Master Thesis

Optimization of Process Corrections

by

Yixin Lu

Supervisor:
Prof.dr. R.M.M. Mattheij (TU/e)

Advisors:
Ir. Martijn Slob (TU/e)
Dr.ir. Erik Wallerbos (ASML)

Eindhoven, August 2008
Summary

This report describes the problem, findings and results of my final project at ASML, the world’s leading provider of lithography systems. The aim is to develop algorithms that enable the automatic determination of optimal process correction settings for overlay, which is defined to be the displacement of an exposed point on the wafer with respect to the position of that same point in the previously exposed layer. In order to reduce the on-product overlay, the overlay of exposed wafers is measured and these measurements are used for modeling. The calculated corrections (process corrections) are applied to subsequent layers when the same product layer is exposed.

The most accurate process correction mechanism is the application of corrections per exposure (CPE). However, this requires a lot of measurement time on an overlay metrology tool. To reduce the measurement effort, higher order process corrections (HOPC) are introduced, which are polynomial approximations of CPE. Now the question is how to achieve an acceptable overlay with a minimum measurement effort. With the current GridMapper tool (a PC-based tool), one has to manually specify the overlay targets to be measured, parameters for modeling, and the polynomial degree for each parameter to be modeled. Then the expected overlay is computed by the PC tool. This tedious and time-consuming process has to be repeated till an optimal setting is found. Therefore, it is desirable to have an algorithm that can do this automatically.

During the course of this project, we first investigated algorithms to automatically select overlay targets to be measured. The goal is to select input points so as to maximally reduce the prediction uncertainty. The analysis of the covariance matrix involved in the HOPC model yielded three important guidelines: zero-mean, symmetry and spatial coverage. These guidelines have been included in the geometry-based overlay target selection algorithm. As an alternative, also a data-driven selection algorithm has been developed. In this algorithm, overlay targets are grouped according to a similarity rule. The validity of both sampling strategies has been verified by empirical tests using data from different fabrication facilities. With the current HOPC models, the geometry-based sampling is the recommended strategy, for its stability and efficiency.

Then, in order to avoid over-fitting with higher order models, several methods have been investigated for selecting appropriate degrees and the relevant terms. Besides two classical methods described in the literature (adjusted $R^2$ and partial sum of squares), a new method, namely, ANOVA-like approach (ALA) has been developed. The effectiveness and stability of these methods have been verified by empirical tests. According to these
test results, ALA is the most successful method.

Therefore, it is recommended to use the geometry-based sampling together with the model selection strategy ALA in order to automatically determine optimal HOPC settings that yield an acceptable overlay.

Finally, also some non-parametric models have been investigated, and their potential benefits have been discussed. Unfortunately, due to the limited time available for the final project, it was not possible to implement and test these non-parametric models. This can be left to the future.
Acknowledgement

I would like to first thank Professor Mattheij for giving me this nice project. I am also thankful to Martijn Slob and Erik Wallerbos for their support, encouragement, and invaluable guidance throughout the course of this project.

Thanks for Joost Smits’s arrangement and support, I had a great time working at ASML with so many brilliant people. My colleagues are so friendly and helpful to me that I would have felt guilty if I didn’t work hard. This internship is definitely one of the most precious experiences for me.

I should also mention my graduate studies in both TU Kaiserslautern and TU/e. I’ve learnt a lot in the past two years and I would like to thank all the lecturers who taught me, all the people who helped me.

Of course, I am grateful to my parents for their love and patience. They are the best of the best parents in the world. I also feel lucky to have a super boyfriend, who is always there encouraging me and sharing his programming skills with me.

Finally, I want to thank the following: Patricio, Oliver (for their friendship in the past two years); Jan Portegijs, Robin Tijssen, Remco Beckers, Geert Simons (for their constant help at ASML); Martijn Antonissen (nice coordinator of our master program); Enna (kind and capable secretary of CASA).

Eindhoven
Yixin Lu
July 20, 2008
List of Tables

2.1 Offsets per exposure ................................................................. 9

C.1 Selected polynomial degrees by adjusted $R^2$ (15 fields measured) ........ 55
C.2 Selected polynomial degrees by adjusted $R^2$ (21 fields measured) ........ 55
C.3 Selected polynomial degrees by adjusted $R^2$ (full-wafer measured) ..... 56
C.4 Coefficients for $T_x, T_y, R_s, R_a, M_s$ and $M_a$ .......................... 56
C.5 Selected terms by PSS and PSS with Zernikes (15 fields measured) ....... 57
C.6 Selected terms by PSS and PSS with Zernikes (21 fields measured) ....... 57
C.7 Selected terms by PSS and PSS with Zernikes (full-wafer measured) ..... 57
C.8 Selected polynomial degrees with ALA (15 fields measured) ............... 58
C.9 Selected polynomial degrees with ALA (21 fields measured) ............... 58
C.10 Selected polynomial degrees with ALA (full-wafer measured) ............ 58
# List of Figures

1.1 How semiconductors are made ........................................ 1
1.2 Intra-field errors: symmetric rotation, asymmetric rotation, symmetric magnification, asymmetric magnification ..................... 2
1.3 Inter-field errors: translation, rotation and scaling ............... 3
1.4 Application of process correction in the production loop .......... 3

2.1 Coordinate system of overlay data ................................. 8

3.1 Geometry-based sampling patterns .................................. 18
3.2 Overlay after “individual” corrections with geometry-based sampling .......................................................... 18
3.3 Overlay after “sub-lot” corrections with geometry-based sampling .......................................................... 19
3.4 Different sampling plans .............................................. 19
3.5 Overlay after “individual” corrections with different sampling patterns .......................................................... 20
3.6 Data-driven sampling patterns ..................................... 21
3.7 Overlay after “individual” corrections with data-driven sampling .......................................................... 22
3.8 Overlay after “sub-lot” corrections with data-driven sampling .......................................................... 22

4.1 Overlay after corrections with different models ................ 26
4.2 Comparison: overlay after full- and reduced($R^2$)-model corrections .......................................................... 28
4.3 Overlay improvements after CPE and adjusted $R^2$ (15 fields) .... 28
4.4 Overlay improvements after CPE and adjusted $R^2$ (21 fields) .... 29
4.5 Overlay improvements after CPE and adjusted $R^2$ (full-wafer) ... 29
4.6 Forward Stepwise Selection ........................................... 31
4.7 Overlay after “individual” full-and reduced(PSS)-model corrections .......................................................... 32
4.8 Overlay improvements after CPE and PSS (15 fields) ............ 32
4.9 Overlay improvements after CPE and PSS (21 fields) ............ 32
4.10 Overlay improvements after CPE and PSS (full-wafer) ............ 33
4.11 Overlay improvements after CPE and PSS with Zernikes (15 fields) .... 34
4.12 Overlay improvements after CPE and PSS with Zernikes (21 fields) .... 34
4.13 Overlay improvements after CPE and PSS with Zernikes (full-wafer) .... 35
4.14 Overlay after “individual” full- and reduced(ALA)-model corrections .......................................................... 36
4.15 Overlay improvements after CPE and ALA (15 fields) ........... 37
4.16 Overlay improvements after CPE and ALA (21 fields) ........... 37
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.17</td>
<td>Overlay improvements after CPE and ALA (full-wafer)</td>
<td>37</td>
</tr>
<tr>
<td>4.18</td>
<td>Overlay after “sub-lot” corrections with different methods</td>
<td>38</td>
</tr>
<tr>
<td>A.1</td>
<td>overlay measurement</td>
<td>49</td>
</tr>
<tr>
<td>C.1</td>
<td>Overlay improvements after CPE and ALA</td>
<td>59</td>
</tr>
<tr>
<td>C.2</td>
<td>Overlay improvements after CPE and ALA</td>
<td>59</td>
</tr>
</tbody>
</table>
# Contents

Summary iii  
Acknowledgement v  
List of Tables v  
List of Figures vii  

## 1 Introduction
1.1 Problem Description ............................................. 2  
1.2 Goals ...................................................................... 4  
1.3 Methodology .......................................................... 4  
1.3.1 Experimental Design ............................................. 5  
1.3.2 Regression Analysis ............................................. 5  

## 2 Higher Order Process Corrections
2.1 Notations and Data Description ................................. 7  
2.1.1 Notations ............................................................ 7  
2.1.2 Data Description .................................................. 7  
2.2 Model Derivation ...................................................... 8  
2.2.1 Basic Assumptions ............................................... 8  
2.2.2 CPE Model ........................................................ 9  
2.2.3 Standard Model .................................................. 9  
2.2.4 HOPC Models .................................................... 10  

## 3 Optimal Distribution of Overlay Targets
3.1 Geometry-based Pattern .......................................... 13  
3.1.1 Basic Principles .................................................. 13  
3.1.2 Ill-conditioning and Centering Effects ..................... 15  
3.1.3 Algorithms and Experimental Results ..................... 17  
3.2 Data-driven Pattern ................................................ 20
### Contents

3.2.1 General Description ........................................... 20  
3.2.2 Experimental Results ........................................ 21  
3.3 Remarks and Further Discussion ............................... 22  

4 Selection of Polynomial Degree and Relevant Terms .......... 25  
4.1 Motivation .......................................................... 25  
4.2 Adjusted $R^2$ ....................................................... 27  
4.3 Partial Sum of Squares ......................................... 29  
  4.3.1 Basic Principles ............................................... 30  
  4.3.2 Forward Selection ........................................... 30  
  4.3.3 Experimental Results: ...................................... 31  
  4.3.4 Modified Partial Sum of Squares with Zernike Polynomials .... 33  
4.4 ANOVA-Like Approach ........................................... 35  
  4.4.1 General Description .......................................... 35  
  4.4.2 Experimental Results ...................................... 36  
4.5 Concluding Remarks ............................................. 38  

5 Non-Parametric Models .............................................. 41  
5.1 Background ........................................................ 41  
5.2 An Overview of Nonparametric Models ......................... 41  
5.3 Potential Benefits ............................................... 42  

6 Conclusion and Recommendation ................................ 45  
6.1 Achievements ..................................................... 45  
6.2 Future Work ...................................................... 46  

Bibliography ............................................................. 47  

A Further Explanation of Metrology Tools ....................... 49  

B Pseudo Codes ....................................................... 51  

C Experimental Results ............................................. 55
Chapter 1

Introduction

In semiconductor manufacturing, integrated circuits (ICs) are fabricated with a number of physical and chemical processes performed on a wafer made of pure semiconducting material, see Figure 1.1.

In general, the various processes used to make an IC can be divided into three categories: film deposition, patterning, and semiconductor doping. Films of both conductors and insulators are used to connect and isolate transistors and their components. Selective doping of various regions of silicon allow the conductivity of the silicon to be changed with the application of voltage. By creating structures of these various components, millions of transistors can be built and wired together to form the complex circuitry of a modern microelectronic device.

