Optimizing hybrid metrology: rigorous implementation of Bayesian and combined regression

Mark-Alexander Henn
Richard M. Silver
John S. Villarrubia
Nien Fan Zhang
Hui Zhou
Bryan M. Barnes
Bin Ming
andrás E. Vladár
Optimizing hybrid metrology: rigorous implementation of Bayesian and combined regression

Mark-Alexander Henn,a,* Richard M. Silver,a John S. Villarrubia,a Nien Fan Zhang,b Hui Zhou,a Bryan M. Barnes,a Bin Ming,a and András E. Vladára

aNational Institute of Standards and Technology, Engineering Physics Division, 100 Bureau Drive MS 8212, Gaithersburg, Maryland 20899-8212, United States
bNational Institute of Standards and Technology, Statistical Engineering Division, 100 Bureau Drive MS 8212, Gaithersburg, Maryland 20899-8212, United States

Abstract. Hybrid metrology, e.g., the combination of several measurement techniques to determine critical dimensions, is an increasingly important approach to meet the needs of the semiconductor industry. A proper use of hybrid metrology may yield not only more reliable estimates for the quantitative characterization of threedimensional (3-D) structures but also a more realistic estimation of the corresponding uncertainties. Recent developments at the National Institute of Standards and Technology feature the combination of optical critical dimension measurements and scanning electron microscope results. The hybrid methodology offers the potential to make measurements of essential 3-D attributes that may not be feasible otherwise. However, combining techniques gives rise to essential challenges in error analysis and comparing results from different instrument models, especially the effect of systematic and highly correlated errors in the measurement on the $\chi^2$ function that is minimized. Both hypothetical examples and measurement data are used to illustrate solutions to these challenges. © The Authors. Published by SPIE under a Creative Commons Attribution 3.0 Unported License. Distribution or reproduction of this work in whole or in part requires full attribution of the original publication, including its DOI.

Keywords: hybrid metrology; electromagnetic simulation; sensitivity and uncertainty evaluation; Bayesian data analysis.

Paper 15086P received May 20, 2015; accepted for publication Oct. 8, 2015; published online Nov. 12, 2015.

1 Introduction

Hybrid metrology, e.g., the combination of distinct measurement techniques to determine critical dimensions (CDs), is an increasingly important approach to meet the needs of the semiconductor industry. A proper use of hybrid metrology may yield not only more reliable estimates for the quantitative characterization of three-dimensional (3-D) structures but also a more realistic estimation of the corresponding uncertainties. Ideally it helps to reduce the overall uncertainties by combining the individual strengths of each of the measurement techniques, making subnanometer uncertainties a realistic goal as CDs approach 10 nm.

Recent developments in hybrid metrology at the National Institute of Standards and Technology (NIST) feature the combination of optical critical dimension (OCD) and scanning electron microscope (SEM) measurements. The challenges and possible solutions have been outlined by some of these authors in a previous proceedings paper. Various methods have been presented to combine measurement results from different tool platforms, revealing two related but distinct challenges. There must be an overlapping parameter set for combined regression such that each individual parametric geometry must share at least one parameter in common (e.g., height). Additionally a priori, each individual method must also yield parametric values that together with their uncertainties are statistically consistent, usually quantified by a Z-test; see Ref. 3 for more details.

This paper can be seen as a continuation of that work and Ref. 3, with an emphasis on the proper treatment of the measurement errors, including highly correlated and systematic errors and their influence on hybrid metrology. Tool-induced errors for OCD and scale errors for SEM are investigated, and it is shown how the parametric uncertainties can be decreased if those issues are addressed accurately in the hybridization.

Since the term hybrid metrology has gained increased significance in the dimensional metrology community outside NIST, we will start this work with a short overview of two of the most common techniques in Sec. 3, namely the Bayesian approach and combined regression. The measured targets and the generalized parameter sets that describe them are discussed in Sec. 4 before we give a detailed description of the performed error analysis for both OCD and SEM in Sec. 4 and their impact on the hybridization in Sec. 5. We will close with the conclusions in Sec. 6.

