

What is a Good Empirical Model?

Eldar Khaliullin, Yaogang Lian, Mark Davey, Xin Zhou*
Luminescent Technologies, Palo Alto, CA, USA 94303

ABSTRACT

An accurately predictive process model is of utmost importance to the traditional Optical Proximity Correction (OPC), the leading-edge Inverse Lithography Technology (ILT), or other simulation software for IC manufacturing. There are many parameters and methods in constructing and calibrating a model. But it is difficult to obtain a good empirical model, partly because the assessment of the final result is lacking in terms of quantitative and objective metrics. We set out to define certain practical guidelines, e.g. Model Effectiveness Standard Index (MESI), for analyzing parameter uncertainty and estimating simulation uncertainty of an empirical model, so that we know what to choose among many similar candidates. The discussion is framed in the estimation theory of statistics.

Keywords: model calibration, estimation, uncertainty, Fisher Information Matrix, Cramér-Rao Lower Bound

1. INTRODUCTION

Calibrating a good empirical model requires a lot of experience, partly because the assessment of the final result is hard to quantify in terms of objective metrics. Application engineers are not to blame, because there are very few quantitative tools available, beyond some jingoistic slogans like “fitting error must be below 1nm rms”, or “model prediction must be very good”. Naturally we get a sinking feeling whenever we are asked to fit the data to within 1nm, knowing full well that the CD-SEM measurement uncertainty is 5nm^[1], let alone that we know that “they” know this too. For model prediction power, practitioners have frequently used a cross-check approach, in which one subset of data is used to verify another subset used in building the model. This is a legitimate but weak form of consistency check. However, when trouble shows up in the cross-check, it does not tell us whether those data points in the verification set are outliers or the model itself ought to be made more robust. Apparently we need more powerful tools to help us face the challenges, and this paper is an attempt at quantifying model assessment.

We have found that the estimation theory of statistics provides a theoretical framework to aid the thinking process, and some of the existing tools and concepts could be borrowed to suit our purpose in model calibration and assessment^[2]. The next section on the mathematical background follows Steven M. Kay^[3].

2. MATHEMATICAL BACKGROUND

If a discrete data set represented by a vector $\mathbf{x} = \{x[0], x[1], \dots, x[N-1]\}$ depends on a single scalar parameter θ , we may design an estimator that finds the true value of θ based on the data set

$$\hat{\theta} = f(x[0], x[1], \dots, x[N-1])$$

Here f is the parameter estimation function, which could be as simple as an algebraic function, or as complex as a lithographic process model calibration process.

2.1 Fisher Information and Cramér-Rao Lower Bound

If we use $p(\mathbf{x}; \theta)$ to denote the probability density function (PDF) of obtaining measurements at \mathbf{x} with the parameter at θ , the Fisher information $I(\theta)$ is defined as

* xzhou@luminescent.com

$$I(\theta) = -E\left[\frac{\partial^2 \ln p(\mathbf{x}; \theta)}{\partial \theta^2}\right] = -\int \frac{\partial^2 \ln p(\mathbf{x}; \theta)}{\partial \theta^2} p(\mathbf{x}; \theta) d\mathbf{x}$$

The theorem of Cramér-Rao Lower Bound (CRLB) states that the variance of any unbiased estimator $\hat{\theta} = f(\mathbf{x})$ must satisfy

$$\text{var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

where the minimum is reached at the true value of θ .

As an example, consider the case of a DC signal in white Gaussian noise

$$x[n] = A + w[n], \quad n = 0, 1, \dots, N-1$$

where A is the true signal DC level, w is the noise with a uniform variance σ^2 . Now the PDF is

$$p(\mathbf{x}; A) = \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x[n]-A)^2}{2\sigma^2}\right] = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{\sum_{n=0}^{N-1} (x[n]-A)^2}{2\sigma^2}\right]$$

And the Fisher information is easily computed

$$I(A) = -E\left[\frac{\partial^2 \ln p(\mathbf{x}; A)}{\partial A^2}\right] = \frac{N}{\sigma^2}$$

Therefore CRLB says $\text{var}(A) \geq \sigma^2/N$. When the usual sample mean is used as the estimator of A in this case, i.e.

$$\hat{\theta} = f(\mathbf{x}) = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$$

the CRLB is reached exactly, in other words, the sample mean is the best estimator of parameter A .