Fundamental to all the above processes is lithography, in which the patterns on the masks (i.e. specific circuit designs) are exposed onto the wafer. In a complex IC (e.g., modern CMOS) fabrication, a wafer will go through the lithographic cycle up to 50 times,
1. Introduction

while each patterned layer being exposed on the wafer must be aligned to the previously formed patterns on the wafer. The displacement of an exposed image field relative to the one exposed in the previous layer is defined as overlay, which must be controlled to within the tolerance incorporated in the IC design to ensure the functionality of the IC.

Since the combination of increasing pattern density and innovative techniques such as double patterning and 193nm immersion lithography creates a novel set of pattern-based yield challenges at the 45nm technology node and below, overlay control has become more and more critical. The diagnosis of causes of overlay and the control of overlay through removal of correctable errors via compensation are important problems in semiconductor manufacturing.

There have been a number of studies on the factors causing the overlay, the mathematical models, and the overlay control methods. However, only a few models are applied in practice. For one thing, the variables considered in the existing overlay models are different from the real data measured through the calibration system. There is a need to bridge the gap between theoretical models and the data in real settings. For another, due to the measurement costs, the number of overlay targets measured per wafer is limited. In order to obtain an as good as possible predicted overlay accuracy with a limited number of overlay measurements, it is important to develop systematic sampling strategies in which one can decide the total number and the locations of the overlay targets to measure.

The rest of this chapter is organized as follows: first, a general description of the problem is given in Section 1.1; then, the goals of the project are defined in Section 1.2; finally the techniques and methods used in the project are explained briefly in Section 1.3.

1.1 Problem Description

There are a number of causes leading to overlay, including system environment, stepper, mask accuracy, and lens distortion [6]. As the overall device size continues to increase, the factors causing overlay become more complex.

In general, the overlay can be divided into two classes: intra-field and inter-field, see the Figure 1.2 and Figure 1.3.

Figure 1.2: Intra-field errors: symmetric rotation, asymmetric rotation, symmetric magnification, asymmetric magnification

The intra-field errors are the displacements caused by fitting problems between mask and earial image¹, while the inter-field errors are the displacements caused by fitting prob-

¹Earial image is the image of mask projected by lens on the wafer.
1.1. Problem Description

Figure 1.3: Inter-field errors: translation, rotation and scaling

lems between earial image and wafer. Nevertheless, only the total overlay are actually measured.

As part of usual production, the overlay of the exposed wafers is checked on an off-line metrology system and process corrections are calculated from the metrology data. Figure 1.4 shows different data flows in the fab environment during the application of process corrections [15].

Figure 1.4: Application of process correction in the production loop

Usually, a sample of a few wafers per lot is measured. Then the measurement data is collected and sent to the APC system for modeling and calculations of new corrections. Then Manufacturing Execution System(MES) updates the process job with the new corrections and sends it to the lithography system where the corrections are applied during exposure of the subsequent lots. Systematic overlay introduced by various lithographic tools are identified and eventually compensated.

The most accurate process correction mechanism is the application of corrections per exposure (CPE). However, this requires a lot of measurement time on an overlay metrology tool. To reduce the measurement effort, higher order process corrections (HOPC) have been introduced, in which only a subset of all fields and/or a subset of overlay targets
per exposure field are measured and polynomials are fitted to these measurements. Now the question is, how to achieve an acceptable overlay with a minimum measurement effort on the overlay metrology tool. With the current GridMapper tool (a PC-based tool), one has to manually specify the overlay targets to be measured, parameters for modeling, and the polynomial degree for each parameter to be modeled. Then the expected overlay is computed by the PC tool. This tedious and time-consuming process has to be repeated till an optimal setting is found. Therefore, it is desirable to have an algorithm that can do this automatically.

1.2 Goals

The goals of this project are defined as follows:

- For a given HOPC model and a given number of overlay targets,
  - determine the optimal distribution of the overlay targets (optimal distribution is the distribution which yields the lowest expected overlay);
  - estimate the confidence interval of the expected overlay;

- For a given HOPC model and a given overlay specification, determine the minimum number of overlay targets that must be measured and their distribution;

- For a given number of overlay targets, determine the HOPC model and the optimal distribution of the overlay targets;

- For a given overlay specification, determine the HOPC model that requires the minimum number of overlay targets to be measured and the optimal distribution;

- Construct alternative models other than HOPCs, compare the overlay improvement after different model-based process corrections.

To achieve the above goals, first we need to develop specific sampling strategies which can be used to effectively measure and compensate for overlay. Next, appropriate criteria should be applied to select the optimal HOPC model. Finally, it is worthwhile to investigate some alternative models other than the current used ones.

1.3 Methodology

This section provides a brief introduction to the techniques and methods used during the course of the project.
1.3.1 Experimental Design

The study of experimental design, which enjoys a long history of theoretical development as well as applications, is extremely crucial in modern industry. The purpose of an experiment in industrial engineering is either improve process yields, reduce the overall costs, or improve the quality of products. In our case, we need well-designed experiments to obtain appropriate data that can be used to optimize the process correction settings.

Traditionally, an experiment is implemented in a lab, a factory, or an agricultural field. This is called a physical experiment, which always suffers from random errors so that one obtains different outputs when the experiment is repeated under identical experimental settings. Existence of random errors creates complexity in data analysis and modeling. Hence experimenters often use some powerful statistical experimental designs.

The statistical approach to the design of experiments is usually based on a statistical model. A good design is one which is optimal with respect to the statistical model under our consideration. There are many statistical models for physical experiments, among which the optimal design based on regression models is widely used in practice. Good designs can provide unbiased estimators of the parameters involved in the models, or even the smallest variance-covariance matrix in a certain sense. The details of these topics will be addressed in Chapter 3.

1.3.2 Regression Analysis

Since only a part of the overlay targets is measured, the overlay of the unmeasured ones must be predicted. In mathematics, this can be achieved by so-called data-fitting based on regression analysis.

Regression acts to minimize the sum of the squared deviations of the experimental values from values calculated using some theoretical model. Basically, there exist two classes of regression: parametric and nonparametric.

The major difference lies in that for parametric regression, there is a priori model form, whereas nonparametric regression does not take a predetermined form but is constructed according to information derived from the data. Thus nonparametric regression requires larger sample sizes than regression based on parametric models because the data must supply the model structure as well as the model estimates.

Typical examples of parametric regression include linear regression and polynomial regression, which is the focus of our study. Examples of nonparametric regression are smoothing splines and neural networks.

Regression analysis provides useful information with respect to the experimental data, for example, the values of the parameters involved in the model that best describe the data. From this information it is possible to judge how well the data fit the model.
1. Introduction
Chapter 2

Higher Order Process Corrections

This chapter deals with the modeling of process corrections. The main purpose of overlay modeling is to determine machine-correctable terms from the measured overlay to minimize overlay between two layers. Well-known machine-correctable terms include rotation, magnification and translation.

Before the detailed explanation of the models, we will first introduce some important notations and describe the overlay data, which serves as input of any used models.

2.1 Notations and Data Description

2.1.1 Notations

The following notations will be used frequently in the rest of this report:

\[
\begin{align*}
\begin{aligned}
dx, dy & : \text{overlay in } X\text{-axis and } Y\text{-axis}, \\
Tx, Ty & : \text{translation in } X\text{-axis and } Y\text{-axis}, \\
Rs, Ra & : \text{symmetric and asymmetric rotation}, \\
Ms, Ma & : \text{symmetric and asymmetric magnification}, \\
deg(\cdot) & : \text{degree of the polynomial(s) for certain parameter(s)}. \\
\end{aligned}
\end{align*}
\]

Of course, overlay can be caused not only by the above factors. As the wafer size continues to increase, the factors causing overlay become more complex. It is beyond the scope of this investigation to develop in detail the relationships between the error sources and their model components. For further information, we refer to [9].

2.1.2 Data Description

In the previous chapter, we have already mentioned, overlay are measured by off-line tools (e.g. KLA machine) at several overlay targets in exposure fields across the wafer. The
position of overlay target $j$ within a field, i.e., relative to the center of the field, is denoted by $(xf_j, yf_j)$, while the position of the center of the field $i$, i.e., relative to the origin of the wafer coordinate system, is denoted by $(xc_i, yc_i)$. See the illustrative figure below:

![Figure 2.1: Coordinate system of overlay data](image)

### 2.2 Model Derivation

In this section, we will describe the data-fitting models for process corrections in detail.

#### 2.2.1 Basic Assumptions

To construct a mathematical model for the process corrections involved in overlay control, we need some basic assumptions:

- the environmental factors are stable across several lots for the same product.
- there is a correlation between overlay of neighboring fields;

The first assumption ensures the applicability of process corrections in any form, since we do not consider the wafer-to-wafer variations in this project. The second assumption
means there exists systematic components of overlay for the fields on the same wafer, which is crucial for the validity of HOPC models.

### 2.2.2 CPE Model

So far, the most accurate process correction mechanism is Correction Per Exposure (CPE), which applies intra-field corrections per exposure. The intra-field model to describe the correction for target $j$ in a field is:

$$dx_j = Tx + (Ms + Ma)xf_j - (Rs + Ra)yf_j,$$

$$dy_j = Ty + (Ms - Ma)yf_j + (Rs - Ra)xf_j,$$

and the results can be formulated as the following table:

<table>
<thead>
<tr>
<th>Exposure Positions</th>
<th>Intra-field Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>X     Y</td>
<td>Tx Ty Rs Ra Ms Ma</td>
</tr>
<tr>
<td>$x_1$ $y_1$</td>
<td>$Tx_1$ ...</td>
</tr>
<tr>
<td>$x_2$ $y_2$</td>
<td>$Tx_2$ ...</td>
</tr>
<tr>
<td>...    ...</td>
<td>...</td>
</tr>
<tr>
<td>...    ...</td>
<td>...</td>
</tr>
<tr>
<td>$x_N$ $y_N$</td>
<td>$Tx_N$ ...</td>
</tr>
</tbody>
</table>

Table 2.1: Offsets per exposure

In order to use this model, overlay data for each field must be available to calculate the full-wafer CPE. And if the full CPE set of six parameters is to be calculated, a minimum of three measurement targets per field is necessary. Consequently, a large amount of measurement time is necessary on the metrology system, which is unacceptable in a full production environment.

### 2.2.3 Standard Model

Another simple but quite useful model in practice is the standard linear model, which reads as follows:

$$dx_{i,j} = Tx + Mx \cdot xc_i - (R + NO)yc_i + (Ms + Ma)xf_j - (Rs + Ra)yf_j,$$

$$dy_{i,j} = Ty + My \cdot yc_i + R \cdot xc_i + (Ms - Ma)yf_j + (Rs - Ra)xf_j.$$

(2.2a)-(2.2b) is actually a combined inter-intra field model (see Section 1.1 for description of inter-and intra-field effects). The first three terms in each equation are the inter-field part. It produces a shift of the field which depends on the position of the field on the wafer.
The last two terms in each equation are the intra-field part. It produces a correction that depends on the position in the field but it is independent of the field position on the wafer (i.e., the same correction for each field).

