Gate ploy-Silicon Critical Dimension DOE

2.830 Control of Manufacturing Processes

Final Project

Ying Pan
Tom Neal
Steve Vasovski
Introduction

The semiconductor manufacturing process is one of the most complex processes in the manufacturing realm. There are numerous processes that happen in serial and parallel in order to go from growing a silicon crystal ingot to an individually packages integrated circuits (IC’s). The following is a breakdown of the various processes involved in semiconductor manufacturing (M.K. Carol Lee and J. Julian Elliot, Semiconductor Manufacturing):

A. Blank Wafer Production
   1. Silicon Crystal Growth
   2. Wafer Manufacturing

B. Semiconductor Fabrication
   1. Overview
   2. Detailed Process Descriptions
      a. Oxidation
      b. Coating Application
      c. Photoresist Exposure
      d. Photoresist Development
      e. Wet Etching
      f. Dry Etching
      g. Photoresist Stripping
      h. Doping - Diffusion
      i. Doping - Ion Implantation
      j. Layering - Epitaxial Growth
      k. Layering - Sputtering
      l. Layering - Chemical Vapor Deposition
      m. Chemical Mechanical Polishing
      n. Solvent Stations
      o. Wet Chemical Stations
      p. Tool / Fab Wipe Cleaning

C. Assembly and Packaging

In addition to the complexity of the process itself, the fact that the IC is so highly integrated makes semiconductor manufacturing that much more complex. For example, transistors, the building blocks of IC’s, are connected together to make functional modules (logic, memory, etc.), which are connected together to make a chip. (As of 2003, Intel’s Itanium® 2 processor was made up of over 400 million transistors!) There are hundreds, sometimes thousands of chips (dies) on each wafer, and a number of wafers (usually less than 25) in a lot. The individual wafers are grouped together in lots and travel through the production process in their respective lot.
As one can see, there are plenty of opportunities, for variation to enter the process and decrease yield. Decreased yield means decreased profitability, therefore yield is very important in semiconductor manufacturing.

This report summarizes and analyses three different papers that all impact yield in some way.

1. Part I - “Statistical Design of Experiments and Analysis on Gate Poly-Silicon Critical Dimension”

2. Part II - “Statistical Bin Analysis on Wafer Probe” & “Challenges for use of statistical software tools in the semiconductor industry”

3. Part III - “Monitoring Defects in IC Fabrication Using a Hotelling T² Control Chart”
Part I: Statistical Design of Experiments and Analysis on Gate ploy-Silicon Critical Dimension

Introduction

The semiconductor manufacturing process is one of the most complex processes in the manufacturing realm. There are numerous processes that happen in serial and parallel in order to go from growing a silicon crystal ingot to an individually packages integrated circuits (IC’s). The following is a breakdown of the various processes involved in semiconductor manufacturing (M.K. Carol Lee and J. Julian Elliot, Semiconductor Manufacturing):

A. Blank Wafer Production
   1. Silicon Crystal Growth
   2. Wafer Manufacturing

B. Semiconductor Fabrication
   1. Overview
   2. Detailed Process Descriptions
      a. Oxidation
      b. Coating Application
      c. Photoresist Exposure
      d. Photoresist Development
      e. Wet Etching
      f. Dry Etching
      g. Photoresist Stripping
      h. Doping - Diffusion
      i. Doping - Ion Implantation
      j. Layering - Epitaxial Growth
      k. Layering - Sputtering
      l. Layering - Chemical Vapor Deposition
      m. Chemical Mechanical Polishing
      n. Solvent Stations
      o. Wet Chemical Stations
      p. Tool / Fab Wipe Cleaning

C. Assembly and Packaging

In addition to the complexity of the process itself, the fact that the IC is so highly integrated makes semiconductor manufacturing that much more complex. For example, transistors, the building blocks of IC’s, are connected together to make functional modules (logic, memory, etc.), which are connected together to make a chip. (As of 2003, Intel’s Itanium® 2 processor was made up of over 400 million transistors!) There are hundreds, sometimes thousands of chips (dies) on each wafer, and a number of wafers (usually less than 25) in a lot. The individual wafers are grouped together in lots and travel through the production process in their respective lot.
Statistical Design of Experiments and Analysis on Gate Poly-silicon Critical Dimension

