# Some Considerations on Choosing An Outlier Method for Automotive Product Lines

Li-C. Wang, Sebastian Siatkowski, Chuanhe (Jay) Shan, Matthew Nero
University of California, Santa Barbara

Nikolas Sumikawa, LeRoy Winemberg
NXP Semiconductors

***Abstract*—Outlier screening is a popular approach employed for automotive product lines. There have been many outlier methods proposed. In practice, it is desirable to choose the "best" outlier method. This work develops a notion of *applicability* associated with an outlier method on a given set of wafers. A measure for applicability is proposed and experiment results are presented to illustrate its effects for finding outliers and for analyzing customer returns based on data collected from several automotive product lines.**

## 1. Introduction

Outlier screening is a popular approach employed in automotive product lines for screening parametric defects [1]. There have been many outlier methods proposed [2]. Well-known methods include the various Part Average Testing (PAT) methods such as Static PAT (SPAT), Dynamic PAT (DPAT), Automotive Electronic Council PAT (AEC DPAT), and Robust DPAT (RDPAT), as well as Nearest Neighbor Residuals (NNR) [3][4] and Location Average (LA) [3][5]. A question frequently asked in practice is: Which method is the best for my test application?

This questions can be asked in different contexts, for example, best for a family of product lines, best for a particular product line, or best for a specific test. Regardless in what context the question is asked, asking the question suggests that there exists a best method for the context. However, this assumption might not be true.

There can also be diverse ways to define what the "best" means. With a given definition, an evaluation is required to obtain some measure associated with each method, and to rank methods using their measured values. Such evaluation results are also subjective to the data in use.

For example, an evaluation can be based on the intuitive thinking: "The best method is the one that screens out the most of defective parts with the lowest yield loss." This objective, however, is hard to implement in practice because (1) the set of all possible defective parts is hard to define and (2) the number of outlier parts screened out by a method depends on the allowable yield loss.

. This work is supported in part by National Science Foundation Grant No. 1618118

In practice, one can modify the objective to: "The best method is the one that screens out the most in a given set of defective parts with a fixed yield loss budget $YL$."

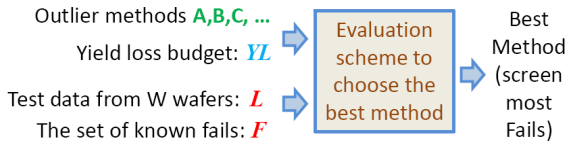


Figure 1. **One way to evaluate outlier methods**

With a given $YL$, one can collect a set of known defective parts, and determine the number of the defective parts that can be screened out under the budget $YL$. However, this approach has two concerns: (1) The evaluation is conducted based on some initial production data. The initial data might not be representative for the future data. (2) Similarly, the set of defective parts might not represent all defective parts that are screenable based on the given $YL$.

Suppose method $\phi$ is determined to be the best based on some initial data $L$ and the set of known defective parts $F$. Let $L'$ represent data produced in the future and $F'$ represent all screenable defective parts from $L'$. The fact that a method is the best based on the data $(L, F)$ does not mean that the method is the best in view of the data $(L', F')$.

One can think of the data $(L, F)$ as the *benchmark* used to compare methods. The two concerns mentioned are both regarding how "representative" the benchmark is. Although using benchmarks is a common practice for comparing methods, benchmarks can also be misleading.

### 1.1. No Free Lunch

Outlier analysis is commonly known as *unsupervised learning*, or can be thought of as *supervised learning* when there are actual outlier samples to verify an outlier model [6]. In the context of machine learning, the No Free Lunch (NFL) theorem [7] had warned about using benchmarks to evaluate machine learning algorithms. Indeed, without a guarantee on the representation of the benchmarks, it is meaningless to say that one algorithm is better than another. In a test application, this guarantee is especially hard to accomplish because characteristics of future data can deviate significantly from the benchmarks in use.


978-1-5386-3413-4/17/$31.00 

NFL is a general theorem. For a specific learning problem such as a particular outlier screening application scenario, it remains unknown if NFL applies. Proving or disproving a NFL property for a particular application scenario can be difficult. However, the theorem does raise the concern how well an evaluation result based on some initial data can generalize into future data.

There are two dimensions of concern for generalization from initial data $(L, F)$ to all (or future) data $(L', F')$: (1) from $L$ to $L'$ and (2) from $F$ to $F'$. This work focuses on the aspect from $L$ to $L'$. If this generalization is already poor, adding the second aspect can only make it worse.

### 1.2. Paper flow and the main idea proposed

Suppose outlier screening is applied on each wafer. The initial data $L$ comprises an initial set of wafers. All data $L'$ comprises a much larger set of wafers. In section 2, we will first show that (1) there is no universally best method, and (2) the generalization is poor from $L$ to $L'$. We call these outcomes the *seemingly NFL* because they are shown experimentally (not proved formally).

Section 3 further illustrates the difficulty of comparing outlier methods, where different methods actually disagree on what the top outliers should be. These results also indicate that for choosing a best outlier method in test, one may be facing a problem similar to that suggested by NFL.

In the rest of the paper, we discuss our approach to overcome the seemingly NFL barriers. The main idea is illustrated in Figure 2.

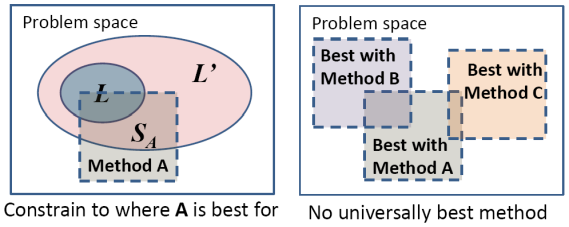


Figure 2. **The main idea of this paper**

There are two fundamental barriers. Given a problem space, a benchmark $L$ comprises a small subset of the instances in the space. The evaluation result using $L$ does not generalize to a larger subset of instances $L'$. The second barrier is that there is no one method better than another if we consider the entire problem space.

In our context, each problem space is based on a given test. Each instance in the space is a wafer. Hence, given a method $A$, the seemingly NFL situation indicates that the performance of $A$ seen on $L$ cannot generalize to the performance of $A$ on $L'$. To overcome this barrier, our idea is to *constrain* the space for this generalization to take place.

Our goal is to determine a subset of wafers, $S_A$, on which A is *applicable*. It is important to note that this goal is *not* to determine the exact applicable subset. Instead, an applicable subset serves only as a constraint to restrict the application of $A$ to those wafers in $S_A$. In other words, if a wafer is in $S_A$, we have a high confidence that it is applicable. If it is outside, applicability is simply *undecided*.

Our conjecture is that while generalization is poor from $L$ to $L'$, generalization can be improved if we constrain the scope to be only from $L \cap S_A$ to $L' \cap S_A$.

Furthermore, with such an applicable scope concept, we no longer need to decide a universally best method. As shown in Figure 2, each method can be the best on its respective applicable subset of wafers.

The key in our idea is therefore the *applicability* concept. In section 4, the *applicability* concept is developed and explained. Section 5 then explains how applicability can be measured with a set of wafers. In section 6, examples are provided to illustrate why inapplicability can occur. Then, in section 7, we show how applicability can be utilized to identify the applicable subset of wafers for a method and show its effect on the result presented earlier in section 3. In section 8, we show that poor generalization result seen in section 2 can be improved if we limit the scope to the subsets of applicable wafers. Finally, Section 9 concludes.