The above formulation of the standard linear model can be rewritten in matrix form:

\[
\begin{bmatrix}
\Delta x_{i,j} \\
\Delta y_{i,j}
\end{bmatrix} =
\begin{bmatrix}
1 & 0 & x_i & 0 & -y_i & -y_i & x_j & x_j & -y_j & -y_j \\
0 & 1 & 0 & y_i & x_i & 0 & y_j & -y_j & x_j & -x_j
\end{bmatrix}
\begin{bmatrix}
T_x \\
T_y \\
M_x \\
M_y \\
R \\
NO \\
M_s \\
M_a \\
R_s \\
R_a
\end{bmatrix}.
\] (2.3)

### 2.2.4 HOPC Models

Process corrections based on the linear model do successfully improve the overlay. However, due to the increased complexity of manufacturing processes, overlay can have both linear and non-linear components. Thus it is no longer sufficient to only correct the linear part with the standard model. New models that can correct for non-linear effects are needed. A natural idea is to replace the linear inter-field part and constant intra-field parameters in (2.3) by higher order polynomials (max. 5th order\(^1\)) of the field position on the wafer. Thus the generic form of HOPC models can be written as:

\[
dx_{i,j} = T_x(x_i, y_i) + (M_s(x_i, y_i) + M_a(x_i, y_i))x_j - (R_s(x_i, y_i) + R_a(x_i, y_i))y_j,
\] (2.4a)

\[
dy_{i,j} = T_y(x_i, y_i) + (M_s(x_i, y_i) - M_a(x_i, y_i))y_j + (R_s(x_i, y_i) - R_a(x_i, y_i))x_j,
\] (2.4b)

where each of the polynomial \(P(x_i, y_i)\) has the form:

\[
P(x_i, y_i) = \sum_{n=0}^{N} \sum_{m=0}^{n} C_{m,n-m} x_i^m y_i^{n-m}, \quad N = 0, \ldots, 5. \tag{2.5}
\]

\(^1\)Since even higher orders would imply we have so many coefficients that the determination of these would require the measurement of almost the same number of overlay targets as corrections per exposure. Then, one better choose the corrections per exposure.
The number of coefficients \( C \) is equal to

\[
N_{\text{coef}} = \frac{1}{2}(N + 1)(N + 2).
\] (2.6)

Obviously, a 5th order polynomial has 21 coefficients.

Substituting the polynomials in vector notation, we can obtain the generic matrix form:

\[
\begin{bmatrix}
  dx_{i,j} \\
  dy_{i,j}
\end{bmatrix} = \begin{bmatrix}
  Ptx_i^T & 0_T & x_j Pms_i^T & x_j Pma_i^T & -y_j Prs_i^T & -y_j Pra_i^T \\
  0_T & Pty_i^T & y_j Pms_i^T & -y_j Pma_i^T & x_j Prs_i^T & -x_j Pra_i^T
\end{bmatrix} \begin{bmatrix}
  Ctx \\
  Cty \\
  Cms \\
  Cma \\
  Crs \\
  Cra
\end{bmatrix},
\] (2.7)

where \( Ptx_i^T, Pty_i^T, Pms_i^T, Pma_i^T, Prs_i^T, Pra_i^T \) have the same formulation as the row vector:

\[
P_i^T = [1, xc, yc, xc^2, xc * yc, \ldots, xc * yc^4, yc^5].
\] (2.8)

and \( Ctx, Cty, Cms, Cma, Crs, Cra \) are column vectors with the polynomial coefficients as:

\[
C = [C_{00}, C_{10}, \ldots, C_{14}, C_{05}]^T.
\] (2.9)

For convenience, we will denote the matrix on the right hand side of the equation by \( \mathbf{M} \). Clearly, the maximum number of columns of the above matrix equals 126: 6 polynomials times 21 coefficients per polynomial. Therefore, in order to compute all these 126 coefficients, we need at least the same number of rows. Then all the polynomial coefficients can be found by pre-multiplying the pseudo-inverse of \( \mathbf{M} \) on both sides of (2.7).

Of course, it is not always beneficial to use polynomials up to 5th order, nor use polynomials of the same degree for all the 6 parameters. From a mathematical point of view, it is better to use lower order polynomials to avoid over-fitting and numerical instability. Furthermore, even if we use 5th order polynomials for some parameters, not all the terms have significant influence on the result. In practice, removing those insignificant terms can improve the efficiency of the model. These issues will be left to Chapter 4.

Next, we will turn to the discussion about the selection of input data, or more exactly, the selection of overlay targets.
2. Higher Order Process Corrections
Chapter 3

Optimal Distribution of Overlay Targets

In the previous chapter, the HOPC models have been introduced. When the concrete formulation of the model is specified, the next step is to determine specific sampling strategies, which can be applied in practice to effectively measure and compensate for overlay with a limited number of measurements.

Clearly, if measurement time on the metrology tool were no issue, then all overlay targets could be measured in order to maximize the accuracy. However, in practice the measurement time is limited, and accuracy and efficiency have to be balanced. Note that for the standard 10 parameter model typically 9-15 fields across the wafer are selected, and per field 4-6 overlay targets are measured. Furthermore, even with a specified number of overlay targets, we still need to decide the distribution of them. For a certain number of overlay targets, say, \( N \), we may have millions of possible distributions and out of all these possibilities, ideally, we want to find an optimal distribution such that the computed regression system can make the best predictions of the overlay performance at the unmeasured targets.

In the following, we will introduce two sampling patterns: the first one, based on the analysis of general regression models, only uses the layout information of the overlay targets; the second one, which groups fields according to a similarity rule, is motivated by clustering algorithms from pattern recognition.

3.1 Geometry-based Pattern

3.1.1 Basic Principles

To determine the optimal distribution of overlay targets, first let’s consider a more general problem, namely, how the choice of inputs \( \{x_1, \ldots, x_n\} \), or equivalently, \( \mathbf{X} \) affects the accuracy of a regression model, which reads as:

\[
y = \mathbf{Xw} + \mathbf{e}, \quad \mathbf{e} \sim N(0, \sigma^2 \mathbf{I}),
\]

\[\text{Equation (3.1)}\]
where \( \mathbf{w} \) is a vector of model parameters, \( \mathbf{e} \) is the normally distributed random noise and the outputs \( \mathbf{y} = \{y_1, \ldots, y_n\} \).

Assume that the training outputs are actually generated by a model in this class with some fixed but unknown parameter \( \mathbf{w}^* \) such that
\[
\mathbf{y} = \mathbf{X}\mathbf{w}^* + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I}),
\]
(3.2)

We want to know, for a given \( \mathbf{X} \), how accurately we are able to recover the “true” parameter \( \mathbf{w}^* \). For a fixed \( \mathbf{X} \), the maximum likelihood (ML) estimator \( \hat{\mathbf{w}} \), is given by
\[
\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \quad (3.3)
\]

In the absence of noise \( \mathbf{e} \), the ML estimator would recover \( \mathbf{w}^* \) exactly:
\[
\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\mathbf{w}^*) \\
= (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})\mathbf{w}^* \\
= \mathbf{w}^*, \quad (3.4)
\]
while in the presence of noise we can still use the relation \( \mathbf{y} = \mathbf{X}\mathbf{w}^* + \mathbf{e} \) to simplify the parameter estimates:
\[
\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\mathbf{w}^* + \mathbf{e}) \\
= (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})\mathbf{w}^* + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e} \quad (3.5)
\]

Therefore, the ML estimate is the correct parameter vector plus an estimate based purely on noise.

Now we have shown that the ML estimator is a linear function of normally distributed noise \( \mathbf{e} \), hence it is also normally distributed. To fully characterize its distribution, given \( \mathbf{X} \), we need to evaluate its mean and covariance.

Based on our assumption in (3.1), the noise is zero mean. It is shown in [11] that our parameter estimator is unbiased:
\[
E\{\hat{\mathbf{w}}|\mathbf{X}\} = \mathbf{w}^* + E\{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e}|\mathbf{X}\} \\
= \mathbf{w}^* + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TE\{\mathbf{e}|\mathbf{X}\} \\
= \mathbf{w}^* + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{0} \\
= \mathbf{w}^*. \quad (3.6)
\]

---

1Equation (3.3) is usually called “normal equation” in Least Squares Method.
3.1. Geometry-based Pattern

Similarly, the covariance can be computed as follows:

\[
E\{(\hat{\mathbf{w}} - \mathbf{w}^\star)(\hat{\mathbf{w}} - \mathbf{w}^\star)^T|\mathbf{X}\} = E\{[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e}][(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e}]^T|\mathbf{X}\} = E\{[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e}]\mathbf{e}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}|\mathbf{X}\} = \sigma^2[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}]
\]

Thus, when the assumptions in the polynomial regression model are correct, the ML estimator follows a simple Gaussian distribution:

\[
\hat{\mathbf{w}} \sim N(\mathbf{w}^\star, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}).
\]  

3.1.2 Ill-conditioning and Centering Effects

According to the previous analysis, the estimation procedure depends on calculation of the covariance matrix \((\mathbf{X}^T\mathbf{X})^{-1}\) and computational difficulty arises when \(\mathbf{X}^T\mathbf{X}\) is nearly singular or ill-conditioned. Ill-conditioning arises if the sample correlation between any pair of independent variables or between any two functionally independent functions of independent variables is numerically near unity.

Ralph and Sushil [2] studied the effects of centering on correlation and ill-conditioning. Given the polynomial regression model in one independent variable:

\[
E(\mathbf{Y}) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p,
\]  

where the rows of the data matrix \(\mathbf{X}\) is \((1, x_i, x_i^2, \ldots, x_i^p), \ i = 1, \ldots, n, n \geq p + 1\). The sample correlation between \(x^\alpha\) and \(x^\beta\) is:

\[
r_{\alpha\beta} = C_{\alpha\beta}/S_{\alpha}^{1/2}S_{\beta}^{1/2},
\]  

where

\[
C_{\alpha\beta} = \sum_{i=1}^{n} (x_i^\alpha - \bar{x}^\alpha)(x_i^\beta - \bar{x}^\beta),
\]

\[
S_{\delta} = \sum_{i=1}^{n} (x_i^\delta - \bar{x}^\delta)^2, \quad \bar{x}^\delta = \frac{1}{n} \sum_{i=1}^{n} x_i^\delta, \quad \delta = \alpha, \beta.
\]  

It is shown in the appendix of [2] that

\[
C_{\alpha\beta} = \sum_{j=0}^{\alpha+\beta-2} a_j c_j, \quad S_{\alpha}^{1/2}S_{\beta}^{1/2} = \sum_{j=0}^{2(\alpha+\beta)-4} b_j c_j,
\]
where the \( a_j \) and \( b_j \) are functions of the central sample moments,

\[
m_s = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^s, \quad s = 0, 1, 2, \ldots,
\]

(3.14) and \( c \) is a standardized location of \( x \),

\[
c = \frac{\bar{x}}{(m_2)^{1/2}}.
\]

(3.15)

Now the correlation \( r_{\alpha\beta} \) can be expressed as a function of \( c \) and thus written as \( r_{\alpha\beta}(c) \). We are interested in the behavior of \( r_{\alpha\beta}(c) \) when \( c \) is numerically large. It can be shown that

\[
a_{\alpha+\beta-2}^2 = b_{2(\alpha+\beta)-4},
\]

consequently, we have

\[
r_{\alpha\beta}(c) = \frac{c_{\alpha+\beta-2}}{c_{\alpha+\beta-2}^2 + f(c)}.
\]

(3.16) where \( \lim_{|c| \to \infty} f(c) = 0 \). Thus, when \( |c| \) is large, we have the following approximate result:

\[
r_{\alpha\beta}(c) = \pm 1,
\]

(3.17) where the sign is determined by the sign of \( c \) or of \( \bar{x} \) if \( \alpha + \beta - 2 \) is odd and positive.