2.2 Extension to a vector parameter

Our model now depends on a series of parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2, \dots, \theta_p\}$. The $p \times p$ Fisher information matrix is defined as

$$[\mathbf{I}(\boldsymbol{\theta})]_{ij} = -E\left[\frac{\partial^2 \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right] = -\int \frac{\partial^2 \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} p(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}$$

The CRLB appears as

$$\text{var}(\hat{\theta}_i) \geq [\mathbf{I}^{-1}(\boldsymbol{\theta})]_{ii} \quad (1)$$

where the true value of $\boldsymbol{\theta}$ is used when the derivatives are taken.

We can also extend the earlier DC signal example to a more sophisticated case of a data set depending on a vector parameter

$$x[n] = s[n; \boldsymbol{\theta}] + w[n], \quad n = 0, 1, \dots, N-1 \quad (2)$$

noting that the signal is non-trivially dependent on the parameter vector θ . A few steps of mathematical maneuvering would produce the Fisher information matrix elements

$$[\mathbf{I}(\theta)]_{ij} = \frac{1}{\sigma^2} \sum_{n=0}^{N-1} \frac{\partial s[n; \theta]}{\partial \theta_i} \frac{\partial s[n; \theta]}{\partial \theta_j} \quad (3)$$

Since we have not explicitly restricted the form of the deterministic signal $s(\theta)$, we shall not derive further the above formula. If θ is a simple scalar, and s is just θ itself, we get back to the earlier DC level example.

2.3 Transformation of parameters

If we want to estimate $\alpha = \mathbf{g}(\theta)$, an r -dimensional function, instead of θ itself, then the covariance matrix of α has a CRLB

$$\mathbf{C}_\alpha \geq \frac{\partial \mathbf{g}(\theta)}{\partial \theta} \mathbf{I}^{-1}(\theta) \frac{\partial \mathbf{g}(\theta)}{\partial \theta}^T \quad (4)$$

where $\partial \mathbf{g}(\theta)/\partial \theta$ is the short hand for $r \times p$ Jacobian matrix

$$\frac{\partial \mathbf{g}(\theta)}{\partial \theta} = \begin{bmatrix} \frac{\partial g_1(\theta)}{\partial \theta_1} & \frac{\partial g_1(\theta)}{\partial \theta_2} & \dots & \frac{\partial g_1(\theta)}{\partial \theta_p} \\ \frac{\partial g_2(\theta)}{\partial \theta_1} & \frac{\partial g_2(\theta)}{\partial \theta_2} & \dots & \frac{\partial g_2(\theta)}{\partial \theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_r(\theta)}{\partial \theta_1} & \frac{\partial g_r(\theta)}{\partial \theta_2} & \dots & \frac{\partial g_r(\theta)}{\partial \theta_p} \end{bmatrix} \quad (5)$$

The diagonal elements of the covariance matrix \mathbf{C}_α correspond to the variance of each transformed parameter α .

3. UNCERTAINTIES IN A FULL-CHIP LITHOGRAPHY MODEL

Typically a full-chip lithography model is calibrated by minimizing the least squares error between simulated CDs and measured CDs. The total number of data points N could range from 100 to 10000. If an empirical model has p parameters denoted by a vector θ , the n -th test pattern would have a simulation CD $s[n; \theta]$, and its difference with measured CD $m[n]$ is the error $e[n]$. The following formula describes this relationship

$$m[n] = s[n; \theta] + e[n], \quad n = 0, 1, \dots, N-1 \quad (6)$$

The objective function to be minimized is

$$F(\theta) = \sum_{n=0}^{N-1} (e[n])^2 = \sum_{n=0}^{N-1} (m[n] - s[n; \theta])^2$$

There are many well-researched optimization algorithms readily available for us to choose from^[5]. We will not go into detail about how we exactly minimize $F(\theta)$ here.

3.1 Uncertainty of model parameters

We consider the entire calibration process of finding the true model parameters from measurements as a problem of computing a p -dimensional estimator θ . Furthermore, we regard the error term in Eq.6 as white Gaussian noise with variance σ^2 , assuming that whatever systematic errors of CD-SEM have been adequately absorbed in $s[n; \theta]$, and that all aspects of lithographic process have been adequately modeled. In estimation theoretical terms, our model is assumed to be sufficient. Now we are ready to apply Eq.1 and Eq.3 to our calibrated model

$$\text{var}(\theta_i) \geq \left[\frac{1}{\sigma^2} (\mathbf{J}^T \cdot \mathbf{J})^{-1} \right]_{ii} \quad (7)$$