## 2. Choosing a method based on known fails

We consider five outlier methods: SPAT, DPAT, AEC DPAT, NNR and LA. For each product line, we obtain a set of customer returns (CQIs - customer quality incidents). We treat them as the set of "known fails," $F$. The evaluation data $L$ comprises the wafers from the lot where the CQI occurs. The total data $L'$ comprises all wafers we collected.

Table 1 shows the list of products and all the data used in this work. In some cases, most data collected are the CQI lots. In other cases more data are collected for non-CQI lots. Hence, data shown in the table has no reflection on the CQI rate for a product.

TABLE 1. DATA USED FOR THE STUDY

| Product code | # of wafers | # of tests | # of CQIs |
|---|---|---|---|
| VP | 51100 | 249 | 32 |
| KM | 9675 | 350 | 23 |
| A2M | 400 | 45 | 11 |
| MPC | 4888 | 620 | 10 |
| ALP | 6996 | 123 | 77 |

In the current experiment, we try to determine the "best" method (and the corresponding outlier model) for each CQI. With each method, we search for the test that results in an outlier model with the minimum yield loss based on $L$ (the CQI lot). We say that a method (and the model) is better for the CQI if this minimum yield loss $YL_{min}$ is smaller.

Let the outlier model achieving $YL_{min}$ be $M$. Then, we use $YL'_{min}$ to denote the yield loss by applying $M$ to all data $L'$. Poor generalization can then be observed if $YL'_{min}$ deviates significantly from $YL_{min}$. Poor generalization can also mean that the best method determined based on $YL_{min}$ is not the best method from the $YL'_{min}$ point of view.

### 2.1. No universally best method across CQIs

Figure 3 summarizes the results on $YL_{min}$ for product VP. Each vertical line corresponds to a CQI. Each marker corresponds to the $YL_{min}$ result from a method. In this picture, the "best" method for a CQI is the one with the lowest marker. From the figure, we observe: (1) different methods can achieve noticeably different results, and (2) there is no universally best method across all CQIs.

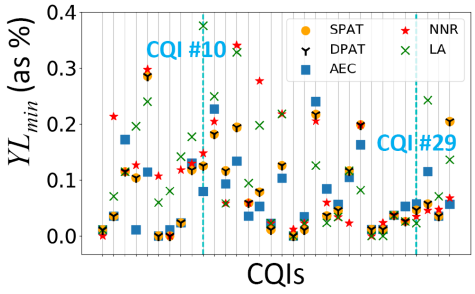


Figure 3. $YL_{min}$ **for each CQI - product VP**

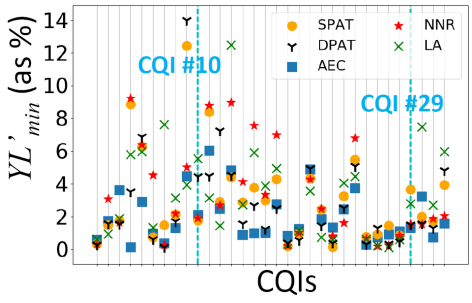


Figure 4. **The corresponding** $YL'_{min}$ **- product VP**

## 2.2. Poor generalization from $YL_{min}$ to $YL'_{min}$

Figure 4 then shows the corresponding $YL'_{min}$ number for each $YL_{min}$. The $YL'_{min}$ results are based on all data, i.e. the 51100 wafers as shown in Table 1. Observe that for most CQIs, the $YL'_{min}$ numbers are substantially larger than those $YL_{min}$ numbers (the scales on y-axis are different), i.e. $YL_{min}$ is a poor predictor for $YL'_{min}$. Furthermore, the best method determined using $YL_{min}$ is not necessarily the best method using $YL'_{min}$.

TABLE 2. $YL_{min}$, $YL'_{min}$, RANKING FOR CQI #10 AND CQI #29

| | | |
|---|---|---|
| CQI #10 | $YL_{min}$(%): | $0.08 < 0.125 < 0.125 < 0.148 < 0.385$ |
| | Ranking | AEC < SPAT < DPAT < NNR < LA |
| | $YL'_{min}$(%): | $1.691 < 1.840 < 2.082 < 4.451 < 5.908$ |
| | Ranking | SPAT < NNR < AEC < DPAT < LA |
| CQI #29 | $YL_{min}$(%): | $0.023 < 0.035 < 0.046 < 0.046 < 0.058$ |
| | Ranking | LA < NNR < SPAT < DPAT < AEC |
| | $YL'_{min}$(%): | $1.309 < 1.405 < 1.466 < 2.532 < 3.184$ |
| | Ranking | AEC < NNR < DPAT < LA < SPAT |

From Figure 3, we select two examples and show their details in Table 2. For example, for CQI #10, based on $YL_{min}$, the best method is AEC with $YL_{min} = 0.08\%$. The worst method is LA with $YL_{min} = 0.385\%$. However, when we move to $YL'_{min}$, the best method becomes SPAT with $YL'_{min} = 1.691\%$, and the worst is still LA.

For CQI #29, based on $YL_{min}$, the best method is LA and the worst is AEC. However, based on $YL'_{min}$, the best is AEC and the worst is SPAT.

TABLE 3. # OF CQIS A METHOD IS BEST FOR, BASED ON $YL'_{min}$

| Product | SPAT | DPAT | AEC | NNR | LA | total CQIs |
|---|---|---|---|---|---|---|
| VP | 3 | 7 | 11 | 5 | 6 | 32 |
| KM | 4 | 2 | 9 | 1 | 7 | 23 |
| A2M | 0 | 3 | 4 | 1 | 3 | 11 |
| MPC | 0 | 4 | 4 | 2 | 0 | 10 |
| ALP | 11 | 20 | 23 | 12 | 11 | 77 |

Table 3 then summarizes the result of "best method" for all products. Here the best method is from the $YL'_{min}$ point of view. Observe that for each product, each method is the best for a subset of CQIs.

The results above show that an evaluation result based on $L$ does not generalize well to all data $L'$. The earlier works [10][11] observe similar poor generalization based on different evaluation objectives. The results above also show that the best method is very much case-dependent. If there is no universally best method, then the next question becomes: Which method should I apply for a given case?

## 3. Disagreement among outlier methods

One might argue that the comparison in the previous section is not meaningful because we do not know if a CQI is a "true" outlier. Take CQI #10 as an example, the AEC model classifies the CQI as an "outlier" with 0.08% yield loss on the CQI lot and the model screens 2.082% of the dies from all lots. A model screening that many dies should not be a "true" outlier model. In other words, we are setting the outlier limit too tight and a comparison based on such a tight limit might not be that meaningful. For example, if one relaxes the outlier limits, it is possible that results from different methods become more agreeable.

To see if this is true or not, we perform a simple experiment. Given a test we apply each method to all the data $L'$. We then identify the top $N$ outliers from $L'$ given by each method. We let $N = 1, 10, 100$ which based on all the data we have, correspond to yield loss 0.063, 0.63, and 6.3 PPM (parts per million), respectively. These numbers are much smaller than those shown in Table 2.

Following the results presented in Table 2 before, for CQI #10 all methods use the same test (call it test $T_{10}$), except for LA which identifies a different test. For CQI #29, all methods use the same test (call it test $T_{29}$).

