

Improving the deployment of inspection tools; tutorial on inspection capacity and sample planning

Dadi Gudmundsson, Prof. J George Shanthikumar
IEOR dept., University of California, Berkeley
Berkeley, CA, USA ,
dadi@ieor.berkeley.edu

Abstract – To help fabs improve their use of inspection tools a practical guide to inspection capacity and sample planning is presented. A comprehensive approach involving data collection, parameter estimation, and model based sample planning is covered. The objective is to enable yield engineers to develop basic inspection planning toolboxes in a spreadsheet and/or more effectively manage inspection planning projects. For parameter estimation, the use and performance of a Hidden Markov Model and the E-M algorithm is introduced.

INTRODUCTION

The role and importance of inspection and metrology (insp/metr) in semiconductor manufacturing continues to increase. For example, fabs spent about 5% of capital expenditure on process control in the early 1990s, by 2000 it was around 10% [1]. This percentage increase, and the ever increasing capex itself, gives greater importance to the science of calculating the insp/metr capacity fabs need and/or how to deploy insp/metr most effectively. Models have been developed to do this [2, 3], but their complexity (relative to general process tool capacity planning, for example) has limited their use. This poster presents a tutorial on basic inspection capacity planning, often referred to as sample planning. The objective is to enable yield engineers to develop basic inspection planning toolboxes in a spreadsheet and/or more effectively manage inspection planning projects.

Emphasis is given to data collection and analysis as those are often cited as roadblocks. The use and performance of a Hidden Markov Model (HMM) and E-M algorithm to parameter estimation from inspection data is presented. Inspection capacity approximation formulas are then presented. These can be used to improve the deployment of current inspection capacity. The paper concludes with a recommendation on the use of the approximation model presented and for what circumstances to consider a more accurate sample planning model.

INSPECTION CAPACITY PLANNING

The model considered here is based on the excursion paradigm. This assumes that a fab/process is in the full production phase and the goal is to maintain a certain yield level. Recurring excursions reduce the yield below target level until they are detected and fixed. The value of

inspection (or metrology) tools is then quantified based on how they can catch these excursions and decrease the amount of yield lost to excursions.

The objective of inspection capacity planning is to identify the inspection capacity that balances inspection costs with the losses due to excursions. That point represents the amount of sampling that is economical. To do this the following main questions need to be answered: 1) inspection points, i.e. after which process steps to send lots to inspection, 2) inspection tool type(s) to use, 3) whether to inspect product or test wafers, 4) what inspection sensitivity to use, 5) which product(s) to inspect, 6) area per wafer to inspect, 7) how many wafers to sample from a lot, and 8) percent of lots to sample. The approximations presented in this paper focus on items 1, 7 and 8. For brevity the output of the model is the expected number of lots at risk to excursions. To balance inspection cost against yield losses, the number of lots at risk can be converted to financial losses with a yield and cost model.

The following sections cover the three main tasks in inspection capacity planning: data collection, parameter estimation, and capacity planning.

DATA COLLECTION

Data is needed from each inspection point. If data does not exist for an inspection point of interest, then the best representative data from another step, process, or fab should be used. The main question is how much data is needed from each inspection point. Excursion related parameters require much more data than is needed to estimate, for example, mean in-control (inc) defect count. Consider the number of lots needed to estimate the excursion probability. The lower the probability, the more data is needed to capture several excursion instances.

A standard model to estimate the sample size for a proportion estimate can be used to approximate the amount of data needed. In Figure 1, the months worth of data needed to estimate different excursion probabilities is shown for a 20,000 wspm fab. Figure 1, assumes an estimate error of +/- 20%. Using this model one assigns confidence c to having the estimate within +/- 20% of the actual value of interest. A general rule of thumb of collecting 2.5 months worth of inspection data would cover many inspection points with acceptable accuracy. Analysis of that data then shows from which inspection points more data may be needed.

