

## 1. Introduction

- (a) Problem Statement
  - Background of Semiconductor Manufacturing
  - Mathematical Model of the Process
  - Semiconductor Manufacturing Data Collection
- (b) **Applications**  
Motivation for research and applications related to the subject.
- (c) **Organization**  
Explain organization of the report, what is included, and what is not.
- (d) **Problem Statement**
  - i. Background of Semiconductor Manufacturing
  - ii. Mathematical Model of the Process
  - iii. Semiconductor Manufacturing Data Collection and Storage

## 2. Literature Survey

- (a) Semiconductor Modeling
- (b) Inconsistent Data
- (c) Large Number of Variables Relative to the Sample Size
- (d) Complexity, Sample Size, and Modeling
- (e) High Dimensional Modeling and Visualization
- (f) Reduction of dimensionality
- (g) Alternate high-dimensional domains and approaches
- (h) Chemometrics
- (i) Text Mining
- (j) **Meta-Analysis in Medical Research**  
Overall and specifics
  - i. Fixed Effect Models  $Y_i = \hat{Y}$
  - ii. Random Effect Models  $Y_i = \hat{Y} + e_i$
  - iii. Random Regression Models  $Y_i = \hat{Y}_{i1} + \hat{Y}_{i2} + \dots + \hat{Y}_{ip} + e_i$

## 3. Methodology

- (a) **Hierarchical Modeling**
  - i. Subcomponents

If a large system can be decomposed into meaningful subunits, understanding may be gained by examining the relationships of subunits to the high level system outputs and between subunits. If users of the system conceive of the system as an agglomeration of subsystems, a modeling methodology that explains the interdependencies and effects of the subcomponents may help build understanding of the system.

- ii. Intermediate variables upper bound the prediction with their accuracy

**For instance:** if the intermediate monitoring variables are used as the sole predictor of the higher level system output, the resulting predictions can only be as good as the best prediction based on the intermediate variables. Several studies have identified an increase in US citizen's heights with an increase in nutrition; using height as a sole surrogate variable to summarize the nutritional variables in prediction of health will limit the results to the information present in the height variable. The intermediate variables can be used, but they should be augmented with other information to produce an improved estimate.

- iii. End result variables

(b) **Meta-Analysis**

- i. risk versus odds

**Risk versus Odds** Two methods of characterizing process losses are an odds and as risk. Odds of failure ( $\frac{\text{failures}}{\text{successes}}$ ) are a simple ratio of failures to successes with the domain  $(-\infty, +\infty)$ , while risks are a ratio of ( $\frac{\text{failures}}{\text{successes} + \text{failures}}$ ) and have the domain  $[0, 1]$ . In a manufacturing process, risks or failure rate might be more commonly used, but for modeling purposes, odds has better numerical properties.

- ii. Random Regression Model of log risk defect  $Y_{risk\ i} = \hat{Y}_{i1} + \hat{Y}_{i2} + \dots + \hat{Y}_{ip} + e_i$
- iii. Random Regression Model of log odds defect  $Y_{odds\ i} = \hat{Y}_{i1} + \hat{Y}_{i2} + \dots + \hat{Y}_{ip} + e_i$

(c) **Proposed Model**

- i. Subcomponents as a risk factor
  - A. Risk, defect rate, and odds are functions of each other
  - B. Baseline risk is the overall, or intercept term.

- C. Equally contributing components will individually have low predictive power, but taken together will have stronger predictive power.

**Improvements due to combining submodels:** Supposing that the estimate of the defect rate due to some submodel  $M_i$   $\hat{y}_i = f_i(\vec{\beta}_i, \vec{X}_i)$  where  $\vec{X}_i$  is a vector of variables in submodel  $i$  and  $\vec{\beta}_i$  is a vector of parameters, and  $f_i()$  is an estimation model, the model  $\hat{y}_i$  will be able to predict the system output with some level of accuracy. Supposing further that  $i = 10$  subsystems are independent and that each is responsible for  $1/20$  of the variance observed in the output, with the remaining  $1/2$  of the variance due to other sources. Each of the 10 models can explain at most 5% of the variation, and may seem insignificant, but a combination of the subsystems has the potential to explain  $1/2$  of the variance in the entire system.

If the entire system is small enough to be modeled in one system, that would assuredly produce better prediction results, however, segmentation into subcomponents can perhaps aid in the interpretation of the resulting model.

Interdependencies between the subsystems can be reflected in the correlations between the subsystem outputs.

- ii. Comparison to intermediate surrogate variables
  - A. (E.g. TEG as an intermediary)
- iii. Continuity Adjustment ( $\pm e$ )
- iv. Combination of models and relative

(d) **Research Plan**

4. **Results / Application**

- (a) Semiconductor modeling

5. **Conclusions**