Introduction

Gate poly-silicon makes up the gate of the transistor which essentially turns the transistor on and off. The length of the gate poly-silicon is a crucial parameter in determining the operating characteristics of a transistor and becomes even more important as:

1. Chip size decreases
2. Operating speed increases
3. Electric power consumption decreases

As such, it is essential to monitor this parameter in a semiconductor wafer fabrication process (i.e. semiconductor fab) to make sure that it does not deviate significantly from its nominal length. The actual measurement is termed a critical dimension (CD).

Typically, for a developed process that is in stable production, statistical process control (SPC) techniques like run charts or Hotelling $T^2$ charts are used to make sure that the process is in control. When the process is out of control however, control charts and the like become less helpful in analyzing the CD variation. Bringing the process in control may require making process changes, which in turn requires precise knowledge of the CD variation components. Designed experiments, more specifically, the analysis of variance (ANOVA) can be used to effectively investigate the CD variation. However, in order to assess process variation or to even do the most elementary estimation or testing methods, a more complicated sampling scheme than simple random sampling is required.

As discussed in the Introduction, process complexity and batch manufacturing make process sampling less straightforward. Traditional sampling methods for assessing variation can be misleading. For example, each run is composed of multiple wafers, each wafer is composed of multiple dies and each die is composed of multiple functional modules (i.e. memory, logic).

A common variance structure in batch processing environments, such as semiconductor fabs, is the nested structure. Park refers to this nested variance components as hierarchical variance components (HVC). One can clearly see the hierarchy that is imposed by the manufacturing process. Measurements made on a wafer are said to be nested within the wafer and measurements taken within a particular lot are said to be nested with the lot. One can attempt to maximize yield first by determining where the variation is occurring, then by modifying the process to reduce variation.

Yield is tremendously important in semiconductor fabs, uniformity throughout all transistors in a device must be achieved because devices that do not meet CD specifications are not passed to subsequent steps in the process, thereby decreasing yield.
Park analyzes the gate poly-silicon critical dimension from a real pilot semiconductor wafer fabrication process and provides a framework for statistical design of experiments (DOE) to guide practitioners in estimating component variance as well as uniformity testing.

**Background**

The gate poly-silicon is build up in the photolithography process by the following steps:

1. application of resist
2. exposure resist to ultraviolet light through a mask
3. removal of soluble photoresist by the develop chemical
4. removal of exposed oxide (not protected by resist) by etching
5. removal of protective photoresist

Uniformity is closely related to yield, i.e. variation in uniformity directly causes die loss resulting in a decreased yield. Process parameters as well as device design are both sources of CD variation. In order to increase yield (i.e. reduce CD variation), it is essential to understand the sources and magnitude of variation. Therefore, HVC estimates are required.

**Analysis and Conclusions**

**Nested two level model**

The length of the gate ploy-silicon was measured on five positions on the wafer, T (top), L (left), C (center), R (right) and F (flat). These measurements are said to be nested under the wafer. In turn, five wafers were randomly selected from each run (or lot). The wafers are said to be nested under the runs. Finally, three runs were randomly selected for the experiment. The nested structure of the experiment is shown below in Figure 1:

![Figure 1. Sampling method for design of experiment to test variation of gate poly-silicon length CD.](image)
The distribution of gate poly-silicon length is modeled with this two level nested model:

\[ y_{ijk} = \mu + R_i + W_{j(i)} + \varepsilon_{(ij)k} \]

where

\[ R_i \sim N(0, \sigma_R^2) \text{ for } i = 1, 2, \ldots, I \]
\[ W_{j(i)} \sim N(0, \sigma_W^2) \text{ for } j = 1, 2, \ldots, J \]
\[ \varepsilon_{(ij)k} \sim N(0, \sigma_\varepsilon^2) \text{ for } k = 1, 2, \ldots, K \]

\[ I = 3, J = 5, K = 5 \]