For $T_{10}$, when $N = 1$, all methods find the same part as the outlier. Figure 5 then show the results for $N = 10, 100$.

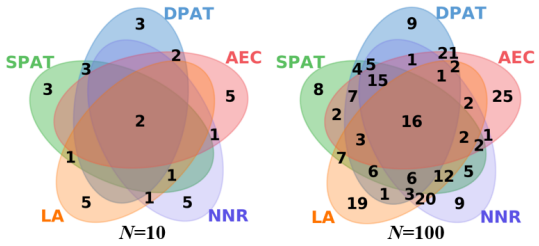


Figure 5. $N$ **outliers found by different methods** ($T_{10}$)

For $N = 10$, all methods agree only on 2 outlier parts. Each method also identifies its own unique outliers. For example, AEC identifies 5 parts as outliers unique to itself. As we increase $N$ to 100, observe that there are unique parts in every intersection of every three and four methods. The union for $N = 10$ has in total 32 outliers and the union for $N = 100$ has 214 outliers.

Figure 6 then shows similar results for test $T_{29}$. For this test, the five methods tend to disagree more on which parts should be the outliers. For example, for $N = 10, 100$, there is no part agreed by all five methods as an outlier. The union for $N = 10$ has 35 outliers and the union for $N = 100$ ends up with 431 outliers.

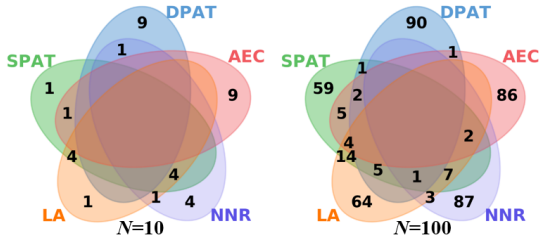


Figure 6. $N$ **outliers found by different methods** ($T_{29}$)

Results above show that even one relaxes the outlier limits to find the very few top outliers, different methods can still disagree on which parts should be the outliers. If two methods cannot even agree on which parts should be the top outliers, how can we decide which one is better? Later in Section 7 we will re-visit the result shown in Figure 5 and Figure 6 and show that while we cannot decide which method is better, with the applicability measure, we can decide which outliers are more trustable.

## 4. Applicability of an outlier method

Results shown above seem to reflect what the NFL theorem had suggested: (1) Benchmarks are misleading, and (2) In general no one method is better than another.

Given a set of wafers, even if one method is not generally better than another across all the wafers, it is still possible that if we restrict to a subset of wafers, one method is better than others. In other words, each method has its own subset of applicable wafers (e.g. see Figure 2 before).

Given a wafer $G_i$ and a method $\phi$ with a test $t_j$, our goal is to develop a method $APP(G_i, t_j, \phi)$ that calculates an *applicability* measure for applying $\phi$ on the wafer data $D_{ij}$. Let $\phi(D_{ij})$ be the result of applying $\phi$ on $D_{ij}$, our goal is to check some properties of the result $\phi(D_{ij})$ in order to decide if $\phi$ is *applicable* or not.

### 4.1. Two properties to check for applicability

Conceptually, let $E(\phi, t_j)$ denote the "expected result" of applying $\phi$ to a wafer based on $t_j$. The first property to check is the difference between $A(D_{ij})$ and $E(\phi, t_j)$, i.e. does the result meet the expectation? As it will be explained later, this property is to ensure that outlier decision made by method $\phi$ is *consistent* across wafers.

Then, based on the assumption that each $\phi()$ should follow a Normal distribution, the second property is to check on average how $\phi()$ deviates from the Normality assumption. This property is to ensure that outlier decision made by method $\phi$ is *justifiable*.

In other words, our notion of *applicability* is that the outlier decision made by applying a method on a given set of wafers is both *consistent* and *justifiable*.

To evaluate the first property, we will develop a *Variance* concept. To evaluate the second property, we will develop a *Bias* concept. Below we will explain their meanings with respect to a given set of wafers.

### 4.2. The basic assumption of an outlier

First, note that in outlier analysis there is no golden definition of what a "true" outlier should be. One of the most basic statements that can be said about an outlier is:

> Given a distribution $D$, a sample is an outlier if its probability of occurrence is so small that it is unlikely the sample is drawn from $D$.

This definition remains subjective to the probability threshold to define how small is small enough. However, this is the minimal about one must assume. The difficulty to apply this definition directly to find outliers is that one need to know the distribution $D$.

For example, suppose $n$ samples are drawn from a Normal distribution $\mathcal{N}(\mu, \sigma)$. For each sample value $x_i$, one can calculate the so-called z-score $z_i = \frac{x_i - \mu}{\sigma}$. A common method to identify outliers is then using the Grubb's test [12] which calculates the probability of the largest z-score given $n$ (Notice that this probability depends on $n$). For example, for $n = 2000$ and probability threshold $10^{-6}$, Grubb's test says if a sample has a z-score $> 6.19$, it is an outlier.

### 4.3. Density estimation vs. Outlier transform

Based on the basic definition above, in practice there can be two approaches to develop an outlier method as illustrated in Figure 7.

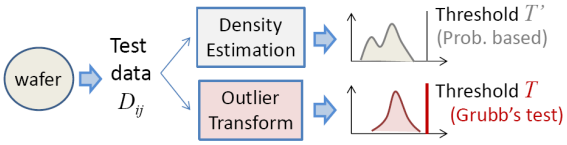


Figure 7. **Two approaches to develop an outlier method**

The first is by applying *density estimation* [8] to estimate the distribution $D$. Once the distribution is known, one applies a probability-based reasoning similar to the Grubb's test to find a threshold. Density estimation, however, is not very reliable for finding outliers because it is very difficult to accurately estimate the probability density on the tails of a distribution [9].

The second apporach is by applying an *outlier transform* to convert each test value into an *outlier score*. As a result, each test value distribution becomes an *outlier score distribution*. Then, a threshold is applied to each outlier score distribution to find outliers.

There are two concerns with the second approach. (1) If outlier score distributions on different wafers are different, then applying the same threshold on them have different meanings. In this case, the outlier decision is *inconsistent* across wafers. (2) To justify using the threshold on each outlier score distribution, we also desire each outlier score distribution to follow the assumed distribution where the probability-based reasoning is applied to derive the threshold. Otherwise, the threshold lacks a justification to be used with the outlier score distributions.

These two concerns reflect the two properties of applicability in section 4.1 above. In the following, we will explain how our *Variance* concept reflects the first concerns and our *Bias* concept reflects the second concern.

### 4.4. Outlier methods in test are outlier transforms

An outlier method in test is basically a transform of a test value into an *outlier score*. For example, given a test

value $t$ for a die $c$, the five methods studied in this work (SPAT, DPAT, AEC DPAT, NNR, LA) can all be thought of as following the abstract transform to obtain an outlier score $O_c$:

$$O_c = \frac{t - E(c)}{\varsigma} \quad (1)$$

where $E(c)$ is the *expected test value* of die $c$, and $\varsigma$ is the quantity to normalize the score. In other words, different methods differ in how they calculate the expected value and what normalization value should be used.