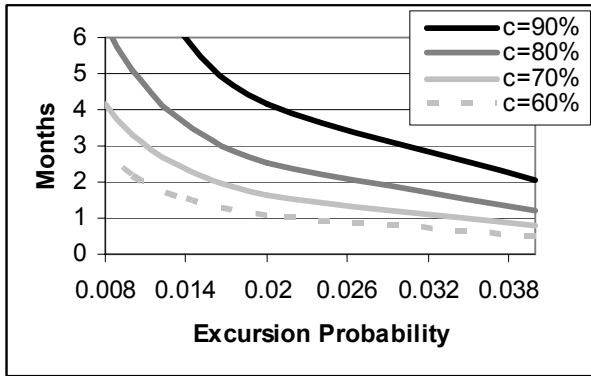


Figure 1. Approximate number of months of inspection data needed to estimate excursion probability for one inspection point. Four confidence levels shown for a 20,000 wafer starts per month fab.

Please note that: 1) the requirements can be scaled for different wspm, e.g. a 10,000 wspm fab/process would double the months needed, 2) 25%-100% lot sampling is assumed for the inspection points during data collection. Lower lot sampling requires an amended guide.

PARAMETER ESTIMATION

Data pre-processing

It is very feasible to do the data pre-processing manually in a spreadsheet. Scripts can then be developed to automate some of the operations.

Defect counts. The approximation method in this paper will split defects into killer and non-killer defects. The defect count to be associated with each wafer should be cumulative count of all killer defects according to the classification information. If no classification information is available then total defect count will have to be used. The killer defect count for each wafer will be referred to as y_{ij} , the count for wafer j in lot t .

Outlier removal. Due to malfunction or misguided operation, inspection tools will occasionally output very high defect counts. These counts are so much higher than counts associated with excursions that they can easily be identified. Lots exhibiting these type of counts need to be removed from the dataset. Figure 2 shows a histogram from a typical dataset. In this dataset, outlier datapoints would have readings in the thousands and hence be easily distinguishable from the main set of data.

Lot averages. In cases where there are more than one wafer reading per lot it is assumed that excursion determination is based on lot averages. For those datasets, the lot average needs to be calculated for each lot. The lot average will be referred to as $y_t = \sum_{i=1}^k y_{ti} / k$ where $k = \#$ of wafers sampled per lot.

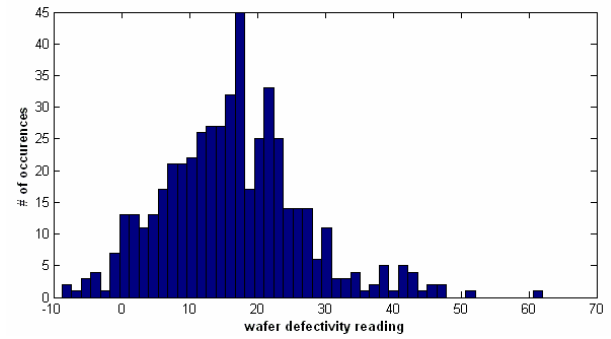


Figure 2. Histogram of defect inspection data with inc and out-of-control distributions partially overlapping each other. Some of the lots belonging to excursions can be seen to the right of the main distribution.

Missing data. In cases where the data comes from an inspection point that currently is not looking at 100% of lots, some lot readings are inevitably missing. For the sake of simplicity in using the HMM and E-M, dummy lots will be inserted for missing lots instead of doing a post analysis correction of the parameters. If the lot sampling at the inspection point providing the data was 50%, then each lot in the dataset needs to be copied once. For ordering purposes the copied lot should be given a timestamp that is midway between the lot it is copied from and the lot in front of it. If the lot sampling was 33% then two lots need to be copied from each lot in the dataset, etc.