\( R_i \) is the factor associated with the run \( i \), \( W_{j(i)} \) is the factor associated with the wafer \( j \) under run \( i \), and \( \varepsilon_{(ij)k} \) is the random error component. \( y_{ijk} \) is the observed CD measurement for \( (i,j,k) \) and \( \mu \) is the overall (constant) process mean. All terms are independent of each other; therefore the variance of a CD observation is as follows:

\[ \sigma_T^2 = \sigma_R^2 + \sigma_W^2 + \sigma_\varepsilon^2 \]

The total variance can be broken up into variance components. \( \sigma_R^2 \) is the run-to-run variance, \( \sigma_W^2 \) is the wafer-to-wafer variance and \( \sigma_\varepsilon^2 \) is the random error (or within wafer variation). The following hypothesis tests determines if these variance components are significant:

1) \( H_0: \sigma_R^2 = 0 \)
   \( H_A: \sigma_R^2 \neq 0 \)
2) \( H_0: \sigma_W^2 = 0 \)
   \( H_A: \sigma_W^2 \neq 0 \)

Table 1 shows the gate poly-silicon length measurements along with intermediate statistics, wafer variance, wafer mean and variance of the wafer means for each lot.
The analysis of variance for the above data is shown below in Table 2.

Table 2. Analysis of Variance and component variance estimates for nested two level design.

<table>
<thead>
<tr>
<th></th>
<th>Variance Source</th>
<th>dof</th>
<th>S.S.</th>
<th>M.S.</th>
<th>F</th>
<th>p-value</th>
<th>Variance Component</th>
<th>% of Total Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lot</td>
<td>2</td>
<td>6872.71</td>
<td>3436.36</td>
<td>6.82</td>
<td>0.011</td>
<td>117.30</td>
<td>49.6%</td>
</tr>
<tr>
<td>Wafer</td>
<td>12</td>
<td>6046.98</td>
<td>503.92</td>
<td>21.70</td>
<td>0.000</td>
<td>96.14</td>
<td>40.6%</td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>60</td>
<td>1393.49</td>
<td>23.22</td>
<td></td>
<td></td>
<td>23.22</td>
<td>9.8%</td>
<td></td>
</tr>
</tbody>
</table>

The variance components are significant (to $\alpha = 0.05$), with lot-to-lot variance ($\sigma_R^2 = 117.30$) accounting for 49.6% of the variation, wafer-to-wafer variance ($\sigma_W^2 = 96.14$) accounting for 40.6% of the variation and the within wafer variation ($\sigma_{\epsilon}^2 = 23.22$) accounting for the remaining 9.8% of the variation. These results indicate that both the wafer-to-wafer and lot-to-lot variations contribute significantly to the CD variation. However upon closer review of the intermediate statistics presented in Table 1, one can see two suspicious statistics.

1. The run 1 variance from wafer means for the gate poly-silicon length CD (see last column of table) is much larger than that for runs 2 and 3. The unusually high variance comes from wafer 1 which has a much larger mean CD than the rest of the wafers in the run. This result brings into question the integrity of the measurements for run 1 wafer 1.

Two regressions were run to determine the impact that wafer one has on the results. The regression in Figure 2 does not include wafer 1 while the regression
in Figure 3 does. It is clear, from Table 3 that the gate poly-silicon length CD does not depend on the wafer (p-value = 0.2102, zero is in the 95% CI of the wafer effect estimate) when wafer 1 is not included. However, when wafer 1 is included in the regression, Figure 3, the wafer effect becomes highly significant (p-value = 0.0012, zero is not even in the 99% confidence interval). In addition, in the second regression, the standard errors of both the intercept term and wafer effect are increased, see Table 4.