where \mathbf{J} is the $N \times p$ Jacobian matrix of simulated CDs w.r.t. parameters

$$\mathbf{J} = \begin{bmatrix} \frac{\partial s[0; \boldsymbol{\theta}]}{\partial \theta_1} & \frac{\partial s[0; \boldsymbol{\theta}]}{\partial \theta_2} & \dots & \frac{\partial s[0; \boldsymbol{\theta}]}{\partial \theta_p} \\ \frac{\partial s[1; \boldsymbol{\theta}]}{\partial \theta_1} & \frac{\partial s[1; \boldsymbol{\theta}]}{\partial \theta_2} & \dots & \frac{\partial s[1; \boldsymbol{\theta}]}{\partial \theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial s[N-1; \boldsymbol{\theta}]}{\partial \theta_1} & \frac{\partial s[N-1; \boldsymbol{\theta}]}{\partial \theta_2} & \dots & \frac{\partial s[N-1; \boldsymbol{\theta}]}{\partial \theta_p} \end{bmatrix} \quad (8)$$

The inequality Eq.7 tells us that the uncertainty of any calibrated model parameter is at least that amount.

3.2 Reciprocity between simulations and parameters

In order to estimate the uncertainty of simulation results, we regard the simulated CDs at each test pattern site as a new set of parameters. In the realm of algebraic equations, if the number of unknowns remains unchanged, then a transformation of parameters would not materially affect the solution. In the more complicated situation of a lithographic process model, the forward operation is mapping from parameters to simulations, and the inverse operation is determining parameters from measurements through calibration.

In the extreme, the simulation results can be tabulated for all possible pattern shapes, which have a finite, albeit large, number of variations if we restrict our attention to a finite bandwidth and a finite domain size in an actual computer-implementation. Then this huge table of simulation results is a valid and complete empirical model, and it is exactly equivalent to the original parameterized model.

3.3 Uncertainty of simulated CDs

Now that we have justified the reciprocity between simulations and parameters, we identify $s[n; \boldsymbol{\theta}]$ of Eq.6 with $\mathbf{g}(\boldsymbol{\theta})$ of Eq.4, which signifies the uncertainty of simulation at each test pattern site

$$\mathbf{C}_s \geq \frac{\partial s(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{I}^{-1}(\boldsymbol{\theta}) \frac{\partial s(\boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}} = \mathbf{J} \left(\frac{1}{\sigma^2} \mathbf{J}^T \mathbf{J} \right)^{-1} \mathbf{J}^T = \sigma^2 \mathbf{J} (\mathbf{J}^T \mathbf{J})^{-1} \mathbf{J}^T \quad (9)$$

The matrix \mathbf{C}_s on the left side of Eq.9 is the $N \times N$ covariance matrix, whose each diagonal element corresponds to the prediction error at a test pattern site. The variance σ^2 on the right side of Eq.9 represents the true variance of noise in the measured data. Of course we don't know the true value of σ^2 , but we have a surrogate in the model fitting error. The usual formula relating true variance to fitting error is

$$\sigma^2 \approx \frac{SSE}{N-p} = \frac{N}{N-p} (rms)^2 \quad (10)$$

where SSE is the sum-squared-error between simulation and measurement, N is the number of data points, p the number of parameters, and rms is the usual fitting error people most often quote when talking about model accuracy.

The formula in Eq.9 can be further simplified, assuming we have performed Singular Value Decomposition (SVD) of $\mathbf{J} = \mathbf{U} \boldsymbol{\Sigma} \mathbf{V}^T$

$$C_s \geq \sigma^2 U \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \vdots & & \vdots \\ 0 & 0 & 1 & \vdots & & \vdots \\ 0 & \cdots & \cdots & 0 & & \vdots \\ \vdots & & & & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \end{bmatrix} U^T \quad (11)$$

where the $N \times N$ diagonal matrix in the middle of two left-singular-vector unitary matrix U has p 1's and $(N-p)$ 0's; in other words, its upper left corner is a $p \times p$ identity matrix, and the rest of the elements are zeros.