For example, SPAT uses the sample mean $\mu_s$ of test values across the entire wafer for $E(c)$ and uses $\varsigma = 1$, i.e. it does not normalize. DPAT also uses $\mu_s$ as $E(c)$ and in addition, uses the sample standard deviation $\sigma_s$ as the normalization value, i.e. $\varsigma = \sigma_s$.

AEC uses the sample median value $M_s$ across the wafer for $E(c)$. AEC uses "$0.43 \times (P_{99} - M_s)$" or "$0.43 \times (M_s - P_{01})$" as the normalization value, depending on which side of the test value $t$ of $c$ is located with respect to the median, where $P_{99}$ and $P_{01}$ are the 99% and 1% quantile values of the test value distribution, respectively.

NNR also uses the sample median value $M_c$, but the samples are not from the entire wafer. Rather, a $k \times k$ window (say $7 \times 7$) is decided centering on the die $c$. Only dies located in the window are used to calculated $M_c$ (for $7 \times 7$, there are up to 49 dies). In this way, $M_c$ can be different for different dies. Typically, NNR uses $\varsigma = 1$.

LA also uses a window to calculate $M_c$ and typically uses $\varsigma = 1$. The difference is that $M_c$ is not calculated based on all available dies in the window. Rather, it is based on only a percentage (say 50%) of the dies whose test values are the closest to the test value $t$ of $c$.

Overall, every method makes an assumption of what the expected test value should be. Then, it makes an assumption of what a fair comparison should be by deciding how to normalize the deviation value "$t - E(c)$."

### 4.5. Concern of lack of consistency across wafers

Given a set of dies on a wafer $G_i$, conceptually their test values based on a given test can be thought of as forming a distribution $D_i$. Note that the discussion from this point on assumes the test $t_j$ is fixed. Hence, instead of using $D_{ij}$ to denote the test data as before, we simply use $D_i$.

An outlier method $\phi()$ transforms each test value into an outlier score. Effectively, the result is another distribution $\phi(D_i)$. Hence, given $W$ wafers, the result of outlier analysis by a method can be represented as a sequence of outlier score distributions $\phi(D_1), \ldots, \phi(D_W)$.

In outlier screening, one decides a threshold $T$ to identify outliers. This $T$ is repeatedly applied to each of the outlier score distributions $\phi(D_1), \ldots, \phi(D_W)$. If these score distributions are different (e.g. $\phi(D_i) \neq \phi(D_j)$ for $i \neq j$), it means that outlier decision made on different wafers are different, i.e. the decision is not *consistent* across wafers.

Ideally, to be consistent we would like to have $\phi(D_1) = \cdots = \phi(D_W) = D'$ as illustrated in Figure 8 where $D'$ can be the *expected* outlier score distribution ($E(\phi, t_j)$) as mentioned in section 4.1 before.

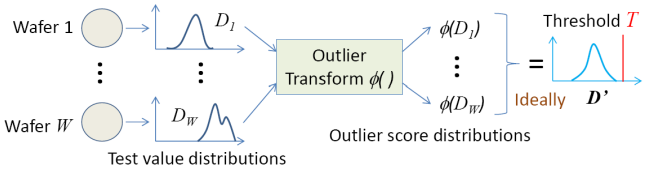


Figure 8. **Consistency with an outlier method** $\phi()$

### 4.6. Assessing Consistency with Variance

When $\phi(D_i) \neq D'$ for some of the $i$, we would like to have a way to estimate the degree of discrepancy.

Suppose we have a distance function $DIST()$ that can measures a *distance* (difference) between two given distributions, i.e. let $d'_i = DIST(\phi(D_i), D')$. Then, we can define the *Variance* of an outlier transform across $W$ wafers as:

$$\text{Variance of method } \phi() \text{ on } W \text{ wafers } = \frac{\sum_{i=1}^{W} (d'_i)^2}{W} \quad (2)$$

A larger Variance means the method is less consistent across the $W$ wafers.

### 4.7. Assessing Justifiability with Bias

The expected distribution $D'$ can be thought of as the *average* distribution across all $\phi(D_1), \ldots, \phi(D_W)$. To make the threshold $T$ a probability-justifiable decision, we would like $D'$ to be close to a known distribution. For example, we assume $D'$ should be Normal to justify using Grubb's test. Let $\mu_D$ and $\sigma_D$ be the sample mean and sample standard deviation of $D'$. Let $D$ be the Normal distribution $\mathcal{N}(\mu_D, \sigma_D)$. We therefore can model the degree of *justifiability* as a *Bias*:

$$\text{Bias of method } \phi() \text{ on } W \text{ wafers } = DIST(D', D) \quad (3)$$

The larger the Bias is, the less justifiable the method is. Figure 9 illustrates the Variance and Bias concepts.

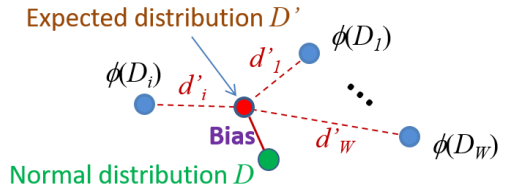


Figure 9. **Variance and Bias in outlier transform**

It is important to note that a large Bias does not necessarily imply the method is not justifiable *in general*. It only means we cannot justify it based on our Normality assumption of $D$. Hence, applicability is constrained by the assumed $D$. Consequently, lack of applicability should not be interpreted as "not applicable." A more meaningful interpretation is that we do not know if it is applicable or not. As mentioned in Section 1.2 with Figure 2 before, we treat applicability as a constraint to identify wafers where a method can be applied with a high confidence.

## 5. Calculating Variance and Bias

In order to calculate Variance and Bias, we require a way to implement the distance function $DIST()$. Our choice for $DIST()$ is the popular Kolmogorov-Smirnoff (KS) test. The

KS test, denoted as $KS(D_a, D_b)$, measures a discrepancy between two given distributions $D_a, D_b$. The result is a value between 0 and 1, where 0 means the same and 1 means the most different. In practice, a KS tool [13] takes $D_a$ and $D_b$ as two vectors of sample values $\vec{v}_a, \vec{v}_b$. Note that the lengths of these two vectors can be different.

Hence, with $W$ outlier score distributions $\phi(D_1)$, ..., $\phi(D_W)$, each $\phi(D_i)$ is represented as a vector $\vec{v}_i$ of outlier scores. The $KS()$ tool enables us to obtain pairwise measures $d_{ij} = DIST(\phi(D_i), \phi(D_j)) = KS(\vec{v}_i, \vec{v}_j)$, for $i \neq j$.

To calculate the Variance and Bias with equations (2) and (3), we need to calculate the *average* distribution $D'$. With the KS tool, $D'$ is simply the vector $\vec{v}'$ comprising the union of outlier scores from $\vec{v}_1, \ldots, \vec{v}_W$, i.e. $\vec{v}' = \vec{v}_1 \cup \cdots \cup \vec{v}_W$. Then, we can obtain each $d'_i = KS(\vec{v}', \vec{v}_i)$.

To calculate a Bias, we simply perform a random sampling based on the assumed Normal distribution $D$ to obtain a vector of scores $\vec{v}$. Then Bias is calculated as $KS(\vec{v}', \vec{v})$.