![Figure 2. Regression plot to determine effect of wafer (2 -5) on CD.](image1)

![Figure 3. Regression plot to determine effect of wafer (1 -5) on CD.](image2)

**Table 3. Parameter estimates for wafer (2-5) effect on CD.**

| Term     | Estimate | Std Error | t Ratio | Prob>|t| |
|----------|----------|-----------|---------|-----|
| Intercept| 202.538  | 3.79476   | 53.37   | <.0001 |
| Wafer    | 1.342    | 1.032862  | 1.30    | 0.2102 |

**Table 4. Parameter estimates for wafer (1-5) effect on CD.**

| Term     | Estimate | Std Error | t Ratio | Prob>|t| |
|----------|----------|-----------|---------|-----|
| Intercept| 234.922  | 6.051114  | 38.82   | <.0001 |
| Wafer    | -6.754   | 1.824479  | -3.70   | 0.0012 |

We cannot simply discard the wafer 1 data because we feel it is suspicious, but it is worth mentioning because it could potentially skew the results of the variance component estimates.

II. The mean CD for run 3 appears to be significantly larger than that of the previous two runs. We can test this with the following two hypothesis tests.

1) \( H_{01}: \mu_{L1} - \mu_{L2} = 0 \)
   \( H_{A1}: \mu_{L1} - \mu_{L2} \neq 0 \)

2) \( H_{02}: \mu_{L1} - \mu_{L3} = 0 \)
   \( H_{A2}: \mu_{L1} - \mu_{L3} \neq 0 \)

If in fact the mean CD for run 3 is different, we will reject the null hypothesis for test 1 and accept the null hypothesis for test 2.
A p-value = 0.2706 for test 1 indicates that we cannot reject $H_0^1$, there is no statistical difference between the mean CD for lot 1 and the mean CD for lot 2. A p-value < 0.0001 indicates that we can reject $H_0^2$, there is a statistical difference between the mean CD for lot 1 and the mean CD for lot 2.

Therefore we can conclude from I that the mean CD measurements from wafer 1 will inflate the wafer variance component. We can also conclude from II that the CD measurements from lot 3 will inflate the lot variance component. If this is the case then, the random error (or within wafer measurements) will account for a larger percentage of the variation.

We recommend that, some more tests be taken and the data be re-analyzed for the following two reasons:

1) The CD measurements on wafer 1 in run 1 are abnormally high; this was not the case on the other two runs. One must determine if this is random or systematic. If the wafer 1 CD measurements are too high for an abnormal reason, it will overstate the wafer variance component estimate.

If in fact the wafer 1 run 1 measurements are abnormally high, then a much smaller portion of the variance is due to the wafer effect. A practitioner may then try
reduce the lot-to-lot variation, which is the largest, to reduce the overall variation in process.

2) There appears to be a mean shift after run 2. If this is the case, the lot-to-lot variance component estimate may be overstated.

If there had been a shift in the mean after run 2, and the variance component due to the lot effect was overstated, a smaller portion of the variation would due to the lot effect (i.e. a larger portion would be due to the wafer effect). A practitioner may then try to reduce the variation by targeting wafer-to-wafer variation.

If both scenarios 1) and 2) are correct, then both the wafer and lot effects are overstated and the within wafer variation is understated.

The preceding analysis highlights the importance of having reliable variance component data in order to understand where a reduction in component variance will have the largest impact on total variance.
Part II: Use of statistical software tool in semiconductor industry

With the number of IC increases significantly in their product, IC customers have higher and higher expectation of the quality of the ICs provided by their suppliers. DPM (defect per million) rates are driven down quickly in recent years, IC suppliers are trying to reduce the number of field returns, at the same time, working on improving the production yield and driving the cost down. With astronomical amount of data to be processed and analyzed, statistical software is playing an increasingly important role in semiconductor industry to help companies improving quality and yield, reducing defects and cost. I will start Part II with a recent example of how statistical software tool are used in wafer/package testing.