3.4 Total variability of a model

Combining Eq.10 and Eq.11 leads to a remarkably simple yet powerful metric for the simulations in aggregate. We now define the trace of the right hand side of Eq.11 to be \mathcal{E} Model Effectiveness Index (**MEI**), which is the lower bound of total variability of a model:

$$\mathcal{E} \equiv \sigma^2 \text{Tr} \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \vdots & & \vdots \\ 0 & 0 & 1 & \vdots & & \vdots \\ 0 & \cdots & \cdots & 0 & & \vdots \\ \vdots & & & & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \end{bmatrix} = \sigma^2 p = \frac{Np}{N-p} (rms)^2 \quad (12)$$

This figure of merit \mathcal{E} characterizes modeling uncertainty by balancing the fitting error (rms) with the number of parameters used (p), and the number of measurements taken (N). We can also use the square root of \mathcal{E} as another figure of merit, which has the same dimension as rms , Model Effectiveness Standard Index (**MESI**):

$$MESI \equiv \sqrt{\mathcal{E}} = \sqrt{\frac{Np}{N-p}} \cdot rms \approx \sqrt{p} \cdot rms \quad (13)$$

The right hand side of this definition holds when $N \gg p$. The fitting accuracy is now married to the formulation efficiency in Eq.12 and Eq.13, and we have arrived at the total effectiveness of the model.

3.5 Identification of measurement outliers

We can use Eq.9 or 11 to compute the lower bound of expected simulation variance for each individual calibration test pattern. If we find that the actual measurement differs from the simulation by a certain factor of the square root of variance, we can be confident that that particular measurement point is an outlier and thus better to be eliminated from consideration for model calibration purpose.

Some people might object that before you have confidence in possessing a final, reliable model, how you can be confident in using any model to qualify measured data points. Alas, we are facing the ancient dilemma of chicken or the egg: do we use our hypotheses to understand our observations, or do we use our observations to form our hypotheses. Our answer is that the best we can do without access to a separate set of independent measurements is to achieve consistency between hypotheses and observations, by iteratively and alternately trusting model and data. Thus calibrating an empirical model is not an entirely empiricist act, but rather an interaction between selecting model parameters and culling measurement data.

4. EXPERIMENTAL

Using a set of about 300 real data points of wafer CD measurements, we calibrated several models to demonstrate our approach. The calibration procedure used a least squares minimization algorithm. For each model we computed variance of parameters and the corresponding distribution of residual errors.

4.1 Uncertainty of the model parameters

First we consider a simplest model that contains only two parameters, i.e. exposure dose and focus offset. Variance of each parameter was computed using Eq.7. Low variance of the threshold indicate that it is the most critical parameter which determines overall model behavior. Second model has additional parameters that capture differences introduced by mask manufacturing. Including mask bias significantly changes the dose value and increases the variance. Overall comparison with original model gives two important conclusions, introduction of extra parameters significantly alters original values and new parameters give higher variance to preexisting parameters.

<i>Model 1 parameter</i>	<i>Dose</i>	<i>Focus</i>	
<i>Value</i>	0.16	141.7	
<i>Variance</i>	0.01	17.4	
<i>Model 2 parameter</i>	<i>Dose</i>	<i>Focus</i>	<i>Bias</i>
<i>Value</i>	0.08	133.5	-6.1
<i>Variance</i>	0.03	30.1	5.6

Table 1 Model parameters and their variances

4.2 Uncertainty of model predictions

All parameters that correspond to actual physical effects could be extracted in some independent measurement and will certainly give a more accurate model prediction. As it was demonstrated above there are only a few parameters that determine the model behavior. If the data set does not accurately capture any of the effects that correspond to actual physical behavior then it may be possible to build several drastically different models using the same set of measurements. A simple model with a mask bias included will significantly alter the estimated value of dose and change model behavior. In such a case a real estimate of the bias or a dose would be the critical information to construct the model. In the absence of such measurements, additional information could be averaged variance of model parameters extracted using Fisher Information Matrix. The goal is to find a model with smallest Model Effectiveness Index \mathcal{E} . This would also mean a smaller uncertainty for each point in the input data set as shown in section 3.3 and shorter error bar for extrapolated points.