### 5.1. Result visualization

In order to visualize a result like Figure 9, we also need to project each distribution $\phi(D_i)$ (i.e. the $\vec{v}_i$) as a point in Euclidean space. The trick is that in this space, the points should be positioned relative to their pairwise distances $KS(\vec{v}_i, \vec{v}_j)$. This is a typical *multidimensional scaling* (MDS) problem. We utilize an MDS tool from [14] to project each distribution into a 3-dimensional space.

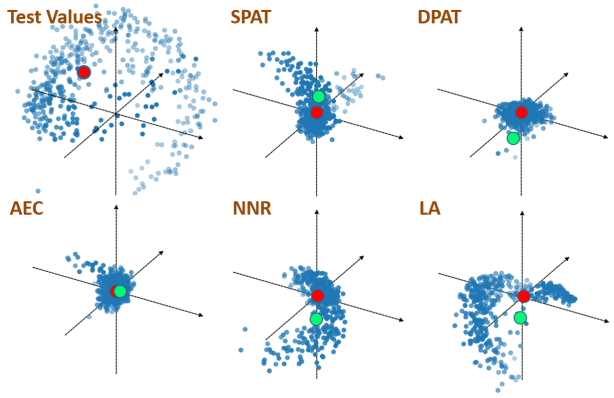


Figure 10. **Visualizing Variance and Bias for one test**

Figure 10 shows plots similar to Figure 9 for one particular test from product VP. The results here are based on 500 wafers. The computation is limited to $W = 500$ wafers because to produce each plot, we need to compute $W$-choose-2 pairwise distances $KS(\vec{v}_i, \vec{v}_j)$.

In Figure 10, each blue-gray dot represents a distribution (from a wafer). The first plot "Test Values" shows the original test value distributions $D_1, \ldots, D_{500}$. Then, in each method plot, the score distributions $\phi(D_1), \ldots, \phi(D_{500})$ are shown. The red dot marks the *expected* distribution $D'$ which is also used to center each plot. The green dot marks the assumed Normal distribution $D$ based on $D'$.

In Figure 10, a larger spread means a larger Variance and hence the result is less consistent. A larger distance between the red dot and the green dot means a larger Bias and hence, the result is less justifiable.

Figure 10 shows that DPAT and AEC have smaller Variance than others. Table 4 shows their Variance and Bias values. First, observe that all methods achieve a *variance reduction*. For example, DPAT brings the variance from its original value 0.223 down to 0.0094. This means that the original test value distributions across wafers are much more diverse. After an outlier transform, the resulting distributions become more similar, enabling one to make a more consistent outlier decision across wafers than using the original test values. This variance reduction can be thought of as a key objective of an outlier transform.

TABLE 4. # VARIANCE AND BIAS FROM DIFFERENT METHODS

| | Original | SPAT | DPAT | AEC | NNR | LA |
|---|---|---|---|---|---|---|
| Variance | 0.223 | 0.046 | 0.0094 | 0.0104 | 0.0543 | 0.0718 |
| Bias | — | 0.163 | 0.158 | 0.0633 | 0.174 | 0.174 |

While DPAT has the smallest variance, its Bias is not the smallest. Overall, we see that AEC is better because its Variance value and Bias value are both small.

### 5.2. Best method for a particular wafer

For a given wafer $G_i$, we can further compare methods based on the distance directly to the Normality assumption $D$, i.e. $d_i = DIST(\phi(D_i), D)$. We say that the smaller the $d_i$ is, the more *applicable* the $\phi()$ is for the wafer because $d_i$ can be thought of as measuring the combined effect of both consistency and justifiability together.

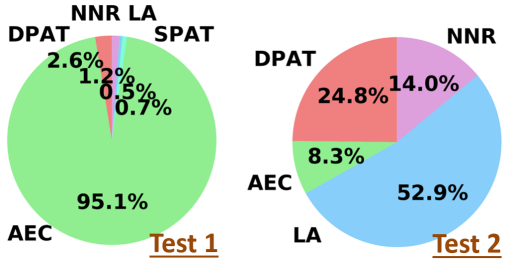


Figure 11. **% of wafers a method is best on**

Based on $d_i$, Figure 11 shows the percentage of wafers a method is the best on. "Test 1" is the test used to produce Figure 10. We see that AEC is the best on 95.1% of the wafers, an expected result based on what we observe in Figure 10. We further select a "Test 2" to show a contrasting result, where LA is the best on 52.9% of the wafers. These examples illustrate that no single method is the best on every wafer and moreover, different tests result in different evaluation results.

### 5.3. Results across tests

For a given test, we can say that a method is the best if its Variance is the smallest or if its Bias is the smallest. These two measures provide different perspectives to examine how overall a method is applicable with the test. Based on the two perspectives, Figure 12 shows the percentage of tests a method is best with, across all 249 tests from product VP. DPAT has the largest percentages in both cases.

Further, if we consider all (1387) tests across the five products, Figure 13 shows similar results.

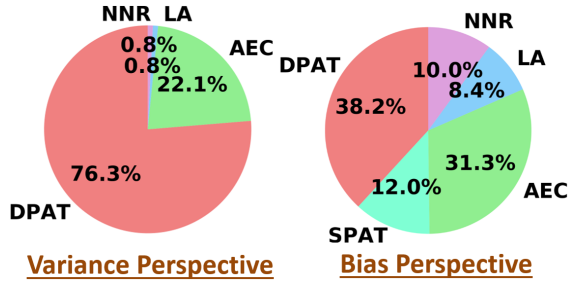


Figure 12. **% of tests a method is best with - product VP**

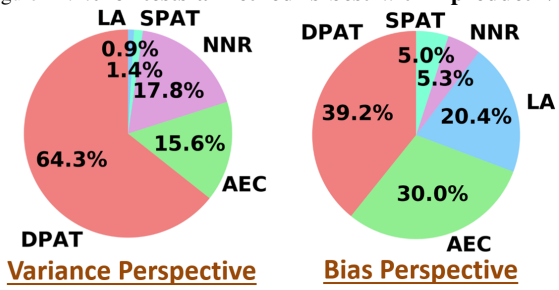


Figure 13. **% of tests a method is best with - all products**

Overall, even though DPAT has the largest percentages, every method shows its unique values. Thus in a test application, one should not discard any of the methods in advance.

## 6. Examples to illustrate Variance and Bias

Figure 14 shows the test value distributions of two wafers selected from Figure 10 presented before. Observe that these two distributions are quite different. In fact, their $KS$ distance is 0.905 (recall that $0 \leq KS() \leq 1$).

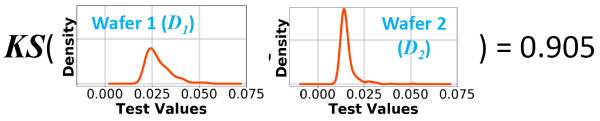


Figure 14. **Original test value distributions** $D_1, D_2$

Figure 14 then shows the respective outlier score distributions resulting from DPAT transform. Their $KS$ distance is 0.151, substantially smaller than 0.905. This illustrates the variance reduction effect with DPAT transform.

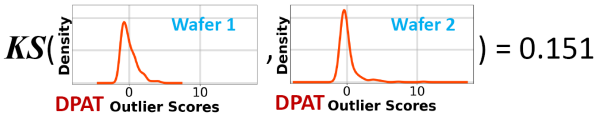


Figure 15. **Effect of DPAT,** $KS(DPAT(D_1), DPAT(D_2))$

In comparison, Figure 16 then shows the respective outlier score distributions resulting from LA transform. Their $KS$ distance is larger than the DPAT distance. From this perspective, we can say that LA is not as effective as DPAT.