Therefore, we can draw the conclusion that the sample correlations between \( x^x \) and \( x^\beta \), \( x < \beta \), \( \alpha, \beta = 1, 2, \ldots, p \), will be numerically near unity when the origin for \( x \) is such that \( c \) is large or \( \bar{x} \) is large relative to \( m_2^{1/2} \). In other words, when the location of \( x \) is not adjusted by centering, the matrix inversion computations will be difficult. Ill-conditioning in \( X^T X \) coincides with ill-conditioning in the correlation matrix \( R = [r_{\alpha\beta}] \), which has rank unity other than \( p \) as \( |c| \to \infty \).

To “center” the predictor variables means that one should translate the origins of the independent variables to or near their means. Suppose that \( x \) has been centered to have zero mean in the polynomial model (3.9) under consideration. Then \( c = 0 \) and it can be proved that

\[
r_{\alpha\beta}(0) = \frac{(m_{\alpha+\beta} - m_{\alpha}m_{\alpha})}{((m_{2\alpha} - m_{\alpha}^2)(m_{2\beta} - m_{\beta}^2))^{1/2}}.
\]

(3.18) If values of \( x \) are symmetrically placed about \( \bar{x} \) (about zero when centered), then \( r_{\alpha\beta}(0) = 0 \) when \( \alpha + \beta \) is odd, for then \( m_{\alpha+\beta} = 0 \) and one of \( m_{\alpha} \) and \( m_{\beta} \) will be zero also. When \( \alpha + \beta \) is even and \( \alpha \) and \( \beta \) are both even, (3.18) does not reduce. When \( \alpha + \beta \) is even and both \( \alpha \) and \( \beta \) are odd, (3.18) reduces to

\[
r_{\alpha\beta}(0) = m_{\alpha+\beta}/(m_{2\alpha}m_{2\beta})^{1/2}.
\]

(3.19) Thus when \( x \) in the polynomial regression is open to choice, symmetric sample should be considered.

Further, according to [10], if the values of \( x \) are limited to a narrow range, there can be significant ill-conditioning or multi-collinearity. Thus the spatial coverage of the sample should also be taken into account.

To sum up, when selecting input for a given polynomial regression model, we should follow three guidelines:
3.1. Geometry-based Pattern

- Firstly, the origin of the independent variable be chosen or adjusted so that the mean of the values used is zero;
- Secondly, values of the independent variable to be selected should be located symmetrically about their mean when possible\(^2\);
- Thirdly, values of the independent variable should have a good spatial coverage of the domain.

3.1.3 Algorithms and Experimental Results

According to the previous analysis, a good sampling pattern of overlay targets for HOPC models should have three properties: zero-mean, symmetry and spatial-coverage. For zero-mean and symmetry, the overlay targets should be chosen such that they are balanced or canceled by some opposite targets. For coverage, the targets to be measured should achieve good spatial distribution in both \( x \) and \( y \) directions so that the “area of interest” is bounded by measurement sites. For these reasons, we develop a geometry-based sampling strategy, which is depicted in Algorithm 1:

\textbf{Algorithm 1: Pseudo Code for Geometry-Based Sampling}

- grouping fields in a zero-mean, symmetric and spatial coverage way;

\textbf{Input}: layout of a wafer, measurement specification, i.e., the number of measurement \( N \)
\textbf{Output}: a list of fields \( L_f \) that are to be measured

1. \textsc{GroupPointsInGeometricalWay};
2. Design a \((2k + 1)^2\)-square cover of the wafer, where \( k \) satisfies \((2k - 1)^2 \leq N \leq (2k + 1)^2\);
3. \textbf{if} The number of groups \((2k + 1)^2\) is strictly larger than the number \( N \) \textbf{then}
4. \hspace{1em} Delete \((2k + 1)^2 - N\) groups in a symmetric way;
5. \textbf{end}
6. Pick up the central field from each group to form the list \( L_f \);
7. \textbf{return} \( L_f \)

As illustration, we plot two sampling patterns for a 195-field wafer: one with 13 selected fields and the other 17. According to the geometry-based sampling strategy, first a 25-square cover of the wafer was calculated. The fields within the same square form a group. Then, for the first sampling pattern, 12 groups were “deleted”, and for the second, 8 groups were “deleted”. The results are shown in Figure 3.1(a)-(b), where the red-marked fields are denoted as selected ones.

\(^2\)For our 2D problem, symmetry should be ensured for both \( X \)-axis and \( Y \)-axis.
To test the validity, we applied the sampling strategy to a 25-wafer lot where each wafer contains 92 fields. The overlay data from the selected fields was used to fit the 2nd order HOPC model and calculate corrections. The expected overlay \((M + 3\sigma)\) after corrections is depicted in Figure 3.2. Note that for CPE, we need full-wafer measurement data, i.e., all 92 fields must be measured.

To verify the stability, we did “sub-lot” test where the 25-wafer lot was divided into 5 sub-lots, each with 5 wafers and only the measurements from 1 sub-lot were used to calculate the corrections. Those corrections were then applied to all the 25 wafers from this lot\(^3\). The overlay after corrections can be seen in Figure 3.3(a)-(b). During this test, first we took the measurements from the 1st sub-lot, then the 2nd, etc.

\(^3\)During the real production loop, only a few wafers are measured and the calculated process corrections are applied to all the wafers across several lots for the same product.
3.1. Geometry-based Pattern

According to Figure 3.2 and Figure 3.3, the geometry-based sampling is quite effective and stable. In practice, if the two assumptions made in Section 2.2 hold, the overlay improvement by 2nd order HOPC model equipped with this sampling strategy can compete with those by the CPE and the measurement effort for the former is much less.

The three guidelines: zero-mean, symmetry and spatial-coverage are very important to ensure the effectiveness and stability of the geometry-based sampling strategy. During our investigation, we did empirical tests to compare the performance between geometry-based sampling patterns and those obtained with manually chosen distributions that don’t meet the symmetry or spatial-coverage requirements. The sampling patterns we used are as follows: Figure 3.4(a), a geometry-based sampling pattern; Figure 3.4(b), a non-symmetric but spatial-coverage pattern; Figure 3.4(c), a sampling pattern with neither symmetry nor spatial-coverage. Each pattern contains 195 fields but only the measurements from 15 fields were used to calculate the corrections.

To compare the overlay improvements after corrections with the above sampling patterns, we applied them to a 16-wafer lot (where the layout of the wafers is the same as shown in Figure 3.4) and the results can be seen from Figure 3.5. Here, the overlay data was again used to fit the 2nd order HOPC model.
According to Figure 3.5, we can find that lack of symmetry or poor coverage in a sample plan inflates the prediction uncertainty and thus deteriorate the overlay after corrections. Further analysis about the advantages of symmetric and good-coverage sampling plans can be found in [4].

### 3.2 Data-driven Pattern

Compared to the geometry-based sampling pattern, which is based on minimizing the condition number of the covariance matrix as a measure of prediction uncertainty, the idea of data-driven sampling seems more natural and straightforward.

#### 3.2.1 General Description

The goal of the data-driven sampling is to select the most representative overlay targets across the wafer. To ensure good representability, we try to group fields according to the similarity between different fields, i.e., fields belonging to the same group should be similar while fields from different groups should not. The question is how to define a measure of similarity. Since one of the most important features for these fields is overlay, we decide to use the difference between the average overlay per field as the measure of similarity. In other words, if the overlay differences between two fields are smaller than a predefined tolerance $\tau$ in both X and Y direction, they are defined to be similar.

The similarity-based sampling algorithm works as follows: First, a list of groups is generated, following the similarity rule. Then, according to the relation between the specified number $N$ of targets allowed to be measured and the number of generated groups, delete or merge some unwanted groups. There are three cases:
3.2. Data-driven Pattern

• the number of groups is equal to \( N \): select the field whose overlay\(^4\) is closest to the mean overlay in that group;

• the number of groups is larger than \( N \): delete the groups with less than two fields and if the number of groups still exceeds \( N \), merge some similar groups;

• the number of groups is smaller than \( N \): decrease the tolerance \( \tau \) by half and generate a new list of groups till one of the first two cases happen;

The deletion rule is based on the fact that, high-variation fields should not be included in the sample, if the given budget of measurement is limited.

3.2.2 Experimental Results

During the experiments, we used \( \tau = 1 \text{ (nm)} \) as the tolerance for both X and Y direction, i.e. when the differences of average overlays between two fields are smaller than \( 1 \text{ (nm)} \) in both X and Y direction, these two fields are similar and as a result, will be grouped together.

As an illustration, Figure 3.6(a)-(b) show two sampling patterns on a 195-field wafer, which are generated by the data-driven algorithm. Fields with different color belong to different groups and the selected fields are marked with red star.

![Figure 3.6: Data-driven sampling patterns](image)

(a) Selection results with 13 fields
(b) Selection results with 17 fields

To verify the effectiveness and stability of the data-driven sampling, we did similar tests with the same 25-wafer lot as those for geometry-based sampling. First, during the “individual” test, the overlay data from the selected 15 fields (out of 92 fields) was used to fit the 2nd order HOPC model and calculate the corrections. Figure 3.7 depicts the

\(^4\)We use the average value of the overlay measured on the targets in a field as the overlay of that field, similar for the overlay of a group.
expected overlay after corrections. Then, during the “sub-lot” test, the measurements from the 1st, 2nd, . . . , 5th sub-lot were used in turns to calculate corrections. The results are shown in Figure 3.8(a)-(b).

Figure 3.7: Overlay after “individual” corrections with data-driven sampling

![Figure 3.7](image)

(a) Overlay in $X$-axis after corrections  
(b) Overlay in $Y$-axis after corrections

Figure 3.8: Overlay after “sub-lot” corrections with data-driven sampling

![Figure 3.8](image)

(a) Overlay in $X$-axis after corrections  
(b) Overlay in $Y$-axis after corrections

According to these test results, the data-driven sampling is not so effective and stable as the geometry-based sampling. Besides, since the full-wafer overlay data must be known beforehand in order to generate the groups of fields, it is more expensive (or less efficient) than the geometry-based strategy.