To improve the quality of ICs going into automobiles and reduce liability issues, the Automotive Electronics Council recommended the use of Part Average Testing (PAT), which is a methodology that prescribes finding test results that fall outside six sigma from the population mean for a given wafer, lot, or group of parts being tested. Any test result outside the six sigma limit for a given device is considered an outlier and removed from the population. Parts that fail the PAT limits are not shipped to the customer, thereby improving quality and reliability.

In a PAT process, recent data from several lots are analyzed to establish static PAT limits for each test. These limits are calculated as Mean +/- 6 Sigma and are incorporated into the test program as the Upper Specification Limit (USL) and the Lower Specification Limit (LSL). However, the preferred method is to calculate dynamic PAT limits for each lot or wafer. The important distinction is that dynamic PAT limits are calculated on a wafer or lot basis, thus the limits are continuously changing based on the performance of the material for that wafer or lot, and cannot be less than the LSL or greater than the USL specified in the test program. Any values outside the dynamic PAT limits but within the LSL and USL limits are considered outliers, and are designated as failures and are binned out to a special outlier software or hardware bin. My work at a semiconductor company gave me some experience on wafer/package testing that it seemed most of the field returns are the parts tested using static limits. Testing using dynamic limits is certainly desirable if software is available.
Two different options are available for implementing PAT in production—real-time PAT and statistical post processing (SPP). Suppliers must decide which option to use for probe and final test. Real-time PAT relies on calculating dynamic PAT limits and making binning decisions in real-time as parts are tested. This process requires robust statistical software capable of taking test data and performing the necessary calculations to generate new limits, passing the new limits and binning information into the test program. Real-time processing works for both probe and final test. Statistical post processing processes statistics from device test and makes binning decisions after a lot has been completed. It can only be used for wafer probe since the test and binning results must be tied to a specific device so that it can be re-binned. At final test, there is no tracking mechanism or serialization once parts have been packaged.

To meet the needs of semiconductor companies, many start-ups are making statistical software to help to address some of the issues. Software company such as Pintail Technologies developed related statistical tools that makes dynamic PAT limits and real-time PAT testing much easier to implement.

However, there are still significant challenges for the use of statistical software tools in semiconductor industry. The rest of Part II will provide some background information on the use statistical tools in other engineering functions within a semiconductor company, and then discuss some the difficulties and challenges.

Typically, within a semiconductor company, product engineers are responsible for the characterization of new products. Characterization involves running of ‘skewed corner lots’ on a new design to check the performance of the product at the extremes of the data sheet specifications. Ideally this is based on a well-defined methodology involving design of experiments. Characterization data is then used to model the variability of the process.
across the specification range at different test conditions and establish guard band limits for product testing.

Process development engineers run designed experiments to characterize the capabilities of new tools. They also work to characterize the stability of new processes through control charting and process capability studies prior to the handoff to manufacturing.

Manufacturing engineers are responsible for ramping volume on new technologies and tools received from the development groups. Their tools of the trade are control charts, ongoing process capability studies, trouble-shooting guides, and designed experiments to test process improvements.

Yield engineers monitor probe yields, analyze low yielding lots, and correlate yield data to process histories to understand the important factors influencing yield. In addition to exploratory data analysis through use of SAS JMP, Statistica, RS/1 or commercial yield management software, some of the methods being used include:

- ANOVA or Kruskal-Wallis analyses to look for tool-to-tool performance differences
- Regressions to monitor the relationships between electrical parameters
- Commonality studies to look for what is in common across a group of bad lots
- ANOVA or Kruskal-Wallis differences analyses to contrast high/low yielding lots
- Routines to correlate SPC violations to prior processing histories

Great difficulties remain in the use of statistical software tool with the semiconductor industry. Companies are facing the dilemma that on one hand the commercial statistical software tools are too general to address semiconductor-specific data access and analysis tasks, on the other hand the commercial semiconductor software applications, such as yield management and SPC packages are too specific to cover the breadth demanded by the spectrum of tasks.