The example illustrates how in Figure 10 LA has a larger spread (a larger Variance) than DPAT. In fact, the two wafers are selected based on Figure 10 where their distances are among the farthest in the LA plot.

### 6.1. Simulated examples to illustrate Bias

In general, there is an implicit assumption with every outlier transform on what properties a test value distribution

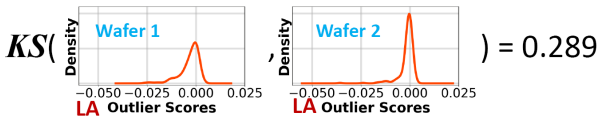


Figure 16. **Effect of LA,** $KS(LA(D_1), LA(D_2))$

should have. While this assumption might not have been explicitly or formally characterized with a method, such an assumption should always exist. Consequently, when this assumption is violated, a large Bias can occur.

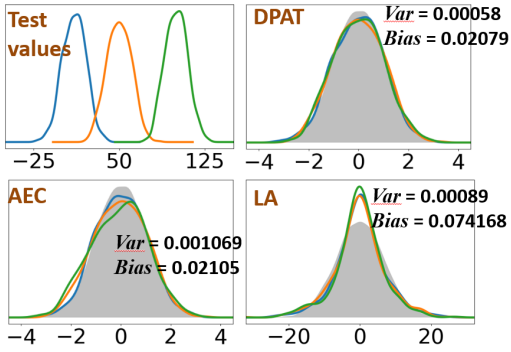


Figure 17. **3-wafer examples, all Normal**

For example, Figure 17 depicts an example with 3 wafers. Each test value distribution is sampled from a Normal distribution. 400 dies are sampled from $\mathcal{N}(10, 10), \mathcal{N}(50, 10)$, and $\mathcal{N}(100, 10)$, respectively.

Because the test value sample distributions are all Normal, it is expected that DPAT would have a small Bias. As seen in the DPAT plot, the 3 resulting outlier score distributions (colored similarly to the "Test values" plot with blue, orange, and green) are close to the assumed Normal distribution $D$ (colored as the shaded gray area).

AEC is intended to improve DPAT by taking asymmetric distribution into account. We see that AEC also has a small Bias. LA has a slightly larger Bias because each outlier score is calculated based on a smaller sample size with dies in a given window. A smaller sample size can cause more noise. As seen, the 3 distributions from LA deviate more from the assumed Normal distribution (shaded gray area).

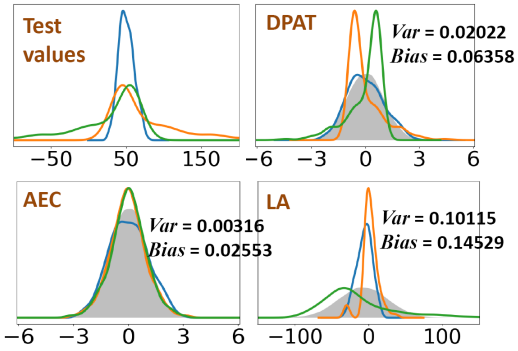


Figure 18. **3-wafer example that favors AEC**

Figure 18 then depicts a different 3-wafer example. The test value distributions are skewed (asymmetric) from a Normal distribution. This asymmetry violates DPAT's assumption but meets the assumption of AEC. As a result,

DPAT has a larger Bias. AEC's Bias remains comparable to that seen in Figure 17. Since LA does not consider asymmetric distribution of test values, its Bias increases from that in Figure 17 as well.

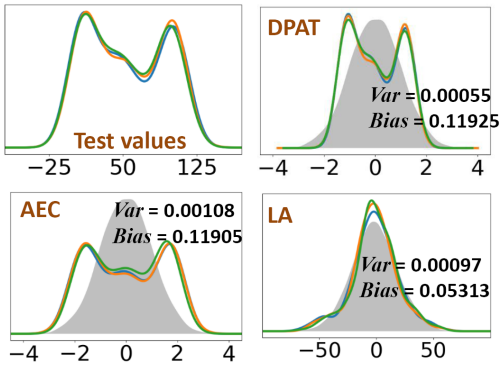


Figure 19. **3-wafer example that favors LA**

The assumption of LA is that there is some location-based wafer pattern across the test values. In Figure 19, on each wafer, we create a horizontal line pattern where each line is repeatedly sampled from one of the three Normal distributions, $\mathcal{N}(10,10), \mathcal{N}(50,10)$, and $\mathcal{N}(100,10)$. Each wafer is sampled in the same way. We expect LA to have a smaller Bias on such an example.

As seen in Figure 19, LA has a small Bias while DPAT and AEC, which do not consider any wafer pattern, both have a large Bias.

The three examples illustrate that a large Bias can be an indication that the test value distribution is out of the consideration by a method (i.e. violating its assumption of how the distribution should be). When a large Bias occurs, using the resulting outlier scores becomes not justifiable.

## 7. Applicability results

Let $D_{ij}$ denote the test value data based on test $t_j$ on wafer $G_i$. Let $D$ be our assumed Normal distribution. Given a method $\phi$, in section 5.2 before, we use the distance $d_{ij} = DIST(\phi(D_{ij}), D)$ to measure the combined effect of consistency and justifiability. Here, we can simply define *applicability* of an outlier method $\phi$ with $t_j$ on $G_i$ as

$$\text{Applicability: } APP(G_i, t_j, \phi) = 1 - d_{ij} \tag{4}$$

Let $H_{app}$ denote an applicability threshold. For example, for $H_{app} = 0.90$ we are interested in seeing how many wafer-test combinations have applicability greater than this threshold. Figure 20 summarizes such results.

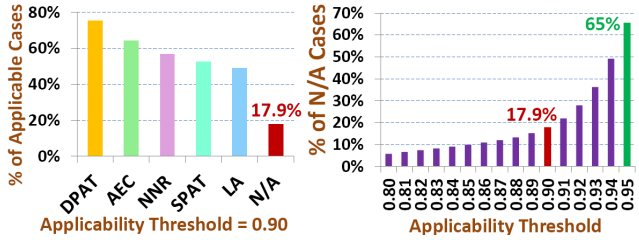


Figure 20. **Summary results with applicability thresholds**

The left chart shows, for all wafer-test combinations across all products, the percentage of wafer-test combinations for which a method is 0.9-applicable (or 90%-applicable). This percentage is not exclusive, i.e. multiple methods can be applicable for the same combination. In addition, we include an "N/A" **bar** to indicate the % of combinations where no method has applicability $\geq 0.90$.

As one changes the threshold $H_{app}$, the percentage of N/A combinations would also change. The right plot in Figure 20 therefore shows how the "N/A" percentage changes with respect to the change of the threshold $H_{app}$.

As seen in the right chart, if we set $H_{app} = 0.95$, then about 65% of the wafer-test combinations become N/A. One can think of $H_{app} = 0.95$ as 95% confidence to apply the outlier method to a wafer-test combination. Hence, the plot shows that for 65% of the wafer-test combinations, we do not have 95% confidence to apply any one of the five outlier methods. However, if we lower our confidence to 90%, the percentage of N/A is reduced to 17.9%.