3.3 Remarks and Further Discussion

In this chapter, we developed two sampling strategies/algorithms to automatically select overlay targets to be measured. The geometry-based sampling can be seen as a dedicated
experimental design for the HOPC models. The empirical tests show its effectiveness and stability. Besides, since it only depends on the specified number of fields to be measured and the layout of the product, the geometry-based sampling doesn’t require any a priori overlay information and thus it is very efficient. The data-driven sampling can be regarded as a cluster sampling strategy, where the entire population is divided into groups, or clusters, and a random sample of these clusters are selected. The advantage of cluster sampling is obvious: it can be implemented economically while retaining the characteristics of a sample. If combined with some non-parametric models (which will be introduced briefly in chapter 5), the data-driven sampling may excel in the effectiveness. However, with the current HOPC models, the geometry-based sampling outperforms the data-driven sampling, especially for its efficiency. Hence we recommend the use of geometry-based sampling strategy during overlay control.
3. Optimal Distribution of Overlay Targets
Chapter 4

Selection of Polynomial Degree and Relevant Terms

The focus of this chapter is the determination of a suitable correction model. As we know, a good model should balance goodness of fit and complexity. Take polynomial regression model for example: although more complex models will be better able to adapt their shape to fit the data (a 6th order polynomial can exactly fit 7 points), it can also be possible that the additional parameters may not represent anything useful (perhaps those 7 points are really just randomly distributed about a line).

The rest of this chapter is organized as follows: first, a short description of the motivation is provided; then, three different selection criteria/methods are introduced, Adjusted $R^2$, Partial Sum of Squares and ANOVA-like Approach, respectively.

4.1 Motivation

Polynomial regression entails an inherent trade-off between accuracy and efficiency. As the degree of the polynomial increases, the fit grows in accuracy, but the time needed to determine it increases as well. For the overlay control problem, the number of measurements required for the fitting of a HOPC model increases exponentially when higher order terms are introduced. In addition, from a numerical point of view, the covariance matrix involved in the HOPC models will become ill-conditioned as the order increases and the calculations of the pseudo-inverse will be inaccurate. As a result, considerable errors may be introduced into the parameter estimates. Hence, it is really important to ensure the higher order terms in the correction model can bring significant overlay improvement.

As an illustration, let’s first take a look at Figure 4.1, which depicts the expected overlay on a wafer after different corrections with full-wafer measurement. The $X$–axis represents different correction models. The two leftmost values in the figure denote the initial overlay in $X$ and $Y$ direction before applying any correction. The rest are the overlays after corrections with different HOPC models (from left to right, we selected $\text{deg}(T_x,T_y) = 1, \text{deg}(M_s,M_a,R_s,R_a) = 0; \text{deg}(T_x,T_y) = 1, \text{deg}(M_s,M_a,R_s,R_a) = $
Selection of Polynomial Degree and Relevant Terms

1; \( \ldots \ deg(Tx,Ty) = 3, deg(Ms, Ma, Rs, Ra) = 3 \) in turn\(^1\) as well as CPE model.

From Figure 4.1, we have two observations. First, it seems that the overlay doesn’t improve much when the polynomial degree of \( Tx \) and \( Ty \) increased from 1 to 3. Second, it doesn’t make much difference when we model \( Ms, Ma \) and \( Rs, Ra \) as constants or \( i \)th order polynomials.

Thus the goal of our investigation is to choose the simplest but adequate model. To determine an appropriate complexity of a polynomial regression model is not easy. For example, if the degree is too high, the regression can suffer from the drawbacks such as over-fitting. On the other hand, if the model is too simple, it will not be able to describe the system under consideration accurately.

In the following, we will present three different selection strategies to automatically identify appropriate degrees and relevant terms for HOPC models with the training data, which in our case is the measured overlay.

\(^1\)With the current GridMapper Tool, one cannot select individual polynomial terms, but only the degree of the polynomials for each modeled parameter
4.2 Adjusted $R^2$:

Denote the total sum of squares, regression sum of squares, and the residual sum of squares (sum of squared errors) by

$$\text{SSTO} = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

(4.1)

$$\text{SSR} = \sum_{i=1}^{n} (f_i - \bar{y})^2,$$

(4.2)

$$\text{SSE} = \sum_{i=1}^{n} (y_i - f_i)^2,$$

(4.3)

where $\bar{y}$ and $\bar{f}$ are the means of the observed data and modeled (predicted) values respectively. The coefficient of multiple determination or model coefficient of determination is defined as the positive square root of

$$R^2 = \frac{\text{SSR}}{\text{SSTO}}.$$

(4.4)

and $R^2$ provides a measure of model fit.

However, a large value of $R^2$ does not necessarily imply that the regression model is a good one. In fact, adding a new variable (term) to the model will always increase $R^2$ [1], regardless of whether the additional variable (term) is statistically significant or not. It is possible for models that have large values of $R^2$ to yield poor predictions of new observations or estimates of the mean response.

Therefore, to better evaluate the goodness of fit of a certain regression model, adjusted $R^2$ is introduced, which reads as

$$R^2_{\text{adj}} = 1 - \frac{\text{SSE}/(n-p)}{\text{SSTO}/(n-1)} = 1 - \frac{(n-1)}{(n-p)} (1 - R^2).$$

(4.5)

where $p$ is the number of coefficients to be decided for a specific model.

Unlike $R^2$, adjusted $R^2$ measures the proportion of the variation in the dependent variable accounted for by the explanatory variables. In general, the adjusted $R^2$ will not always increase as variables (terms) are added to the model. In fact, if unnecessary terms are added, the value of $R^2_{\text{adj}}$ will often decrease. For this reason, adjusted $R^2$ is usually considered to be a more accurate goodness-of-fit measure than $R^2$.

During the early investigation, we applied the adjusted $R^2$ to decide the appropriate complexity of an HOPC model by computing optimal polynomial degree for each parameter. The pseudo code is included in Appendix B.

To test the effectiveness and stability of the adjusted $R^2$, we applied it to the 25-wafer lot where each wafer contains 92 fields as before. With the measurements from certain fields (geometry-based sampling was used to select these fields) on the selected wafers (during the
“sub-lot” test, the whole lot was divided into 5 sub-lots and only the measurements from 1 sub-lot were used, the polynomial degree for each parameter was determined and then the corrections were calculated.

Figure 4.2 shows the average overlay improvements after “individual” corrections with reduced model by adjusted $R^2$, different HOPC models and also CPE model for each wafer (15-fields were measured). We can find that the reduced model only performs better than the 4th order HOPC model and worse than the other three.

During the “sub-lot” test, first, the geometry-based sampling strategy was applied to select certain fields from the 5 wafers under consideration (first the overlay data from 1st sub-lot was used, then 2nd sub-lot, etc) and the overlay data from the selected fields was used by adjusted $R^2$ to derive the reduced model. Then, the measurements were fitted to both the reduced model and CPE model. Finally, the calculated corrections were applied to all the 25 wafers in the lot. Figure 4.3, 4.4 and 4.5 show the average overlay improvement results for this lot when 15 fields, 21 fields and all fields were measured respectively. The polynomial degrees calculated during the test are included in Appendix C.
According to these experimental results, the simple selection criteria adjusted $R^2$ is neither effective nor stable\(^2\) when there is not enough measurements available (see Figure 4.3). However, if sufficient measurement data is provided, we can obtain quite effective and stable correction models with this method (in our case, when 21 fields were measured, the overlay improvement is quite significant and stable). In the following, we will use the adjusted $R^2$ together with Partial Sum of Squares during the forward selection process.

### 4.3 Partial Sum of Squares

The Partial Sum of Squares strategy [3] we will introduce in this section can be seen as a type of stepwise regression, where new terms are added based on their computed statistical significance.

\(^2\)In fact, there are two different kinds of stability here: one is about the selected polynomial degrees and the other is about the overlay. Since the goal is to minimize overlay as much as possible, the stability of the overlay after corrections is what we really care.
4.3.1 Basic Principles

Let \( x_{(1)} \) be a sub-vector of independent variable \( x \). Let \( \text{SSR}(x_{(1)}) \) and \( \text{SSE}(x_{(1)}) \) be the regression sum of squares and the error sum of squares, respectively, when only \( x_{(1)} \) is included in the regression model. For example, take \( x_{(1)} = (x_1, x_2) \); then \( \text{SSR}(x_1, x_2) \) is the regression sum of squares when only \( x_1 \) and \( x_2 \) are included in the regression model. The difference \( \text{SSR}(x_1, x_2) - \text{SSR}(x_1) \) reflects the increase of regression sum of squares associated with adding \( x_2 \) to the regression model when \( x_1 \) is already in the model. Thus we define the \textit{extra sum of squares} \( \text{SSR}(x_2|x_1) \) as follows:

\[
\text{SSR}(x_2|x_1) = \text{SSR}(x_1, x_2) - \text{SSR}(x_1).
\] (4.6)

Note that for polynomial regression models, we have

\[
\text{SSTO} = \text{SSR}(x_1, x_2) + \text{SSE}(x_1, x_2) = \text{SSR}(x_1) + \text{SSE}(x_1).
\] (4.7)

Thus,

\[
\text{SSR}(x_2|x_1) = \text{SSE}(x_1) - \text{SSE}(x_1, x_2),
\] (4.8)

Similarly, we have

\[
\text{SSR}(x_{(2)}|x_{(1)}) = \text{SSR}(x_{(1)}, x_{(2)}) - \text{SSR}(x_{(1)}),
\] (4.9)

and

\[
\text{SSR}(x_{(2)}|x_{(1)}) = \text{SSE}(x_{(1)}) - \text{SSE}(x_{(1)}, x_{(2)}).
\] (4.10)

From the above results, we can see, the extra sum of squares \( \text{SSR}(x_{(2)}|x_{(1)}) \) measures the marginal reduction in the error sum of squares when \( x_{(2)} \) is added to the regression model, given that \( x_{(1)} \) is already in the model. In particular, taking \( x_{(2)} = x_i \) and \( x_{(1)} = x_{-i} = \{x_1, \cdots, x_{i-1}, x_{i+1}, \cdots\} \), then

\[
p_i = \text{SSR}(x_i|x_{-i}), \quad i = 1, \cdots
\] (4.11)

measures the importance of input variable.

4.3.2 Forward Selection

The relevant terms of a correction model can be determined by the so-called \textit{forward selection} process, which starts with an initial model (in our case, we start with 0th order polynomials for each parameter), and be “greedy” in the whole selection process. “Greedy” means at each step, only the terms with large\(^3\) extra sum of squares are included in the model. See the illustrative flowchart in Figure 4.2.

At \( i \)th step, we first test all the \( i \)th order polynomial terms for \( Tx \) and \( Ty \), then test those terms for \( Rs, Ra \) and \( Ms, Ma \). Note that the selection loop will continue to iterate at \( i \)th step even some terms with order \( i \) have already been selected. In other words, if the extra sum of squares for the under-processing term is large enough, it will be included, no matter whether the same order terms already exist in the model or not.