The full dataset will now have in each row t : lot ID, lot timestamp, $y_{t1}, y_{t2}, \dots, y_{tk}$ and y_t , where k is the number of wafers sampled per lot. Define T as the number of lots and define \mathbf{y} to be the subset, $\mathbf{y} = (y_1, y_2, \dots, y_T)$. Note that even if the out-of-control (ooc) distribution is well separated from the inc distribution, a manual separation is not recommended. That would require a manual processing of all data points to verify which lots should be clustered to belong to the same excursion and/or to find p_{exc} which is not defined as: # of excursion lots/ T .

HMM formulation and E-M algorithm

Background. During the processing of lot t the process step is assumed to be in state q_t which is either inc ($q_t=0$) or ooc ($q_t=1$). That state is, however, "hidden" from the inspection tool. The only indicator of the state is the noisy lot average y_t . Parallel to the vector \mathbf{y} there is the hidden vector $\mathbf{q}=(q_0, q_1, \dots, q_T)$ that needs to be uncovered, i.e. a 1 or a 0 needs to be assigned to all elements of \mathbf{q} . When that is complete each parallel value in \mathbf{y} is known to be either an inc or out-of control value which is necessary for parameter estimation. Experience has shown that fab defectivity (and metrology) data in the full production phase has excursion signatures [3, 5]. The process steps exhibit a pattern of staying inc until they shift to ooc requiring an intervention to fix the excursion. After the intervention the process step is back inc and starts the excursion cycle over again. The y_t values exhibit this

pattern, i.e. y_{t+1} is not independent of y_t . If y_t is created while the process is out of control then y_{t+1} has a greater probability of coming from the out of control distribution. Therefore, values adjacent to y_t can and should be used to identify whether y_t (and its corresponding q_t) is in or out of control. The HMM formulation does this.

HMM and E-M algorithm. Parameter estimation with HMM and E-M are an instance of the known statistical tool of maximum likelihood (ML). In ML a likelihood function is defined. This function measures the likelihood that a set of given parameters will result in the data acquired. The parameters that maximize the likelihood function are the ones most likely to produce the data. A HMM arises in our context as the formalism for a likelihood function. The resulting function could be passed to any black-box numerical optimization routine, but this particular function has been shown to have a structure that the E-M algorithm can optimize efficiently. Generally speaking the E-M algorithm is a coordinate ascent algorithm. Mathematical details on HMMs and the E-M algorithm can be found, for example, in [4, 5].

For each inspection point the following parameters are needed from the data: p_{exc} = excursion probability, μ_{inc} = mean inc defectivity, μ_{ooc} = mean ooc defectivity, and σ^2 = defectivity variation. Let $\theta = (p_{exc}, \mu_{inc}, \mu_{ooc}, \sigma^2)$. In our case we want to find the θ that maximizes the likelihood of observing the data y .

HMM/E-M implementation and performance. Many statistics programs contain a HMM/E-M feature. For example, HMM/E-M functionality comes with the statistical toolbox in MATLAB. In this implementation we are interested in separating two distributions that we assume are normal/Gaussian. Therefore, we use, in HMM terminology, a ‘‘HMM with a mixtures of Gaussians output’’. To verify that the algorithm performs adequately a large experiment with simulated data is performed. The performance of the algorithm is then compared to the known parameters used to create the dataset and an unconditional mixture model (UMM). UMM is an analysis method that does not take the dependency of y_t and y_{t+1} into account. Note that UMM does not provide a p_{exc} estimate and is, therefore, not fully comparable to HMM.