### 7.1. Removing untrustable outliers

Recall that in Section 3 we show the undesirable disagreement results among the five methods. With the applicability measure, a simple resolution to the disagreement can be to remove "untrustable" outliers from each method.

Suppose we set $H_{app} = 0.9$ and consider any outliers found on a wafer-test combination with applicability below $H_{app}$ as "untrustable." Figure 21 shows the result after removing those untrustable outliers from the "$N = 100$" Venn diagrams of Figure 5 and Figure 6, respectively.

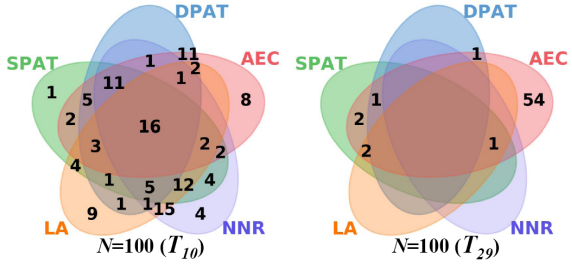


Figure 21. **Results compared to Figure 5 and Figure 6**

For $T_{10}$ the total number of outliers is reduced from 214 in Figure 5, to 121 in Figure 21. For $T_{29}$, the total number is reduced from 431 to 61. It is interesting to observe that for the result with more disagreement in $T_{29}$, the reduction is also more significant. It is also interesting to observe that for $T_{10}$, the 16 common outliers found by all five methods in Figure 5 also stay in the Venn diagram in Figure 21.

### 7.2. Usages of the applicability measure

Given a collection of outlier methods, Figure 20 shows that applicability can be used to identify the subset of wafer-test combinations for which each method is applicable, and the subset of combinations for which no method is applicable. Identifying the applicable subset with each method enables one to choose the best method for each combination with more confidence. Identifying the N/A subset enables one to assess if the collection of methods is sufficient.

Figure 21 further shows that applicability can be used to discard "untrustable" outliers found by a method. This avoids screening low-confidence outliers and hence, can help reduce yield loss. In the next section, we will re-visit the results presented in Section 2 and show how those total yield loss $YL'_{min}$ numbers can be reduced.

## 8. Re-visiting results in Section 2

In Section 2, the "best" outlier model is established for each CQI. For each CQI, we can use applicability to evaluate if the model (based on the particular method with the particular test) is indeed applicable on the CQI wafer. With an applicability threshold $H_{app} = 0.95$, Table 5 shows the # of CQI cases that are not applicable ("N/A") and the # of cases that are applicable. In total, about 38% ($= \frac{59}{153}$) of the CQI cases where their models are applicable.

TABLE 5. APPLICABILITY RESULT ACROSS ALL CQIS

| Product | VP | KM | A2M | MPC | ALP | Total |
|---|---|---|---|---|---|---|
| # CQIs | 32 | 23 | 11 | 10 | 77 | 153 |
| # "N/A" | 17 | 22 | 5 | 7 | 43 | 94 |
| # Applicable | 15 | 1 | 6 | 3 | 34 | 59 |

### 8.1. Improvement on total yield loss $YL'_{min}$

For the 59 applicable CQI cases, Figure 22 shows how much the total yield loss $YL'_{min}$ is reduced if the CQI outlier model is applied to only wafers with applicability $\geq 0.95$. For each CQI, the "Without applicability" bar denotes the $YL'_{min}$ from the original experiment described in Section 2. The "With applicability" marker denotes the new $YL'_{min}$. For a fair comparison, note that this new yield loss % is based on the total number of parts from only those applicable wafers, not the entire set of wafers as before. In other words, both $YL'_{min}$ numbers are the percentage from the same formula: $\frac{\text{total \# of parts screened out}}{\text{total \# of parts applied on}}$.

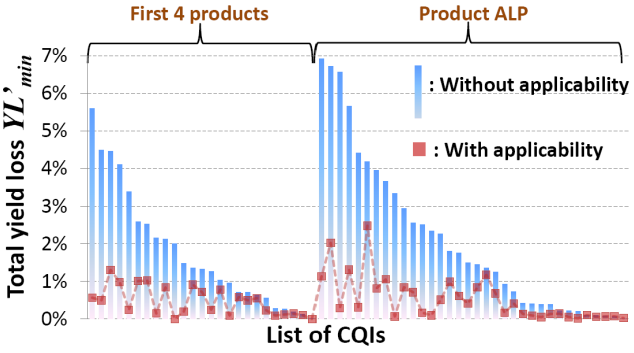


Figure 22. **Improvement on total yield loss** $YL'_{min}$

The improvement on the total yield loss can be clearly observed. This is interesting because applicability check reduces the total number of parts applied on. By reducing the denominator in the $YL'_{min}$ formula above (can be a significant reduction as indicated from Figure 20 that many cases (65%) are not applicable with $H_{app} = 0.95$), we observe that the resulting $YL'_{min}$ is also reduced. This implies that the % of yield loss on those not-applicable wafers is greater than the % of yield loss on those applicable wafers. This suggests that the wafers removed by the applicability check indeed contain the excessive yield loss.

### 8.2. Improvement on yield loss difference

In Section 2, poor generalization is shown by the significant difference between $YL_{min}$ and $YL'_{min}$. Based on this difference, "$YL'_{min} - YL_{min}$," Figure 23 shows how applicability check improves this generalization.

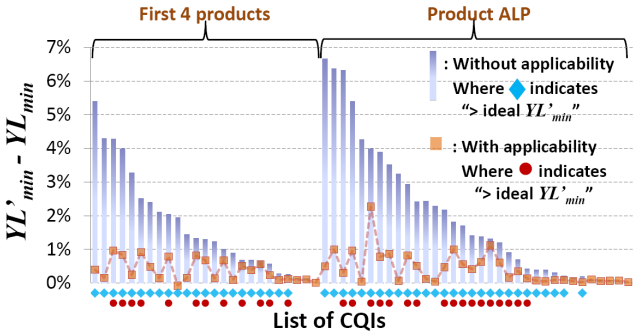


Figure 23. **Improvement on yield loss difference** $YL'_{min} - YL_{min}$

### 8.3. Ideal $YL'_{min}$ for each given $YL_{min}$

It is important to note that because $YL'_{min}$ is based on a (much) larger number of parts, it is expected that $YL'_{min} > YL_{min}$. Suppose all wafer outlier score distributions follow our assumption of Normal distribution $D$. Even with this ideal situation, we still expect to see $YL'_{min} > YL_{min}$.

To see this, for a given $YL_{min}$, we perform Monte Carlo simulation of the ideal situation based on the assumption $D$ and using the same number of parts in each case (e.g. on CQI lot, on all lots, without applicability, with applicability). Through this simulation, we calculate an *ideal* $YL'_{min}$ for each case. Table 6 shows some example results based on the first five CQIs shown in the figures above.