Due to the diversity of data and complexity of its interrelationships, the ease and flexibility of data access is still a major problem for many of the engineering groups. Data is often stored in different systems, parameters follow various naming conventions, and there are often no easy methods for merging the data together. My past work experience with data analysis was typical in a semiconductor company, that engineers spent significant time on data collection, cleaning the data and modifying the data format to combine them, as a result, the data collection time was four five times longer than data analysis. As commercial tools directed at the semiconductor engineering market evolve, they will obviously need to address this issue.

There are some notable trends in recent years, such as the availability of semiconductor-industry specific commercial tools with more flexible and powerful integrated data access and analysis capabilities. While general statistical software such as MS/Excel, SAS, S-Plus, Statistica, and others will continue to improve in the depth and breadth of their features, they might still fall short in providing the industry-specific automated routines.
needed by engineers. In addition to one or more general statistical software tools, every company will need a complementary in-house or commercial yield management tool. Since many semiconductor companies will develop new data sources over time, analysis applications will need to be flexible enough to interface to new data sources as they evolve.

With the large amount of data being generated to support semiconductor manufacturing, new tools are needed bring diverse types of data together and provide comprehensive analysis. Point solutions need to evolve into integrated capabilities, and new statistical technologies, such as data mining, and database technologies, such as data warehousing and online analytical processing will act to enable portions of this evolution.
Part III: Monitoring Defects in IC Fabrication Using a Hotelling $T^2$ Control Chart

One significant problem that can seriously hinder the yield of a semiconductor product line is the problem of false alarms during testing. By “false alarm” we are referring to process control limits being exceeded without the process actually being out of control. This may negatively affect yield by reducing the capacity of the product line due to downtime for unnecessary troubleshooting and error resolution. False alarms also have a negative effect on workers who may get accustomed to so many false alarms that they no longer take quality metrics seriously.

Many false alarms in semiconductor manufacturing come from a phenomenon known as clustering. The simplest thing to apply process control techniques to in semiconductor processing is the number of defects per wafer, where a defect is a defective die. One would expect these defects to be spatially distributed in a random fashion for a process that is in control. However, defects on wafers are often grouped together in clumps, called clusters. This phenomenon changes the true yield of the wafer from that expected due to the number of defects. As a simplified example, consider a wafer with ten randomly distributed defects that are serious enough to cause failure of the die on which they occur. By counting the number of defects, one knows how many dies are defective. If there were ten defects, then there should be ten defective dies. However, say the ten defects are arranged such that four of them are contiguous die in a box shape, another four are another contiguous box shaped set, and the remaining two are two contiguous die next to each other. In order for this to happen, the actual defect (say a particulate contaminant) may be in the center of the cluster of dies and barely touching each. If this were the case, and the clusters were distributed randomly, then the process may very well still be in control, even though by only monitoring the number of defects — since it is so high — would result in exceeding a control set-point, and thus result as a false alarm.

So, how should clustering be taken into consideration when determining wafer yield due to defects? Jun$^1$ proposed one method that devised what he called a “clustering index” (CI). This CI was more useful in depicting the clustering phenomenon than previous attempts because the CI value is independent of chip size and any assumptions about the distribution of the clusters or the defect distribution within the clusters.

Jun’s method involved indexing defects on an x and y coordinate plane on the wafer. The defects were then sorted in ascending order along both the x and y axis. The intervals between the defects in the x and y coordinates were defined by a sequence where $v_i = x(i) - x(i-1)$ and $w_i = y(i) - y(i-1)$ for $i=1,2,3,...,n$, where $x(0) = y(0) = 0$. The figure below from Jun shows graphically how these intervals are created.