Each simulated dataset has 2000 lots which corresponds, for example, to 2.5 months of data for the $c=70\%$ curve in Figure 1. A total of 1500 datasets were created = fifteen replications of 100 different parameter combinations. As it turns out, $\delta = \mu_{ooc} - \mu_{inc}$ is what really matters when it comes to calculating the probability of catching an excursion. Therefore, a range of five δ values, is used along with four p_{exc} values, and five σ values to get $5 \times 4 \times 5 = 100$ different parameter combinations. μ_{exc} = mean excursion length was kept at 5 lots. The absolute values may not apply to all process steps, but the ratios of δ and σ have been observed to be common across many process

steps. The values used are: $p_{exc} = (0.01, 0.02, 0.03, 0.04)$, $\delta = (15, 22.5, 30, 37.5, 45)$, and $\sigma = (3, 6, 9, 12, 15)$. In the comparison, one point is assigned to each method that has both δ and σ estimates within an error tolerance for all 15 replications of a parameter combination. This makes 100 points a maximum score, see Table 1. The HMM methodology performs considerably better and overall scores are good considering the strict scoring scheme.

Method	Error tolerance			
	5%	10%	15%	20%
UMM	31	45	47	48
HMM	43	63	68	73

Table 1. Scores for the two methods. A method gets a point for each parameter combination that has the δ and σ estimates within an error tolerance.

Since UMM does not provide an estimate of p_{exc} , HMM is only compared with the known values generating the data, see Figure 3. For $p_{exc} > 0.01$ HMM delivers acceptable accuracy for delta/sigma ratios of 2-4 where most datasets reside. However, as expected from the guide in Figure 1, there is not enough data for the HMM to estimate for $p_{exc} = 0.01$ in the critical 2-4 delta/sigma ratio range.

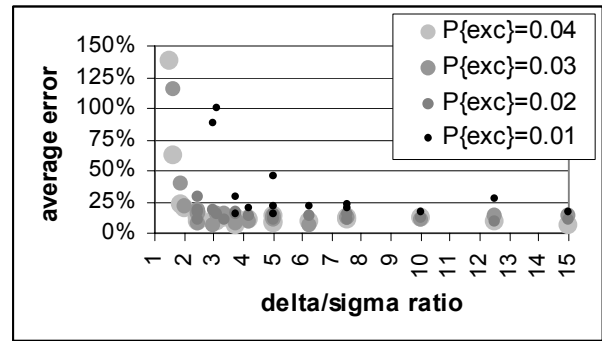


Figure 3. Ability of HMM to estimate excursion probability from 2.5 months worth of data in a 20,000 wspm fab.

The strategy to derive from this result is as follows: 1) collect data using the rule of collecting at least 2.5 months of data per inspection point in a 20,000 wspm fab. 2) perform analysis and notice which inspection points give $p_{exc} > 0.1$ or < 0.02 . These values are indicators that more data is needed. If more data is not available, then this can be compensated with sensitivity analysis in subsequent analysis.

Assuming the HMM assigns state 0 to the inc data, the parameter estimation can be completed as follows for each inspection point: $p_{exc} = 1 - p_{00}$, $\mu_{inc} = \mu_0$ = mean of all y_t labeled as inc, $\mu_{ooc} = \mu_1$ = mean of all y_t labeled as ooc, and $\sigma^2 = k \sigma_{00}^2$. Note that p_{00} , μ_0 , μ_1 and σ_{00}^2 are standard outputs from a HMM with Gaussian outputs.

CAPACITY PLANNING

Solution to the inspection capacity planning problem can be approximated with the following model.

$$\alpha = 1 - \Phi\left(x - \mu_{inc} / \sqrt{\sigma^2/k}\right) \quad \beta = \Phi\left(x - \mu_{ooc} / \sqrt{\sigma^2/k}\right)$$

α = false alarm probability = P{excursion signal | no excursion in progress}, β = probability of missing excursion signal = P{no excursion signal | excursion in progress}, x = control limit on control chart, and $\Phi(\bullet)$ is the cumulative standard normal distribution.

$$E[N_{inc}] = \frac{1 - p_{exc}}{p_{exc}} \quad E[N_s] = \frac{(1 - p_{exc})^h}{1 - (1 - p_{exc})^h}$$

$E[N_{inc}]$ = expected # of lots produced before excursion starts, $E[N_s]$ = expected # of lots sampled before process shifts, h = sampling interval (integer), $(1/h)*100$ = percent of lots sampled.