2 Hybrid Metrology

Hybrid metrology has gained significant recognition in recent times as an approach to considerably reduce parametric uncertainties by combining different measurements of the same measurand. We want to use this section to identify the main differences and similarities of two of the most common hybrid approaches, namely the use of a priori information in a Bayesian sense and combined regression. We start with the Bayesian approach and continue with combined regression. In order to keep the notation simple, throughout this section we will assume that only two measurement techniques are combined with each individual method, yielding a statistically consistent set of parameters. Note that even if
some of the notations are different, the presented approaches are equivalent to those given in Ref. 2. Additional information for those who are not familiar with all of the terminology of Bayesian data analysis can be found in Refs. 8–11.

2.1 Bayesian Approach

The Bayesian approach treats information provided by each of the measurement tools quite differently. The first tool provides \( m \) values of measurement data that can be described as a vector in an \( m \)-dimensional real vector space

\[
y = (y_1, \ldots, y_m)^T \in \mathbb{R}^m
\]

that contains only indirect information about the quantity of interest. We therefore need to analyze the data in terms of an inverse problem. A common approach to solve an inverse problem is to set it up as a regression problem. Initially, we need to provide a model function

\[
f : \mathbb{R}^n \rightarrow \mathbb{R}^m, \quad f(p) = [f_1(p), \ldots, f_m(p)]^T
\]

that maps the parameters of interest (e.g., the height, the width, etc.), that are identified with a vector in an \( n \)-dimensional real vector space \( p = (p_1, \ldots, p_n)^T \), to the simulated quantities. These simulated data are generated in the same \( m \)-dimensional space defined by the measurement setup similar to the actual experiment at NIST.\(^{14,15}\)

By subtracting the prior distribution or, more precisely, its logarithm, from the function in Eq. (5), the negative log-likelihood, we get a function that is proportional to the negative logarithm of the posterior probability distribution. If we again ignore normalization constants we get the function that serves as the modified \( \chi^2 \) function

\[
\chi^2(p) = [f(p) - y]^T V^{-1} [f(p) - y] + (p - \mu_p)^T \Sigma_p^{-1} (p - \mu_p).
\]

Note that the second term that reflects the prior information acts as a penalty or regularization term, penalizing possible solutions to the inverse problem for measurement tool 1 that are not consistent with the prior information. The function in Eq. (5) is finally minimized to find the parameter vector estimate \( \hat{p} \).

2.2 Combined Regression

In combined regression we start with two distinct sets of measurement data, \( y_A \) and \( y_B \), that come from different measurement techniques. Their model functions, \( f_A(p_A) \) and \( f_B(p_B) \), depend on parameter vectors \( p_A \) and \( p_B \), respectively. The models must have at least one common model parameter in order to perform combined regression. Determining what the common parameters of the two models are can be a challenging task; see Ref. 2 for further details. In combined regression we define the combined \( \chi^2 \) function to be the sum of the individual \( \chi^2 \) functions for each of the tools. Note that this is only possible if we assume that the two measurements are independent of each other

\[
\chi^2(p_{AUB}) = \chi_A^2(p_A) + \chi_B^2(p_B).
\]

Here \( p_{AUB} \) is the vector that consists of the union of the elements in \( p_A \) and \( p_B \). The solution \( \hat{p}_{AUB} \) to the inverse problem in combined regression is then found by minimizing the above combined \( \chi^2 \) function.