---

Figure 7 shows how clustering can be identified as clumps along the x and y coordinates. In order to quantify this clumping along the axis, Jun used the coefficients of variation of the intervals. Therefore normal distributions of defects will result in a coefficient of variation that is close to one, while heavy clustering will result in coefficients of variation of greater than one. In particular, the CI is defined as follows:

\[ CI = \min \left( \frac{S_v^2}{\bar{V}^2}, \frac{S_w^2}{\bar{W}^2} \right) \]

\[ \bar{V} = \frac{\sum V_i}{n}, S_v^2 = \frac{\sum (V_i - \bar{V})^2}{(n-1)} \]

\[ \bar{W} = \frac{\sum W_i}{n}, S_w^2 = \frac{\sum (W_i - \bar{W})^2}{(n-1)} \]

\[ V_i = x_i - x_{i-1}, W_i = y_i - y_{i-1} \]

Jun went on to show that using this clustering index for modeling predicted yield actually matches actual yield much better than previous methods such as a negative binomial model. Therefore Jun provided not only a method for analyzing the phenomenon of clustering, but also an accurate way to model yield given only defect numbers and locations on a wafer.

Unfortunately, this CI alone and the yield model that goes along with it does not help to reduce the number of false alarms from process control. This is where Tong comes in with an approach to integrate Jun’s CI into a control chart. Simply stated, Tong proposes to monitor both the number of defects and the CI of the defects in a Hotelling T^2 multivariate control chart. He reasons that since both factors affect actual yield both should be monitored. He further argued that monitoring them individually could lead to misleading results since the interaction between the two effects is not analyzed by univariate charts.

Before we delve into Tong’s method we will quickly review the Hotelling T^2 control charts. Multivariate, like univariate control is based on the assumption that a process is in control if the output corresponds to some sort of fixed distribution (typically assumed to be a normal distribution). For multivariate control, the distribution is defined as a combination of the two individual distributions combined through the covariance of

---


the two variables to a single distribution of dimensions equal to the number of variables monitored. Below is a graph from Montgomery that shows a multivariate normal distribution for two factors (bivariate distribution). Next to it, also from Montgomery, is a depiction of how two factors can be plotted together.

![Bivariate Distribution](image1)

![Hotelling Control Chart](image2)

It should be apparent that the blue region in Figure 9 represents the control limits of the Hotelling $T^2$ control chart, and that this region is defined by the bivariate distribution, just as the control limits for a standard univariate chart come from the normal distribution. Instead of using a two axis scatter plot, though, the $T^2$ statistic boils the multiple variables down into a single statistic that can be plotted just like a standard univariate control chart. The general equations for the Hotelling $T^2$ statistic and control limits are:

\[
T^2 = (X - \bar{X})'S^{-1}(X - \bar{X})
\]

\[
UCL = \frac{p(m-1)}{m-p} \times F_{\alpha, m-n-p}
\]

$X$ is the $p$ dimensional quality characteristics ($p=2$ in this case), and $\bar{X}$ and $S$ are the common estimators for the mean vector and covariance matrices.

It should also be apparent looking at Figure 9 that a measurement that is out-of-specification on the univariate chart might be in specification on the Hotelling chart (the upper right and lower left corners of the chart), and that a sample that is in specification on the univariate chart might be out of specification on the Hotelling chart (the upper left and lower right corners of the chart).

This fact—that univariate out-of-specs may be in-spec for multivariate control—is important because it illustrates that by using a Hotelling chart for the factors of number of defects and CI, that there could be samples where the number of defects would normally cause an alarm on a standard control chart but do not cause an alarm on the Hotelling chart. This would occur if there were a large number of defects with a high CI, or a very low number of defects with a low CI (assuming the two factors have positive correlation). Assuming that high CI is a valid excuse for having a high number of defects, this will clearly reduce the number of false alarms, which was the entire point of the exercise.

Previously, instead of examining defects and clustering together, process engineers attempted to find a distribution that fit clustered defects. Without clustering the distribution is assumed to be Poisson which assumes that each defect is independent of the rest. Clustering, however, demonstrates that each defect is clearly not independent of
the rest. Friedman\textsuperscript{4} proposed that the Neyman type-A distribution better represents the number of defects expected on a wafer in the presence of the clustering phenomenon. The Neyman distribution is analogous to assuming that the number of clusters is Poisson distributed with mean $\lambda$ and the number of defects in each cluster is Poisson distributed with mean $\phi$. Therefore the mean number of defects is $\lambda \phi$ and the variance is $\lambda \phi(1+\phi)$, which has the effect of widening the normal control chart limits by a constant factor. Friedman went on to show in examples of clustered defects vs. an out-of-control process that this method reduces the number of false alarms.