$$E[N_{det}] = 1/1 - \beta \quad E[N_c] = h(E[N_s] + E[N_{det}])$$

$E[N_{det}]$ = expected # of samples needed to detect excursion, $E[N_c]$ = expected # of lots in a inc-ooc cycle.

$$E[N_{LR}] = E[N_c] - E[N_{inc}] \quad \omega = \alpha E[N_s] r / E[N_c]$$

$E[N_{LR}]$ = expected # of lots at risk, ω = # of false alarms/week, r = lots per week produced.

$$\gamma = rkt_{insp} / h 420$$

γ = inspection capacity needed in hours/day, t_{insp} = inspection time in min/wafer.

Let us refer to a choice of h and k as a sample plan. To get a fair comparison between different sample plans it is necessary to find x^* = the optimal control limit associated with each pair of h and k . One can also optimize h , k , and x simultaneously. Optimization is done by minimizing $E[N_{LR}]$ while constraining ω and γ to be less than a user specified value. With the appropriate settings, the "solver" in Microsoft Excel optimizes this formulation in a fraction of a second. The above can be extended to study multiple inspection points in parallel. In those cases one should minimize/use $E[N_{LR}]/E[N_c]$ to get a relative comparison of the lots at risk for different inspection points. Also, ω and γ should be summed up over all the inspection points. If desired, it is not hard to add a basic yield and cost model to the above formulation. That would allow balancing of the cost of inspection and the costs due to excursion yield losses.

USE OF THE APPROXIMATION

For mature and relatively low priced products, the model herein can help fabs quickly assess and improve deployment of current inspection capacity. When greater uncertainty or monetary value is involved, more accurate

models are recommended. This includes planning for less mature products, medium-to-high priced products, and/or inspection capacity purchase decisions. More accurate models capture the higher order effects on α and β from, for example: 1) difference in lot-to-lot and wafer-to-wafer variance, 2) difference in inc and ooc variance, 3) excursion signal propagation (non-independent inspection points), 4) multiple defect/excursion types, and 5) inspection tool capture rate(s). Then there are also higher order dynamics that affect other aspects than α and β . Some major ones are listed in [2], for example: ability of preventive maintenance and probe to fix/catch excursions and the impact of sampling on fab cycle-time and capacity of process tools.

CONCLUSION

A comprehensive, yet manageable, approach towards improvement of inspection capacity deployment has been presented. The guidelines ranging from data collection to calculation of lots at risk can be implemented directly while simultaneously serving as a foundation for understanding of more advanced modeling. Other new results include the good performance of a Hidden Markov Model and the E-M algorithm for parameter estimation

REFERENCES

- [1] KLA-Tencor, "Summary Annual Report," 2004.
- [2] Gudmundsson, D., et al., "Integrated process and inspection/metrology capacity planning," In proceedings of ISSM, Published by IEEE, Piscataway, NJ, USA, 2005.
- [3] Williams, R., et al., "Optimized sample planning for wafer defect inspection," In proceedings of ISSM, Published by IEEE, Piscataway, NJ, USA, 1999.
- [4] Jordan, M. I., *Learning in Graphical Models*, Kluwer Academic Publishers, Boston, MA. 1998.
- [5] Lee, S. J., *Inspection tool selection and capacity allocation for In line process control*. Ph.D. Thesis. University of California, Berkeley, 96 pages, 1999.

AUTHOR BIOGRAPHY

Dadi Gudmundsson, received the B.S. degree in Industrial Engineering from the Univ. of Arkansas, Fayetteville in 1995. In 1997 he received the M.S. degree in Industrial Engineering and Operations Research from the Univ. of California, Berkeley. This fall he will receive a Ph.D. from Berkeley in the same field. From 1998 to 2004 he worked at KLA-Tencor researching and implementing models for inspection and metrology capacity planning.