TABLE 6. EXAMPLES OF IDEAL $YL'_{min}$ BASED ON $YL_{min}$

| CQI | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Results without applicability check | | | | | |
| $YL_{min}$ = | 0.197% | 0.209% | 0.191% | 0.124% | 0.12% |
| Ideal $YL'_{min}$ = | 0.768% | 0.754% | 0.649% | 0.611% | 0.523% |
| Actual $YL'_{min}$ = | 5.597% | 4.496% | 4.461% | 4.107% | 3.39% |
| Results with applicability check | | | | | |
| $YL_{min}$ = | 0.172% | 0.343% | 0.341% | 0.154% | 0.00% |
| Ideal $YL'_{min}$ = | 0.683% | 1.06% | 1.124% | 0.653% | 0.184% |
| Actual $YL'_{min}$ = | 0.567% | 0.497% | 1.31% | 0.984% | 0.241% |

Note that with the applicability check, the $YL_{min}$ on the CQI lot can change as well, even though we do not change the outlier model. This is because some wafers from the CQI lot can be not-applicable.

As seen in the table, each ideal $YL'_{min}$ is greater than the $YL_{min}$, but not too much greater. For each case, if the actual $YL'_{min}$ is greater than the ideal $YL'_{min}$, then this means the total yield loss is not ideal.

Without applicability, we see that all cases have an actual $YL'_{min}$ above the ideal yield loss. And the difference between the ideal and the actual is quite large.

With applicability check, for CQI 1 and CQI 2, the actual $YL'_{min}$ is smaller than the ideal. Even for the other three cases, the actual $YL'_{min}$ is much closer to the ideal than those numbers without the applicability check.

Refer back to Figure 23. In the figure we use a diamond marker to note the CQI cases where the actual $YL'_{min}$ is greater than the ideal yield loss in the "Without applicability" experiment. We use a circle marker to note those similar CQI cases in the "With applicability" experiment.

As seen from those makers in Figure 23, without applicability many CQI cases have a total yield loss greater than the ideal. With applicability, many of these cases (without a marker) become ideal or smaller than the ideal yield loss. Note that for most of those CQIs with the circle marker, the differences between the actual $YL'_{min}$ and the ideal yield loss are quite small, similar to those shown in Table 6.

### 8.4. Consistency without justifiability

In the above experiment, we use the applicability equation (4). As mentioned above, this applicability reflects the combined effect of both consistency and justifiability. Suppose we change this to measure only consistency and discard justifiability. In other words, we use $CON(G_i, t_j, \phi) = 1 - d'_{ij}$ where $d'_{ij} = DIST(\phi(D_{ij}), D')$, i.e. distance to the average distribution $D'$, rather than to the Normal distribution $D$. By replacing $APP()$ with $CON()$, we re-perform the experiment for the 59 CQI cases by keeping all the other aspects in the experiment the same.

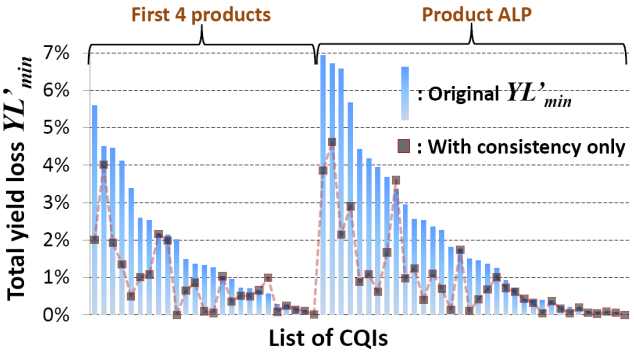


Figure 24. **Less improvement on total yield loss** $YL'_{min}$

Figure 24 shows a similar chart as that shown in Figure 22 above. Notice that, although the total yield loss $YL'_{min}$ is still reduced for many CQIs, the reductions are smaller than those shown in Figure 22. For a few CQIs, the new $YL'_{min}$ is the same as before or even worse.

Figure 24 shows the importance to include justifiability (i.e. Bias) in our applicability measure. Without checking the justifiability, we would not be able to obtain the substantial yield loss improvements as seen in Figure 22.

## 9. Conclusion

This work is motivated by the seemingly No-Free-Lunch phenomenon observed in the experiment results presented in Section 2 and Section 3. Our resolution to the seemingly NFL problem is by developing an *applicability* measure. The paper discusses the reasoning behind this applicability measure in terms of the Variance and Bias concepts.

Given an outlier method, the applicability is measured for each wafer-test combination individually. With an applicability threshold $H_{app}$, one can therefore decide if a given outlier method is applicable to each wafer-test combination or not. Intuitively, we explain that if a method is applicable, then it means the outlier decision is both consistent and justifiable for the wafer-test combination.

In Section 7 we provide examples to demonstrate the effect of our proposed applicability check and explain its possible usages. Then, in Section 8, based on experiments for analyzing 153 customer returns from five automotive product lines, we demonstrate that our applicability definition is indeed meaningful. In particular, we show that the applicability check can substantially improve the generalization of a yield loss result by an outlier model seen on a CQI lot, to the entire set of lots.

In practice, the applicability equation (4) can be implemented in an online fashion. Let $O$ be the set of outlier scores calculated for the parts from all lots up to a production point. Let $\mu_o$ and $\sigma_o$ be the sample mean and sample standard deviation of $O$. The $D$ distribution is simply the Normal distribution $\mathcal{N}(\mu_o, \sigma_o)$. In this way, $D$ can be adjusted incrementally. With this $D$ and an applicability threshold $H_{app}$, one can therefore determine the applicability of each outlier method to each wafer-test combination. In the future work, we plan to pursue such an implementation.

## References


[1] Guidelines for Part Average Testing. Automotive Electronic Council, AEC-Q001 Rev-D, December 9, 2011.

[2] Manuel J. Moreno-Lizaranzu and Federico Cuesta. Improving Electronic Sensor Reliability by Robust Outlier Screening. *Sensors* 2013, 13, pp. 13521–13542

[3] Ken Butler, et. al. Successful Development and Implementation of Statistical Outlier Techniques on 90nm and 65nm process driver devices IEEE *IRPS*, 2006, pp. 552–559.

[4] W. R. Daasch, J. McNames, D. Bockelman, K. Cota, R. Madge. Variance Reduction Using Wafer Patterns in IDDQ Data. International Test Conference, Oct 2000, pp. 189–198.

[5] R. Daasch, K. Cota, J. McNames, and R. Madge. Neighbor Selection for Variance Reduction in IDDQ and Other Parametric Data. IEEE ITC, Oct 2001, pp. 92–100.

[6] Charu C. Aggarwal. Outlier Analysis. Springer, 2013.

[7] David Wolpert. The Lack of A Priori Distinctions between Learning Algorithms. *Neural Computation*, 1996, pp. 1341-1390.

[8] B.W. Silverman. Density Estimation for Statistics and Data Analysis. Monographs on Stat. and App. Prob., v26, 1986.

[9] Jan Beirlant and Luc Devroye. On the impossibility of estimating densities in the extreme tail. *Statistics & Probability Letters*, 43, 1999, pp. 57-64.

[10] Sebastian Siatkowski, et al. Generalization of an outlier model into a "global" perspective. ITC, 2015.

[11] Sebastian Siatkowski, at el. Consistency in wafer based outlier screening IEEE VLSI Test Symposium, 2016.

[12] F. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics* 11, 1969, pp. 1-21.

[13] Eric Jones, Travis Oliphant, Pearu Peterson, and others. SciPy: Open source scientific tools for Python, http://www.scipy.org/, 2001-.

[14] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, 2011, pp. 2825-2830.