allowable interaction between different predictor variables. In the example, it was chosen as two.

The "mars" routine has multiple outputs for a MARS model. However, only a few of those are necessary to interpret the model. These part of the output include:

1. **factor**. This is a matrix. The number of columns in factor is the same as that of the input matrix x. Table 11 shows the "factor" matrix for the case example.

	X_1	X_2	M_1	M_2
[1,]	0	0	0	0
[2,]	0	0	1	0
[3,]	0	0	-1	0
[4,]	1	0	0	0
[5,]	-1	0	0	0
[6,]	0	0	0	1
[7,]	0	0	0	-1

Table 11. "factor" matrix for MARS model

The rows in Table 11 represent the different basis functions. The value zero in a particular row and column means that the predictor variable represented by that column does not contribute to the basis function represented by that row. For example, no variable contributes to the first basis function as illustrated by the first row in Table 11. The first row represents the constant term in the MARS model. The presence of a "1" or a "-1" in a row implies that the variable represented by the column contributes to the basis function represented by the row. The second row in Table 11 contains a "1" in the column under M_1. This means that M_1 contributes to the second basis function. The significance of "1" and "-1" is explained in a later paragraph.

Notice that all rows in the "factor" matrix (except the first row) have only one non-zero element. This implies that there are no interaction terms in any basis function. The "mars" routine did develop MARS models with two-way interactions, as shown by Equation 21. However, the MARS model without any interaction term was chosen because it was more accurate than the one which included two-way interaction terms.

2. **cuts**. This is a matrix. Its dimensions are the same as those of the "factor" matrix. Table 12 shows the "cuts" matrix for the case example. The rows represent the different basis functions and the columns represent the different predictor variables. In Table 12, the four columns represent X_1, X_2, M_1 and M_2 respectively.

	[,1]	[,2]	[,3]	[,4]
[1,]	0.000	0.000	0.000	0.000
[2,]	0.000	0.000	1.386	0.000
[3,]	0.000	0.000	1.386	0.000
[4,]	0.454	0.000	0.000	0.000
[5,]	0.454	0.000	0.000	0.000
[6,]	0.000	0.000	0.000	-1.868
[7,]	0.000	0.000	0.000	-1.868

Table 12. "cuts" matrix for MARS model

The "cuts" matrix gives the cut-off values of the predictor variables that contribute to different basis functions. As such, only a few positions in the "cuts" matrix are non-zero. These positions have a "1" or a "-1" in the "factor" matrix. Compare Table 11 and Table 12. Since the first row represents the constant term in the model, it contains only zeroes in both matrices. The second row has a cut-off value of 1.386 for M_1 with a "1" in the second row of the "factor" matrix. The "1" (without the minus sign) means that the second basis function contributes only for values of M_1 greater than 1.386. The third row has a cut-off value of 1.386 for M_1 with a "-1" in the third row of the "factor"

matrix. The "-1" means that the third basis function contributes only for values of M_t less than 1.386.

3. **selected.terms**. This is a vector. It contains the row numbers in the "factor" matrix (which are the same as those in the "cuts" matrix) that become part of the final MARS model. The number of elements in the "selected.terms" vector is less than or equal to the number of rows in the "factor" matrix. Table 13 shows the selected terms in the case example. All basis functions contribute to the final MARS model in the example considered here.

[1] 1 2 3 4 5 6 7

Table 13. Selected terms in the final MARS model

4. **coefficients**. This is a vector. It contains the coefficients that multiply with the different basis functions. The number of coefficients is the same as the number of elements in the selected.terms vector. Table 14 shows the coefficients in the case example. The first coefficient multiplies with the first basis function. It turns out to be the model constant. The second coefficient multiplies with the second basis function and so on.

	Coefficients
[1,]	3.404
[2,]	1.997
[3,]	-2.001
[4,]	7.302
[5,]	-7.205
[6,]	-0.751
[7,]	0.745

Table 14. Coefficients in the final MARS model

5. **residuals**. This is a vector of residuals. The number of elements in "residuals" is that same as the number of observations used to develop the MARS model. In the case example, "residuals" has fifty elements. These residuals are later used to determine the standard error of the MARS model.

The final MARS model can now be constructed with the use of Table 11 through Table 14. The model is constructed by adding the selected basis functions in the "factor" matrix. Since the first row is a selected term, and it does not have any predictor variables, only the coefficient contributes to the model. The coefficient value is 3.404 from Table 14. The second

basis function has contribution from M_1, see Table 11, only when M_1 is greater than 1.386. This is mathematically represented as $[M_1 - 1.386]_+$. The coefficient value 1.997 multiplies with the basis function. The resulting expression is $1.997[M_1 - 1.386]_+$. This expression is added to the model constant 3.404. The next basis function is from the third row of the "factor" matrix. Here, M_1 contributes to the MARS model, only for values less than 1.386. This is mathematically represented as $[M_1 - 1.386]_-$. The multiplying coefficient for the third basis function is -2.001. The total contribution of the third basis function to the final MARS model is $-2.001[M_1 - 1.386]_-$. The contributions of the rest of the basis functions are also calculated likewise. These contributions are added together to construct the final MARS model. The standard error is calculated from the residuals and added to the MARS model. Equation 22 shows the final MARS model for the case example.

$$Y = 3.404 + 1.997[M_1 - 1.386]_+ - 2.001[M_1 - 1.386]_- + 7.302[X_1 - 0.454]_+ - 7.205[X_1 - 0.454]_- - 0.751[M_2 - (-1.868)]_+ + 0.745[M_2 - (-1.868)]_- + \varepsilon, \quad \varepsilon \sim N(0, 0.0061)$$

Equation 22. Final MARS model for the case example

By substituting the expressions for M_1 and M_2 from Equation 18 and Equation 19 respectively in Equation 22, Y can be expressed purely in terms of the process variables X_1 and X_2.