3 Measured Targets and Generalized Parameter Sets

The investigated targets and the geometrical parameterizations used have already been described in detail in Ref. 4, so we will only give a brief overview. We investigate finite 30-line arrays of Si on Si with a thin native conformal oxide (see Fig. 6). The nominal widths are 14, 16, and 18 nm. A schematic representation of the geometrical parameterizations can be found in Fig. 6. For the OCD analysis, the geometry is fully characterized by the height, width, \( \Delta_{\text{top}} \), and \( \Delta_{\text{bot}} \) of a single line. The physics by which this geometry interacts with incident light to produce a signal was approximated by the rigorous coupled-wave analysis (RCWA) model that is based on a semianalytical treatment of Maxwell’s equations.\(^{\text{19,21}}\) The OCD data used in this study have been generated using this RCWA model for 30 lines and a measurement setup similar to the actual experiment at NIST\(^{12,22}\).
The middle and hybridization was limited, therefore, to two parameters: the width at half height. Similarly, line position was not important for OCD. Hybridization with OCD is a technique that averages over a comparatively large area, making it susceptible to certain types of errors. Consequently, for hybridization with OCD, we choose the INA to be normally distributed with a mean of 0.13 and a standard deviation of 0.01. Figures 3 and 4 show a graphical representation of the first 752 entries of the sample covariance matrix that correspond to the first four correlated errors. For example, one can imagine that being slightly off axis in the illumination will affect the symmetry of the entire set of collected images. Determining such correlations from measurement data alone is often not possible, and other methods need to be developed in order to give a quantitative description of those effects. We use the Monte Carlo method, which is based on the following reasoning. If we denote by $\nu$ the vector of $k$ fixed parameters of the measurement setup, we can make use of the following more general model function:

$$\mathbf{f}: \mathbb{R}^{n-k} \rightarrow \mathbb{R}^m, \quad \mathbf{f}(\mathbf{p}, \nu) = [f_1(\mathbf{p}, \nu), \ldots, f_m(\mathbf{p}, \nu)]^T.$$  (8)

The effect that a slight deviation of the parameters $\nu$ from the nominal values $\nu_0$ has upon the simulated image can then be estimated from the following five steps:

1. Assume a distribution for $\nu$, based on tool specifications, expert knowledge, etc.
2. Draw a sample $\{\nu_i\}_{i=1}^N$ from the distribution.
3. Calculate $\{\mathbf{f}(\mathbf{p}, \nu_i) = [f_1(\mathbf{p}, \nu_i), \ldots, f_m(\mathbf{p}, \nu_i)]^T\}_{i=1}^N$.
4. Define $\mathbf{r}_i = \mathbf{f}(\mathbf{p}, \nu_0) - \mathbf{f}(\mathbf{p}, \nu_i)$.
5. Calculate the sample covariance matrix via

$$\mathbf{V} = \frac{1}{N-1} \sum_{i=1}^N (\mathbf{r}_i - \bar{\mathbf{r}})(\mathbf{r}_i - \bar{\mathbf{r}})^T.$$  (9)

In this paper, the $\nu$ vector included components for the collection numerical aperture (CNA), illumination numerical aperture (INA), focus heights, and phase. This Monte Carlo procedure, therefore, estimates the effect of errors propagated from variations in those instrument parameters, e.g., we choose the INA to be normally distributed with a mean of 0.13 and a standard deviation of 0.01. Figures 3 and 4 show a graphical representation of the first 752 entries of the sample covariance matrix that correspond to the first four focus heights in $X$ polarization for phase, focus, illumination, and CNA variations, along with the respective concatenated images.

One can clearly see both the positive and negative correlation between errors as colored areas off the diagonal. In order to have a reliable model for the measurement errors, it
is therefore very important to account for this effect in the covariance matrix that is being used in the \( \chi^2 \) function [see Eq. (10)]. This has an influence not only on the best fit values, but also on the estimation of the parametric uncertainties or, more precisely, the covariance matrix \( \Sigma \), diagonal elements of which are the uncertainties in the estimated parameters, given by

\[
\sum = (J^T V^{-1} J)^{-1}.
\]

Here \( J \) denotes the Jacobian matrix, i.e., the matrix of all first-order partial derivatives of the vector-valued model function \( f \), at the best fit vector \( \hat{p} \). A comparison of the parametric uncertainties based on using a diagonal \( V \) and the full \( V \) in Eq. (10) can be found in Table I.