\(^3\)Here we use a predefined tolerance to determine whether the extra sum of squares for a certain term is large enough
4.3.3 Experimental Results:

To compare the effectiveness and stability of the reduced correction model using PSS strategy\(^4\) and the original HOPC models as well as CPE model, again, we applied them to the 25-wafer lot. During the “individual” test, the measurements from 15 fields on each wafer were first used to determine the “reduced” model and then fit both original models and the reduced model. From Figure 4.7, we can find that for an individual wafer, the reduced model (usually with less than 10 terms, see Appendix C) performs almost equally as the 2nd order full model and better than 3rd or 4th order full models. Although the CPE model seems much superior than the other correction models, it is the most expensive one, since it requires the measurement of overlay in each field.

Next, during the “sub-lot” test, the 25-wafer lot was divided into 5 sub-lots and the

\(^4\)Actually, the adjusted \(R^2\) is also used implicitly, see the flowchart Figure 4.3.2.
4. Selection of Polynomial Degree and Relevant Terms

Figure 4.7: Overlay after “individual” full-and reduced (PSS)-model corrections

measurements from each sub-lot was used to determine the model and calculate the corrections in turns. The results can be seen from Figure 4.8, 4.9 and 4.10.

Figure 4.8: Overlay improvements after CPE and PSS (15 fields)

Figure 4.9: Overlay improvements after CPE and PSS (21 fields)

According to these test results, the PSS strategy is quite effective and generally more stable than the adjusted $R^2$. Though the selected terms (see Appendix C) are not so
4.3. Partial Sum of Squares

stable when different sub-lots are used or different sample sizes are specified, the overlay improvements are usually significant. Finally, it provides a feasible way to identify the relevant terms.

4.3.4 Modified Partial Sum of Squares with Zernike Polynomials

The defect with the forward selection process is that there exists correlation between higher and lower order terms. This can possibly explain the instability of the selected relevant terms. Since the monomial polynomials used in the forward selection process are not independent with each other, it is quite possible that a specific 2nd or 3rd order term alone doesn’t improve the overlay much but when combined with another linear term, it can bring significant improvement. Thus if the linear term is not included in the model by chance, the 2nd or 3rd order term will not be included either. This kind of chain effect is unfavorable. For this reason, we investigate Zernike polynomials, which form a complete set of functions or modes that are orthogonal over a circle of unit radius and are convenient for serving as a set of basis functions.

The odd and even Zernike polynomials are given by

\[ U_{m}^{n}(\rho, \phi)^{o} = R_{m}^{n}(\rho) \sin(m\phi), \]
\[ U_{m}^{n}(\rho, \phi)^{e} = R_{m}^{n}(\rho) \cos(m\phi), \]

where the radial function \( R_{m}^{n}(\rho) \) is defined for \( n, m \in \mathbb{N} \) and \( n \geq m \) by

\[
R_{n}^{m}(\rho) = \begin{cases} 
\sum_{l=0}^{(n-m)/2} \frac{(-1)^{(n-l)!}}{l!(n+m-l)!l!(n-m-l)!(n-2l)!} \rho^{n-2l} & \text{for } n - m \text{ even} \\
0 & \text{for } n - m \text{ odd}
\end{cases}
\]

The original Zernikes are represented in polar coordinates, but with simple transformations, we can calculate Zernike polynomials in Cartesian. Grouping them according to the orders, we can obtain the first 21 terms as follows:

- Order 0: 1;
• Order 1: \(x, y\);
• Order 2: \(-1 + 2(x^2 + y^2), x^2 - y^2, 2xy;\)
• Order 3: \(-2x + 3x(x^2 + y^2), -2y + 3y(x^2 + y^2), (x^2 - 3y^2)x, (3x^2 - y^2)y;\)
• Order 4: \(1 - 6(x^2 + y^2) + 6(x^2 + y^2)^2, (4x^2 + 4y^2 - 3)(x^2 - y^2), (4x^2 + 4y^2 - 3)2xy, x^4 + y^4 - 6(x^2)(y^2), 4(x^3)y - 4xy^3;\)
• Order 5: \(3x - 12x(x^2 + y^2) + 10x(x^2 + y^2)^2, 3y - 12y(x^2 + y^2) + 10y(x^2 + y^2)^2, -4x^3 + 12xy^2 + 5x^3(x^2 + y^2) - 15xy^2(x^2 + y^2), -12(x^2)y + 4y^3 + 15(x^2)y(x^2 + y^2) - 5y^3(x^2 + y^2), x^5 - 10(x^3)y^2 + 5xy^4, 5(x^4)y - 10(x^2)y^3 + y^5.\)

Different from non-orthogonal set of monomials, for which the best-fit coefficients of lower terms change as higher order terms are added, the best-fit coefficients for Zernike polynomials remain the same, no matter how many or which polynomials are used in the fitting formula. Thus Zernikes are supposed to be a good alternative to monomials when using PSS strategy.

We applied the new PSS strategy with Zernikes to the 25-wafer lot and the “sub-lot” correction results can be seen from Figure 4.11, 4.12 and 4.13.
4.4. ANOVA-Like Approach

Figure 4.13: Overlay improvements after CPE and PSS with Zernikes (full-wafer)

Compare these results with those in the previous section, we can find that the resulting models by PSS with Zernikes do have some improvements in the stability of overlay improvement. However, in some cases (e.g., when 15 and 21 fields are measured, the overlay in Y-axis) the overlay improvement with PSS is more significant than PSS with Zernikes. For the stability of the selected terms (see Appendix C), unfortunately, the results are not so promising as expected either. Possible explanation can be that the Zernikes are orthogonal only in a continuous sense over the interior of a unit circle, and in general they will not be orthogonal over a discrete set of data points within a unit circle (in our case, only a small sample of fields were measured).

4.4 ANOVA-Like Approach

During our investigation, the most successful model selection method is the ANOVA-Like Approach, which will be introduced in detail in this section.

4.4.1 General Description

The ANOVA-like approach (ALA) starts with a first degree polynomial regression model and finds the appropriate degree by continually monitoring the fit. To determine if the degree needs to be increased, a global prediction error and a local noise estimate are compared. Though a high global prediction error indicates a poor fit, suggesting that the polynomial degree should perhaps be increased, such an error might also be attributed to a large amount of random noise in the observed data, in which case increasing the degree will not help. The comparison between global prediction error and local noise estimate finds the lowest degree polynomial that achieves a satisfactory fit.

Suppose that \( y_i \) is the observed data and can be expressed as \( f(x_i) + \omega_i \), where \( f \) is the function we want to approximate and \( \omega_i \) is zero-mean random noise (\( \omega_i \sim N(0, \sigma^2) \)). The expected squared error, \( C^2 \), is \( E([\hat{f}(x) - y]^2) \), where \( \hat{f} \) is the current estimate of \( f \). Since the noise is assumed to be zero-mean and independent of the function error, the expected
squared error can be decomposed as:

\[
C^2 = E([\hat{f}(x) - y]^2) = E([\hat{f}(x) - f(x)]^2) + E(\omega^2).
\] (4.15)

The first term on the right, which represents the function estimate inaccuracy, is what we want to minimize. The second term, which denoted by \(D^2\), is unavoidable, since even for a perfect fit function, the random noise still exists. Because of this, it is desirable to be able to distinguish between these two sources of errors. This can be achieved by estimating \(D^2\), the variance of the random noise, directly.

In order to estimate \(D^2\), we can partition the input domain into a pre-set number of equal-sized small sub-domains (for our problem, the fields on the wafer can serve as sub-domains automatically). Each data point \((x, y)\) is then classified according to which sub-domain \(x\) falls into. The variance of \(y\) over the points in any given sub-domain is a good estimate of the noise in the data at each point. Thus we maintain the variance of \(y\) over each sub-domain and take their average as the mean squared noise, \(D^2\).

From (4.15), we know that \(C \geq D\). The ratio between these two quantities is a good measure of the extent to which the prediction error, \(C\), can be accounted for by the random noise, \(D\). Therefore, to determine whether it is necessary to increase the polynomial degree, we compare \(C\) and \(D\) using a threshold ratio, \(\theta > 1\). If \(C > D\theta\), there is a significant amount of error that cannot be explained by random noise and thus the degree is increased by one. Pseudo code for ALA is included in Appendix B.

### 4.4.2 Experimental Results

To test the effectiveness and stability of the ALA strategy, we applied it to the 25-wafer lot. The results for the “individual” test are shown in Figure 4.14. We can find that for an individual wafer, the corrections computed by the reduced model with ALA performs almost equally as the 2nd full model and better than 3rd and 4th order full models, similar as the corresponding results of PSS strategy.

![Figure 4.14: Overlay after “individual” full- and reduced(ALA)-model corrections](image-url)
For the “sub-lot” test, first the reduced model was determined with the measurements from the sub-lot under consideration. Then the same measurements were used to fit both the reduced model and CPE model. The overlay improvements can be seen from Figure 4.15, 4.16, and 4.17.

**Figure 4.15:** Overlay improvements after CPE and ALA (15 fields)

**Figure 4.16:** Overlay improvements after CPE and ALA (21 fields)

**Figure 4.17:** Overlay improvements after CPE and ALA (full-wafer)

According to the test results, the ALA strategy is not only effective, but also very stable: there exists much less differences in overlay when different sub-lots were used. Further, it
is very efficient: even less than 20 percent of overlay data from some wafers are available, we can obtain useful correction models (in our case, the overlay improvements when 15 fields, 21 fields, or full-wafer were measured are almost the same).

4.5 Concluding Remarks

In order to determine an appropriate model, we investigated existing criteria (methods): adjusted $R^2$ and partial sum of squares (PSS). We also developed a new one, namely, ANOVA-like approach (ALA). The adjusted $R^2$ measures the proportion of the variation in the dependent variable accounted for by the explanatory variables and theoretically, an appropriate model should have a large value of adjusted $R^2$. PSS, as a stepwise selection method, is a useful tool for the selection of relevant terms of the polynomials. Although it suffers from correlation problem during the forward selection process, PSS can generally produce effective and efficient (consider the number of relevant terms) models. The most successful model selection method according to our investigation is ALA, of which the resulting models are verified to be not only effective, but also stable and efficient by empirical tests. Figure 4.18 shows the comparison of the overlay that may be expected when applying the different methods during the “sub-lot” test where 15 fields were measured.

![Figure 4.18: Overlay after “sub-lot” corrections with different methods](image)

We can find that with the same measurements, the reduced model determined by ALA outperforms the models by other selection methods in both effectiveness and stability. As for the CPE model, though it is still the most accurate one, the measurement cost is 5 times higher than others. As we pointed out at the beginning of this chapter, a good model should balance complexity and accuracy. If the extra 10 percent overlay improvement is not that crucial, we recommend the use of reduced models derived by ALA, for its greatly reduced measurement cost.