Clearly Tong’s use of a Hotelling control method is superior in theory to Friedman’s standard method with adjusted control limits because the CI that Tong uses is completely independent of the distribution of clusters or defects within clusters. In order for Friedman’s method to be accurate, the assumption that clusters are Poisson distributions and that the defects within a cluster are also Poisson distributions must hold true. While Friedman did show some empirical evidence that supports this claim, a method—such as Tong’s—that doesn’t rely on such tenuous claims is much better supported.

Yet, Tong’s method is not without its faults. First, Tong still relies on the assumption—as does Friedman—that with a high degree of clustering, large numbers of defects can result from a process that is in control. This assumption makes sense to a certain degree. As in the example given above, if the clusters are caused by single defect mechanism, then the number of defects will be artificially high and not representative of the actual process. Yet, taking this example to the extreme demonstrates how the assumption can break down. Say, for instance, that the entire wafer is completely defective. Then it would have a large number of defects, but would also have a very large CI (since it would essentially be one big cluster). In reality, this situation would be recognized for what it is, and in fact Tong’s method has a step that removes outliers and a step that ensures normality of the defect data and CI data, but this still shows that the assumption has it’s limits. It would seem that monitoring a third factor that takes cluster size into account would also improve the process, since clearly a cluster that is abnormally large is indicative of a process that is not in control.

A second drawback to Tong’s method is the complexity of calculations involved in his method. Multivariate charts require manipulation of matrices of order equal to the number factors involved. Although 2\textsuperscript{nd} order linear arithmetic is not prohibitively difficult, it does add complexity. Multivariate control also requires the calculation of covariance terms which is another nuisance. The most notable contributor to complexity however is determining the cause of a control limit violation. For univariate analysis the factor being monitored is the cause of the limit violation and troubleshooting begins naturally from this knowledge. For the Hotelling control charts though, since two factors are combined into a single limit, one must first decompose the data to reveal which factor caused the violation and therefore where to begin troubleshooting. Again, this involves straight forward, although tedious calculations according to the following equations:

\footnotesize
\begin{eqnarray}
\text{Equation 1} & \quad & \\
\text{Equation 2} & \quad & \\
\text{Equation 3} & \quad & \\
\end{eqnarray}

\[ T^2 = T_{1,2}^2 + T_{2,1}^2 = T_{2,2}^2 + T_{1,1}^2 \quad \text{and} \quad UCL_j = \frac{m+1}{m} F_{\alpha,1,m-1} \]

The only remaining critique of Tong’s method is to ask why Hotelling $T^2$ charts were chosen for multivariate control. Exponentially weighted moving average (EWMA) control charts have been used for many years and are quite popular in the field of semiconductor manufacturing for process control. EWMA, like cusum charts, allow for identification of a slowly shifting process mean, and are therefore quite useful for monitoring processes where there is a very gradual shift in the process. Semiconductor manufacturing has many processes where this is the case, such as a lowering plasma CVD deposition rate due to buildup on the chamber walls. In addition to the standard EWMA charts, there is a multivariate form that could be used in essentially the same way that Tong is using the Hotelling $T^2$ charts, with the added benefit of being able to monitor the process for a slowly changing mean. Multivariate EWMA is probably not as well developed or known as multivariate Hotelling, but there has been thorough analysis of the subject. $^3$

Overall, Tong’s method of using Hotelling $T^2$ control charts to merge clustering index with defect number monitoring seems very promising. Although it is computationally rigorous, it surpasses Friedman’s use of the Neyman distribution to account for clustering because it doesn’t depend on knowing the distribution of clusters or defects within clusters, and it has a direct effect on yield because it reduces the number of false alarms.
References:

1. www.pintailtechnologies.com