Note that there is a notable difference between the estimated parametric uncertainties, with the most significant change in the uncertainty of the width. It is also very important to note that a given parametric uncertainty might increase or decrease if correlations in the measurement data are taken into account. Thus, a general statement about the effect correlated errors have on the estimated parametric uncertainties cannot be given, and it must be investigated separately for each new problem.

### 4.2 Scaling Errors in Scanning Electron Microscope

The biggest contribution to the SEM’s measurement error in this experiment is due to pixel calibration. Errors in this calibration directly influence the obtained values for the CDs. The usual approach is to simply add the uncertainty due to the calibration in quadrature with the estimated parametric uncertainty after the reconstruction. The estimation of the parameters of interest, i.e., the vector \( p \) and their uncertainties in combined regression, are based on the combination of the OCD and the SEM data, while the error induced by pixel calibration only affects SEM data; hence, it is not possible to simply include the uncertainty due to the calibration afterward. We will therefore model the effect a variation in the scale has on the measurement in a simple way, multiplying the model parameter vector \( p = (\text{width}, \Delta \text{top})^T \) by a scaling parameter \( \kappa \) such that the modified model is given by

\[
\tilde{f}: \mathbb{R}^3 \rightarrow \mathbb{R}^m, \quad \tilde{f}(\kappa, p) = f(\kappa \cdot p)
\]

with \( f \) being the SEM’s model function. In this description, \( \kappa = 1 \) corresponds to no scale error, \( \kappa = 1.01 \) to 1% scale error, etc. It is clear that this approach leads to a high parametric correlation in the parameters of the model. We will explain this effect using a simple model with an added scaling factor in the following:

Let

\[
f: \mathbb{R} \rightarrow \mathbb{R}^m, \quad f(x) = [f_1(x), \ldots, f_m(x)]^T, \quad \text{and} \quad J = (J_{i1})_{i=1, \ldots, m}, \quad J_{i1} = Df_i
\]

be a model function depending on only one parameter \( x \) and denote by \( Df_i \) the derivative of \( f_i \) with respect to this parameter. If we assume the measurement errors to be independent and identically distributed (i.i.d.) with unit variance [hence \( V = (\delta_{ij})_{i,j=1, \ldots, m} \)] we have for the estimated covariance matrix

\[
\sum = (J^T V^{-1} J)^{-1}, \quad \text{with} \quad J^T V^{-1} J = \sum_{i=1}^m |Df_i|^2,
\]

which is well defined if \( \det(J^T V^{-1} J) = \sum_{i=1}^m |Df_i|^2 \neq 0. \)
Now assume that we add a scale parameter to the above model by defining a slightly modified function

\[ F: \mathbb{R}^2 \to \mathbb{R}^m, \quad F(\kappa, x) = f(\kappa \cdot x), \]  
and  
\[ J = (J_{i,j})_{i=1,\ldots,m; j=1,2}, \quad J_{i,1} = x \cdot Df_i, \]
\[ J_{i,2} = \kappa \cdot Df_i. \quad (14) \]

The estimated covariance matrix for this two-parameter model is then given by

\[ \sum = (J^T \mathbf{V}^{-1} J)^{-1}, \quad \text{with} \]
\[ J^T \mathbf{V}^{-1} J = \left[ \begin{array}{c} \sum_{i=1}^m x^2[Df_i]^2 \\ \sum_{i=1}^m \kappa x[Df_i]^2 \\ \sum_{i=1}^m \kappa x^2[Df_i]^2 \end{array} \right]. \quad (15) \]

![Fig. 4 Estimated covariance matrices for variations in the (a) illumination numerical aperture (INA) and (b) collection numerical aperture (CNA). INA normally distributed with a mean of 0.13 and a standard deviation of 0.01, CNA normally distributed with a mean of 0.95 and a standard deviation of 0.1. The variances induced by INA and CNA variations are one magnitude less than those from phase errors.](image)

Table 1: Estimates of the parametric uncertainties for simulated OCD data using diagonal \( \mathbf{V} \), i.e., only accounting for uncorrelated random noise, and full \( \mathbf{V} \), i.e., taking correlations into account.