4.3 Distribution of residual error

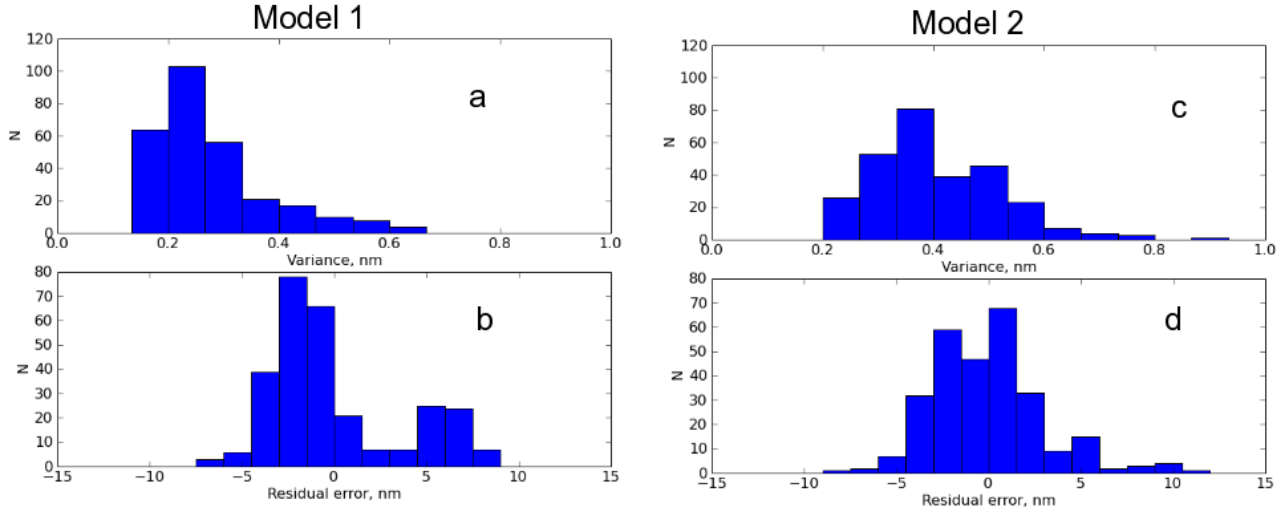


Figure 1 Data point predicted variability and residual error distribution. a - variability for model1, b - residual error for model1, c - variability for model2, d - residual error for model2

Given a set of data with minimal information it is important to recognize that some of the point may have unusually high uncertainty. Unless there is a certain amount of confidence about each point in the data set, a stable model should recognize such points. Although additional parameters may give a better overall fit to the data, that shouldn't always be the primary choice for the model. Variability (a, c) and residual error (b, d) distribution of data points was calculated using Eq. 9 for the two models discussed in 4.1. First model uses two parameters, and the second model uses three giving a slightly lower *rms* error. The tail in the residual error and variance distribution for the first case would be considered as bad points and likely contributes to higher overall *rms*. The second model is attempting to fit those points, resulting in a more symmetric distribution. Distribution of residual error of data points is a good indicator of the model sufficiency.

5. CONCLUSION AND DISCUSSION

The importance of a full-chip model makes its assessment a high-stake exercise. We attempt to define some quantitative methods to help evaluating the quality of a calibrated model. In particular, the trace of simulation covariance lower bound matrix defines a new figure of merit for modeling, the Model Effectiveness Index $\bar{\mathcal{E}}$, or Model Effectiveness Standard Index (*MESI*): $s = \sqrt{p(rms)}$. The lower the *MESI*, the better the model. The metrics are rooted in the estimation theory of statistics, and adapted to our modeling practice. We have used a couple of model-building results to demonstrate the methods.

This paper has focused on the diagonal elements of various matrices. The off-diagonal elements of Fisher information matrix, the parameter covariance matrix, and the simulation covariance matrix carry important information. This will be further studied and reported in the future.

Our analysis of simulation uncertainty and the reciprocal relationship between simulated CDs and model parameters has implication for calibration pattern design and wafer measurement plan. Fewer data points than customarily used are sufficient to building a good enough empirical model^[4]. We will explore this important area in a future study.

REFERENCES

- [1] Narender Rana, Chas Archie, Wei Lu, and Bill Banke, "The measurement uncertainty challenge of advanced patterning development", Proc. SPIE, Vol.7272, 727203 (2009).
- [2] Samit Basu, Yoram Bresler, "The Stability of Nonlinear Least Squares Problems and the Cramér-Rao Bound", IEEE Transactions on Signal Processing, Vol.48, No.12, 3426-3436 (2000).
- [3] Steven M. Kay, [Fundamentals of Statistical Signal Processing: Estimation Theory], Prentice Hall PTR (1993).
- [4] Ralph E. Schlieff, "Effect of data selection and noise on goodness of OPC model fit", Proc. SPIE, Vol.5754, 1147-1158 (2005).
- [5] R. Fletcher, [Practical Methods of Optimization], 2nd ed., John Wiley & Sons (1987).