In the statistical literature, many other “criteria” have been proposed to automatically select models. Popular examples include the Akaike Information Criteria, the Bayes Information Criterion, Mallow’s $C_p$ and cross-validation. These methods optimize parameters
strictly in terms of prediction error. The ALA differs from them in that it balanced the prediction accuracy and the complexity in accordance with the amount of noise.
4. Selection of Polynomial Degree and Relevant Terms
Chapter 5

Non-Parametric Models

In this chapter, we will give a general discussion about non-parametric models, which serve as an alternative to the current polynomial models.

5.1 Background

In the last decade non-parametric models have been used to tackle regression and classification problems, with several notable successes. It has also been widely recognized that they form a part of a wide variety of non-linear statistical techniques that can be used for these tasks.

One of the attractions of non-parametric models is their flexibility, i.e., their ability to model a wide variety of functions. The flexibility is extremely helpful in a preliminary and exploratory statistical analysis of a data set. If no a priori model information about the regression is available, the nonparametric analysis could help in suggesting simple parametric formulations of the regression relationship.

However, this flexibility comes at a cost, in that a large number of parameters may need to be determined from the data, and consequently that there is a danger of “over-fitting”. Over-fitting can be reduced by using weight regularization, but this leads to the problem of specifying how to set the regularization parameters.

5.2 An Overview of Nonparametric Models

In general, the traditional regression model fits the model

\[ y_i = f(\beta; x'_i) + \epsilon_i, \]  \hspace{1cm} (5.1)

where \( \beta = (\beta_1, \ldots, \beta_k)' \) is a vector of parameters to be estimated, and \( x'_i = (x_1, \ldots, x_m) \) is a vector of predictors for the \( i \)th of \( n \) observations; the errors \( \epsilon_i \) are assumed to follow the distribution \( NID(0, \sigma^2) \). The function \( f(\cdot) \) is specified in advance.
The nonparametric regression model [13] is written in a similar manner, but different from the parametric models, the function $f$ is now left unspecified:

$$y_i = f(x'_i) + \epsilon_i. \quad (5.2)$$

And the objective of nonparametric regression is to estimate the regression function $f(\cdot)$ directly, instead of estimating parameters. Most methods of nonparametric regression implicitly assume that $f(\cdot)$ is smooth and continuous. Moreover, the assumption of $\epsilon_i$ still holds.

Since it is difficult to fit the general nonparametric regression model when there are many predictors, and display the fitted model when there are more than two predictors, more restrictive models have been developed. One example is the additive regression model,

$$y_i = \alpha + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_m(x_{im}) + \epsilon_i \quad (5.3)$$

where the partial-regression function $f_j(\cdot)$ are assumed to be smooth, and are to be estimated from the data. This model is much more restrictive than generalized nonparametric regression models, but less restrictive than either linear or polynomial regression models.

There are two main types of nonparametric techniques: *kernel estimation* and *artificial neural networks*.

- Kernel estimation specifies $y = m(x) + e$, where $m(x)$ is the conditional expectation of $y$ with no parametric form, and the density of the error $e$ is completely unspecified. The $N$ observations $y_i$ and $x_i$ are used to estimate a joint density function for $y$ and $x$. The density at a point $(y_0, x_0)$ is estimated by seeing what proportion of the $N$ observations are “close to” $(y_0, x_0)$. This procedure involves the use of a function called a kernel to assign weights to nearby observations.

- Artificial neural networks model an unknown function by expressing it as a weighted sum of several sigmoids (usually chosen to be logit curves), each of which is a function of all the relevant explanatory variables. This amounts to an extremely flexible functional form for which the estimation requires a nonlinear least-squares iterative search algorithm based on gradients.

### 5.3 Potential Benefits

In Chapter 2 and 3, we considered polynomial regression models in which the form of the mapping from input $x$ (position of overlay targets) to output $y$ (overlay) is governed by a vector $w$ of adaptive parameters (such as $T_x, T_y, \ldots$). During the learning phase, a set of training data (overlay data from measured targets) is used to obtain estimates of the parameter vectors. Then the training data is discarded, and predictions for new inputs (unmeasured overlay targets) are based purely on the learned parameter vector $w$ [14].

However, if some nonparametric model, for example kernel estimation, is taken, those training data points, or a subset of them, will be kept and used also during the prediction.
phase. Considering the gap between the variables considered in the existing overlay models and the real data measured through the metrology tools, it might be beneficial to use nonparametric models, which exploit the training data more sufficiently.
Chapter 6

Conclusion and Recommendation

In this chapter, we will look back on the accomplishments of this project and evaluate which part of the goals have been achieved. Furthermore, we will discuss about possible directions of future work. Before presenting the achievements, let’s review the project goals, which can be simplified as the following three:

- develop a strategy to determine a suitable (if possible, an optimal) sampling pattern;
- develop a strategy to determine a suitable (if possible, an optimal) correction model;
- construct alternative models other than HOPC.

6.1 Achievements

In Chapter 3, we developed two sampling algorithms for the overlay targets on a wafer. The first one is geometry-based. Given the layout of overlay targets, the sampling pattern is determined automatically under three principles: zero-mean, symmetry and spatial coverage. As a dedicated design to the HOPC models, the geometry-based sampling allows the number of overlay measurements for a given model to be greatly reduced without a significant increase of overlay (usually, a 15-field sampling already yields acceptable overlay). As an alternative, we also developed a data-driven sampling algorithm, with which the fields on a wafer are grouped according to some predefined similarity rule and then the most representative field is selected automatically from each group. According to the empirical tests, the data-driven sampling algorithm is generally not so stable as the geometry-based one. Further, considering the efficiency, the geometry-based algorithm also outperforms the data-driven one, since the latter requires for a priori overlay information.

In order to determine an adequate and efficient correction model, we investigated existing model selection methods (adjusted $R^2$ and partial sum of squares) and also developed a new one: ANOVA-like approach (ALA). Overlay data from different fab facilities were used to verify the validity of these methods. Basically, we did two different kinds of tests, namely, local-model tests and global-model tests. For the local-model tests, first the
6. Conclusion and Recommendation

geometry-based sampling algorithm was used to select a number of fields on each wafer and the overlay data from these fields were used to derive the reduced model under a certain selection strategy (adjusted $R^2$, PSS or ALA) independently. Then the same measurements were fitted to the reduced model. On the contrary, for the global-model tests, a unique reduced model was determined for all the wafers from the same lot. During this kind of tests, a certain number of fields on a sub-lot of wafers were selected to be measured (with the geometry-based sampling algorithm). The measurements were first used to derive the reduced model and then fitted in the model to calculate the process corrections, which would be used by the whole lot. The empirical tests demonstrate that ALA is the most successful one among the three model selection methods. By taking the amount of random noise into account, ALA avoids over-fitting and achieves a good balance between prediction accuracy and the complexity of correction models.

Therefore, as the answer to the question how to achieve an acceptable overlay with a minimum measurement effort on the overlay metrology tool, we recommend the use of geometry-based sampling together with the model selection strategy of ALA.

6.2 Future Work

In Chapter 5, we did a general survey of nonparametric models and discussed about the potential benefits of those models. However, we have not implemented any of these models yet because of the time limitation. More investigations about nonparametric models can be done in the future. Particularly, in some test data, there seems to be too much non-linear errors within a field, even the application of CPE cannot reduce the overlay to acceptable tolerance (See Appendix C, Figure C.1 and Figure C.2). It is desirable to have some alternative models which are more powerful and effective than the current ones.

Furthermore, the effectiveness of geometry-based sampling is based on its dedication to polynomial regression models. Hence, we may need new sampling strategies to collect overlay data and fit the nonparametric models. According to literature, space-filling designs, especially uniform designs [3] can be good choices.
Bibliography


Appendix A

Further Explanation of Metrology Tools

The overlay is measured with the displacement of an exposed point on the wafer with respect to the position of that same point in the previously exposed layer. Particularly, the “box in box” feature [9] exposed on the boundary of the field during each exposure (i.e., the strip between exposed fields) is designed to measure the overlay, see Figure A.1(a). The inner box is designed to be patterned in the center of the outer box. Thus if there is no overlay, we would have \( x_1 = x_2 \) and \( y_1 = y_2 \).

![Measurement boxes and total overlay](image)

**Figure A.1: overlay measurement**

Though overlay can be divided into inter- and intra-field\(^1\), only the total overlay are actually measured by the \( x\)- and \( y\)- axis displacements between the two boxes patterned in the present and previous layers, i.e.,

\[ (d_{x,1}, d_{y,1}) \]

---

\(^1\)Here, we use \( d_X, d_Y \) to denote inter-field errors and \( d_x, d_y \) intra-field errors
A. Further Explanation of Metrology Tools

\[ d_{x+X} = \frac{x_1 - x_2}{2}, \quad (A.1) \]
\[ d_{y+Y} = \frac{y_1 - y_2}{2}. \quad (A.2) \]

In addition to the displacements of the sampled overlays, the coordinates of sampled locations are the data that can be used to detect the assignable causes of the overlay.
Appendix B

Pseudo Codes

Pseudo Code for the Strategy of Adjusted $R^2$

**Algorithm 2**

**Input**: overlay data from the $N$ selected fields of a wafer $W$

**Output**: appropriate degrees of polynomials for each parameter used by HOPC models

Initialize the degrees of polynomials for each parameter;

$N \leftarrow$ the number of fields selected to be measured;

$error \leftarrow$ the original overlay on $W$;

$SSTO_x \leftarrow \sum_{i=1}^{N} (error_{i,x} - \overline{error_x})^2$;

$SSTO_y \leftarrow \sum_{i=1}^{N} (error_{i,y} - \overline{error_y})^2$;

for degree $k_x, k_y \leftarrow 1$ to 5 do

$d_x \leftarrow$ the overlay after the $k_x$-th order polynomial corrections in $X$-axis;

$d_y \leftarrow$ the overlay after the $k_y$-th order polynomial corrections in $Y$-axis;

$SSR_x \leftarrow \sum_{i=1}^{N} (d_{i,x} - \overline{error_x})^2$;

$SSR_y \leftarrow \sum_{i=1}^{N} (d_{i,y} - \overline{error_y})^2$;

$R^2_x = SSR_x / SSTO_x$;

$R^2_y = SSR_y / SSTO_y$;

$P_x \leftarrow$ degree of freedom in $X$-axis;

$P_y \leftarrow$ degree of freedom in $Y$-axis;

$adjusted R^2_x = 1 - (1 - R^2_x) \cdot (N - 1)/(N - P_x)$;

$adjusted R^2_y = 1 - (1 - R^2_y) \cdot (N - 1)/(N - P_y)$;

Store $adjusted R^2_x, adjusted R^2_y$;

end

Find out the $k_x, k_y$ with the maximum $adjusted R^2_x, adjusted R^2_y$;

return $k_x, k_y$
Pseudo Code for the ANOVA-Like Approach

**Algorithm 3**

**Input:** overlay data from the $N$ selected fields of a wafer $W$

**Output:** appropriate degrees of polynomials for each parameter used by HOPC models

Initialization;

for each field $f_i \in W$, $i \in \{1, \ldots, N\}$ do

\[ \text{Variance}_{i,x} \leftarrow \text{the variance of the overlays in X-axis for } f_i \; ; \]

\[ \text{Variance}_{i,y} \leftarrow \text{the variance of the overlays in Y-axis for } f_i \; ; \]

end

\[ D_x \leftarrow \sqrt{\sum_{i=1}^{N} \text{Variance}_{i,x}} \; ; \]

\[ D_y \leftarrow \sqrt{\sum_{i=1}^{N} \text{Variance}_{i,y}} \; ; \]