<table>
<thead>
<tr>
<th></th>
<th>Width (nm)</th>
<th>( \Delta )top (nm)</th>
<th>( \Delta )bot (nm)</th>
<th>Height (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\rho} )</td>
<td>17.09</td>
<td>2.04</td>
<td>0.10</td>
<td>35.37</td>
</tr>
<tr>
<td>( \sigma ) diagonal ( \mathbf{V} )</td>
<td>0.05</td>
<td>1.37</td>
<td>0.53</td>
<td>0.97</td>
</tr>
<tr>
<td>( \sigma ) full ( \mathbf{V} )</td>
<td>0.14</td>
<td>1.29</td>
<td>0.63</td>
<td>0.92</td>
</tr>
</tbody>
</table>

However,

\[ \det(J^T \mathbf{V}^{-1} J) = \sum_{i=1}^m x^2[Df_i]^2 \cdot \sum_{i=1}^m \kappa x[Df_i]^2 \]
\[-\left\{ \sum_{i=1}^m \kappa x[Df_i]^2 \right\}^2 = 0. \quad (16) \]

Since the above term is always equal to zero, the estimated covariance matrix \( \sum \) is not defined and we cannot assign a parametric uncertainty. Since we have prior information about the scale, the actual error lies between 1% and 2%; we can use the Bayesian approach as described in Refs. 2 and 9 under the premise that the prior information can be expressed in terms of normal distributions. The prior information on the parameter \( \kappa \) is treated as an additional data point the model function has to account for, such that we have a function that still depends on \( \kappa \) and \( x \) but now maps into an \((m+1)\)-dimensional space, with the \((m+1)\)th value simply being \( \kappa \). This also adds additional terms to the Jacobian, \( J_{m+1,1} = (\partial x/\partial \kappa) = 1 \) and \( J_{m+1,2} = (\partial x/\partial \kappa) = 0 \), such that

\[ \hat{F}: \mathbb{R}^2 \to \mathbb{R}^{m+1}, \quad \hat{F}(\kappa, x) = [F(\kappa, x), \kappa]^T, \quad \text{and} \]
\[ J_{m+1,1} = 1, \quad J_{m+1,2} = 0. \quad (17) \]

Since we know that \( \kappa \sim \mathcal{N}(1, \sigma^2) \), we add an additional entry to \( \mathbf{V} \), namely \( V_{m+1,m+1} = \sigma^2 \) with \( \sigma = 0.01 - 0.02 \), and obtain the estimated covariance matrix, again using Eqs. (14) and (15),

\[ \sum = (J^T \mathbf{V}^{-1} J)^{-1}, \quad \text{with} \]
\[ J^T \mathbf{V}^{-1} J = \left[ \begin{array}{c} \sum_{i=1}^m x^2[Df_i]^2 + \frac{1}{\sigma^2} \sum_{i=1}^m \kappa x[Df_i]^2 \\ \sum_{i=1}^m \kappa x[Df_i]^2 \end{array} \right]. \quad (18) \]
and
\[
\det(J^T V^{-1} J) = \left\{ \sum_{i=1}^{m} \chi^2[Df_i]^2 + \frac{1}{\sigma_e^2} \right\} \cdot \sum_{i=1}^{m} \chi^2[Df_i]^2 - \left\{ \sum_{i=1}^{m} \kappa \chi[Df_i]^2 \right\}^2.
\]
(19)

Equation (19) implies that as long as the prior knowledge about \( \kappa \) is not too vague, i.e., \( \sigma_e \) is not too large, the above term is in general not equal to zero and the estimated covariance matrix is well defined. A graphical representation of the above-described phenomenon for the measured SEM data is shown in Fig. 5. Note that the \( \chi^2 \) surface with the added prior information about \( \kappa \) has a distinct minimum at \( \hat{p} = (1.17.01 \text{ nm}, 2.86 \text{ nm}) \), while it is hard to determine where the minimum is for the \( \chi^2 \) surface without prior information. In fact, there is not a single distinct minimum but an infinite set of possible minima. Obviously, \( \hat{p} = (1.17.01 \text{ nm}, 2.86 \text{ nm}) \) is also the minimum for the \( \chi^2 \) without prior information, but so is any vector \( \hat{p} = [\kappa, (17.01/\kappa) \text{ nm}, (2.86/\kappa) \text{ nm}] \) with \( \kappa \neq 0 \). Defining a parametric uncertainty is therefore not possible. In contrast, the error estimation for the model with prior information for the scale \( \kappa \) yields a 2% parametric uncertainty if we assume an error of 2% in the scale as expected; here, this strict linearity only holds since the random errors in the SEM data are much smaller than those attributed to the scale error.

5 Results
We now combine the results that we found in the previous section with the hybridization of OCD and SEM data by combined regression. As pointed out in Sec. 4, this is done by minimizing the sum of the respective \( \chi^2 \) functions. Note that we use prior information about the scale \( \kappa \), so that the \( \chi^2 \) function for the SEM data is modified as shown in Eq. (9). The individual \( \chi^2 \) surfaces in dependence on the width and \( \Delta \text{top} \) are shown in Figs. 6 and 7. For these plots, the height and \( \Delta \text{bot} \) have been fixed for OCD. The plots also show the individual minima and the assigned parametric uncertainties. The \( \chi^2 \) surface from the combined regression is shown in Fig. 8, and the results from the combined regression are presented in Table 2. The combined minimum is close to the SEM’s minimum and the parametric uncertainty...
uncertainties for combined regression are lower than the individual ones, even for the parameters that are only present in the OCD model. This is due to the strong parametric correlations in the models.

### Table 2  Parameter estimates and parametric uncertainties obtained from combined regression.

<table>
<thead>
<tr>
<th></th>
<th>Width (nm)</th>
<th>Δtop (nm)</th>
<th>Δbot (nm)</th>
<th>Height (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{\text{hybrid}}$</td>
<td>17.01</td>
<td>2.94</td>
<td>0.59</td>
<td>36.11</td>
</tr>
<tr>
<td>$\sigma_{\text{hybrid}}$</td>
<td>0.23</td>
<td>0.04</td>
<td>0.58</td>
<td>0.21</td>
</tr>
<tr>
<td>$\sigma_{\text{SEM}}$</td>
<td>0.34</td>
<td>0.06</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>$\sigma_{\text{OCD}}$</td>
<td>0.14</td>
<td>1.29</td>
<td>0.63</td>
<td>0.92</td>
</tr>
</tbody>
</table>

### 6 Discussion and Conclusion

Following the approach outlined in Ref. 3, we studied the challenges in hybrid metrology due to measurement errors. Those included highly correlated tool-induced errors for the OCD data and systematic errors due to scaling errors in SEM. We have demonstrated how slight variations in the measurement setup for OCD, e.g., in the focus heights, the phase, INA, and CNA lead to highly correlated errors in the measurement data that manifest themselves as nonzero elements in the sample covariance matrix $V$. Including those off-diagonal elements in the estimation of a parametric uncertainty can lead to either an increased or a decreased parametric uncertainty compared to the case where only the diagonal of the $V$ matrix is used, depending on the individual nature of that particular full $V$ matrix. Furthermore, we demonstrated the influence of scaling errors on the analysis of SEM data. Attempting to account for such scale errors by including the scale as a fully free parameter would lead to...
unreasonable results due to strong, or even perfect correlations. This problem has been solved using prior information about the scale in a Bayesian approach. Finally, we demonstrated how the more sophisticated error analyses could be used in the hybridization of OCD and SEM data. With the proper treatment of those errors, we could achieve a subnanometer parametric uncertainty. It is important to note that the presented framework can be extended to include additional measurement techniques, such as atomic force microscopy or CD small angle X-ray scattering. In addition, it may also be applied across a homogeneous multiple-tool platform. However, for every added measurement technique, it is crucial to perform a careful error analysis in order to use its full capabilities.

References