$\theta \leftarrow$ an empirical constant ;

for degree $k_x \leftarrow 1$ to 5 do

\[ r_x \leftarrow \text{the residue in X-axis after the corrections with } k_x\text{-th order model} ; \]

\[ \text{SquareSum}_x \leftarrow \sum_{i=1}^{N} r_x^2 \; ; \]

\[ C_x \leftarrow \sqrt{\text{SquareSum}_x / N} \; ; \]

if $C_x < D_x \times \theta$ then

break ;

end

end

for degree $k_y \leftarrow 1$ to 5 do

\[ r_y \leftarrow \text{the residue in Y-axis after the corrections with } k_y\text{-th order model} ; \]

\[ \text{SquareSum}_y \leftarrow \sum_{i=1}^{N} r_y^2 \; ; \]

\[ C_y \leftarrow \sqrt{\text{SquareSum}_y / N} \; ; \]

if $C_y < D_y \times \theta$ then

break ;

end

end

return $k_x, k_y$
Pseudo Code for the Strategy of Partial Sum of Squares

**Algorithm 4:**

**Input:** overlay data from the $N$ selected fields of a wafer $W$

**Output:** a list of polynomial terms for each parameter used by HOPC models

Initialization;

\[ ovr_{old} \leftarrow \text{the original error before any correction} \; \]

for each direction $t = x$ or $y$ do

for degree $k = 1$ to 5 do

\[ L \leftarrow \text{the selected terms with order less than } k \; \]

repeat

for each term $v$ with order $k$ do

\[ L \leftarrow L + v \; \]

\[ ovr_{new} \leftarrow \text{the overlay after corrections with the model including term } v \text{ in } L \; \]

Store $ovr_{new}$

end

\[ ovr_{min} \leftarrow \text{the minimum of } ovr_{new} \; \]

\[ ovr_{improvement} \leftarrow (ovr_{old} - ovr_{min}) \; \]

\[ ovr_{old} \leftarrow ovr_{min} \; \]

$v_{min}$ is stored in the list $L_t$;

until $ovr_{improvement} < \tau$  /* $\tau$ is an empirical constant */

end

end

return $L_x, L_y$
Appendix C

Experimental Results

Correction Results by Adjusted $R^2$

The selected polynomial degrees when the adjusted $R^2$ is applied to the 25-wafer lot can be seen from Table C.1, C.2 and C.3.

<table>
<thead>
<tr>
<th># sub-lot</th>
<th>Tx</th>
<th>Ty</th>
<th>Rs</th>
<th>Ra</th>
<th>Ms</th>
<th>Ma</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table C.1: Selected polynomial degrees by adjusted $R^2$ (15 fields measured)

<table>
<thead>
<tr>
<th># sub-lot</th>
<th>Tx</th>
<th>Ty</th>
<th>Rs</th>
<th>Ra</th>
<th>Ms</th>
<th>Ma</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table C.2: Selected polynomial degrees by adjusted $R^2$ (21 fields measured)
C. Experimental Results

<table>
<thead>
<tr>
<th># sub-lot</th>
<th>Tx</th>
<th>Ty</th>
<th>Rs</th>
<th>Ra</th>
<th>Ms</th>
<th>Ma</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table C.3: Selected polynomial degrees by adjusted $R^2$ (full-wafer measured)

Correction Results by PSS + Adjusted $R^2$

For convenience, we use indices to denote different terms. The mapping between indices and polynomial terms are formulated as follows:

\[
\begin{align*}
\text{HOPC design matrix} &= \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ c00 & c10 & c01 & c20 & c11 & c02 & c30 \\ \text{ones}(m,1) & xc & yc & xc.~2 & xc.*yc & yc.~2 & xc.~3 \\ 8 & 9 & 10 & 11 & 12 & 13 & 14 \\ c21 & c12 & c03 & c40 & c31 & c22 & c13 \\ xc.~2.*yc & xc.*yc.~2 & yc.~3 & xc.~4 & xc.~3.*yc & xc.~2.*yc.~2 & xc.*yc.~3 \\ 15 & 16 & 17 & 18 & 19 & 20 & 21 \\ c04 & c50 & c41 & c32 & c23 & c14 & c05 \\ yc.~4 & xc.~5 & xc.~4.*yc & xc.~3.*yc.~2 & xc.~2.*yc.~3 & xc.*yc.~4 & yc.~5 \end{bmatrix};
\end{align*}
\]

Further, the coefficients for $T_x$ (translation in $x$ direction), $T_y$ (translation in $y$ direction), $R_s$ (symmetric rotation), $R_a$ (asymmetric rotation), $M_s$ and $M_a$ (symmetric and asymmetric magnification) are ordered as shown in Table C.2.

<table>
<thead>
<tr>
<th>parameters:</th>
<th>$T_x$</th>
<th>$T_y$</th>
<th>$R_s$</th>
<th>$R_a$</th>
<th>$M_s$</th>
<th>$M_a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>indices:</td>
<td>1 - 21</td>
<td>22 - 42</td>
<td>43 - 63</td>
<td>64 - 84</td>
<td>85 - 105</td>
<td>106 - 126</td>
</tr>
</tbody>
</table>

Table C.4: Coefficients for $T_x, T_y, R_s, R_a, M_s$ and $M_a$
We applied the ordinary PSS strategy (with monomials) and the PSS with Zernikes to the 25-wafer lot. The selected relevant terms can be see from Table C.5, C.6 and C.7. Note that the mapping between indices and polynomial terms for PSS with Zernikes are different from those of ordinary PSS. It follows the grouping results in Section 4.3.4.

<table>
<thead>
<tr>
<th>#sub-lot</th>
<th>Selected Terms by PSS</th>
<th>Selected Terms by PSS with Zernikes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 2 7 22 25 27 29 43 64 85 106</td>
<td>1 2 22 26 27 43 64 85 106</td>
</tr>
<tr>
<td>2</td>
<td>1 2 7 22 24 25 43 64 85 106</td>
<td>1 2 22 27 43 64 85 106</td>
</tr>
<tr>
<td>3</td>
<td>1 7 8 22 24 31 43 64 85 106</td>
<td>1 4 22 27 28 43 64 85 106</td>
</tr>
<tr>
<td>4</td>
<td>1 22 24 28 31 43 64 85 106</td>
<td>1 4 22 43 64 85 106</td>
</tr>
<tr>
<td>5</td>
<td>1 5 6 8 22 24 28 31 36 43 64 85 106</td>
<td>1 22 25 26 28 43 64 85 106</td>
</tr>
</tbody>
</table>

Table C.5: Selected terms by PSS and PSS with Zernikes (15 fields measured)

<table>
<thead>
<tr>
<th>#sub-lot</th>
<th>Selected Terms by PSS</th>
<th>Selected Terms by PSS with Zernikes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 9 10 22 24 25 29 31 43 64 85 106</td>
<td>1 22 27 29 43 64 85 106</td>
</tr>
<tr>
<td>2</td>
<td>1 9 22 24 25 43 64 85 106</td>
<td>1 22 31 43 64 85 106</td>
</tr>
<tr>
<td>3</td>
<td>1 9 10 22 24 43 64 85 106</td>
<td>1 4 22 43 64 85 106</td>
</tr>
<tr>
<td>4</td>
<td>1 9 22 24 31 43 64 85 106</td>
<td>1 4 22 31 43 64 85 106</td>
</tr>
<tr>
<td>5</td>
<td>1 8 9 22 24 43 64 85 106</td>
<td>1 4 22 43 64 85 106</td>
</tr>
</tbody>
</table>

Table C.6: Selected terms by PSS and PSS with Zernikes (21 fields measured)

<table>
<thead>
<tr>
<th>#sub-lot</th>
<th>Selected Terms by PSS</th>
<th>Selected Terms by PSS with Zernikes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 3 22 24 29 31 43 64 85 106</td>
<td>1 3 22 43 64 85 106</td>
</tr>
<tr>
<td>2</td>
<td>1 9 10 22 24 31 43 64 85 106</td>
<td>1 22 30 31 43 64 85 106</td>
</tr>
<tr>
<td>3</td>
<td>1 3 22 24 31 43 64 85 106</td>
<td>1 3 22 43 64 85 106</td>
</tr>
<tr>
<td>4</td>
<td>1 22 24 30 31 43 64 85 106</td>
<td>1 22 43 64 85 106</td>
</tr>
<tr>
<td>5</td>
<td>1 3 22 24 30 31 43 64 85 106</td>
<td>1 3 22 43 64 85 106</td>
</tr>
</tbody>
</table>

Table C.7: Selected terms by PSS and PSS with Zernikes (full-wafer measured)
Correction Results by ANOVA-like Approach

The selected polynomial degrees when we apply ALA to the 25-wafer lot can be seen from following tables.

<table>
<thead>
<tr>
<th># sub-lot</th>
<th>Tx</th>
<th>Ty</th>
<th>Rs</th>
<th>Ra</th>
<th>Ms</th>
<th>Ma</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table C.8: Selected polynomial degrees with ALA (15 fields measured)

<table>
<thead>
<tr>
<th># sub-lot</th>
<th>Tx</th>
<th>Ty</th>
<th>Rs</th>
<th>Ra</th>
<th>Ms</th>
<th>Ma</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table C.9: Selected polynomial degrees with ALA (21 fields measured)

<table>
<thead>
<tr>
<th># sub-lot</th>
<th>Tx</th>
<th>Ty</th>
<th>Rs</th>
<th>Ra</th>
<th>Ms</th>
<th>Ma</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table C.10: Selected polynomial degrees with ALA (full-wafer measured)

As we have seen, for those wafers with assignable overlay, both CPE and ALA-based correction can improve the overlay significantly, but the latter is much cheaper with respect
to the measurement costs. During the test, we also found some lot for which both CPE and ALA-based correction failed to improve the overlay on some wafers, see Figure C.1 as illustration. The test lot contains 12 wafers, each with 149 fields. We used the measurement data from the first 4 wafers and apply the calculated corrections to the whole lot later. In this case, we can find that 15-field measurement is not enough to ensure significant overlay improvement any more.

![Figure C.1: Overlay improvements after CPE and ALA](image)

Further examination showed that there exists some outliers for certain wafers from this lot. Thus we tried to remove those outliers and did the test again. From Figure C.2, we can find that for \( X \)-axis, the overlay are still not improved much, but for \( Y \)-axis, both CPE and ALA improved the overlay significantly. The possible explanation for this can be that there exists too much wafer-to-wafer variations and this is a contradiction to our basic assumptions.

![Figure C.2: Overlay improvements after CPE and ALA](image)