Use of Knowledge Discovery from Wafer Fab Data for Yield Improvement

Patrick J Carroll

RF Micro Devices, Inc. 7628 Thorndike Rd. Greensboro, NC 27409.
Email: pcarroll@rfmd.com, Phone: 336-678-5395

KEYWORDS: Yield Improvement, Knowledge Discovery, Data Mining, Data Analysis

Abstract

In manufacturing RF products from III-V semiconductors a massive amount of data is or potentially could be collected. The knowledge that can be obtained from that data is extremely valuable. The process of discovering knowledge from manufacturing data using data mining techniques is described. Proper treatment of the data can result in improved efficiency very quickly. A brief example will be given.

INTRODUCTION

When yield improvement is desired, given the background of most scientists and engineers working in compound semiconductor device manufacturing, it may seem natural to immediately hypothesize about the cause of the parametric performance leading to the yield loss based on device physics knowledge and start testing the hypotheses by performing experiments on wafers. These experiments may take a full manufacturing cycle to complete and may or may not lead to the final solution. This, after all, is the scientific method, isn't it? In a manufacturing environment this method, as described, is an inefficient way to make improvements. The initial input to the method is not nearly comprehensive enough. It ignores the wealth of knowledge that can be gained from the full set of manufacturing data.

Surely it would be nice to know what the manufacturing data has to say to us, but the process of producing RF components made from compound semiconductors involves a great number of steps, starting with the growth of the semiconductor boule, and including among other things: epitaxy, wafer fabrication, die singulation, assembly and test of the final components. Each step is quite complex involving multiple operations and a large amount of important data, including: process control information, electrical test data, time of each operation, tools used, process conditions, batches, shifts and operators. These steps may occur at various sites around the globe. The data may be stored in various databases, using different protocols. The information may be stored in ways that may make sense to the one who stores the information, not necessarily the user of it. Should data associated with the initial process and materials really be considered when evaluating the performance and quality of the end product? We believe that it should and could be done by using the process of Knowledge Discovery from manufacturing data, which is described here.

Data Mining is considered to be a step in Knowledge Discovery in Databases. But the terms are also often used interchangeably and it is now to the point where the acronym KDD, seems to have evolved from meaning Knowledge Discovery in Databases to Knowledge Discovery and Data Mining. KDD [1] is an active and growing field of study and application. By applying the principles of KDD to a manufacturing environment producing RF products from III-V semiconductors significant yield and quality improvements have resulted. KDD in this case can be a very efficient process. In this paper we will describe it in four steps: Assemble, Fix, Explore and Validate.

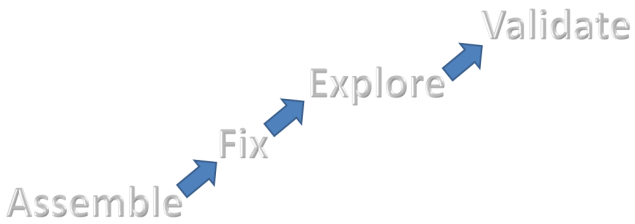


Figure 1. A simple KDD process description

ASSEMBLE

Data retrieval and assembly for almost any project is considered to be a very lengthy step but it doesn't have to be. Commercial systems can be purchased that provide a common platform from which to access data and perform the necessary data analysis. However, they can be quite costly and may be inflexible, not allowing the user to tailor the software to specific needs. At RFMD, JMP® is used as a standard platform for data analysis. We therefore decided to develop a JMP® script to retrieve data from existing sources that we call the Universal Data Query (UDQ). The chief value of the UDQ is that the user does not need to know all the intricacies of the many data tables and databases to obtain useful, properly linked data. Queries of the individual tables are optimized within the script to be efficient. Special knowledge needed to transform the data into a useful form is built in.

The UDQ starts by establishing a list of the wafers or product lots of interest. That list can either be obtained from

a table that contains the list of starts in the fab or may be provided by the user. The user then selects the data desired from a menu similar to the simplified version in Figure 2. Each menu item looks at one or more tables for the data. Each table may have a different requirement to produce quick efficient retrieval. For example, wafers may be called by different names depending on the table. The fab database tables are indexed by one particular wafer name. Before the wafer enters the fab, it is identified by the substrate name. After a wafer is shipped from the fab it is identified by a tracking number. This name is used to obtain data from die associated with that wafer at assembly and packaging. A comprehensive query obtained from the UDQ can take a matter of minutes and result in a table in JMP® containing many thousands of rows of wafers and many hundreds of columns of information ranging from boule information through package testing results.

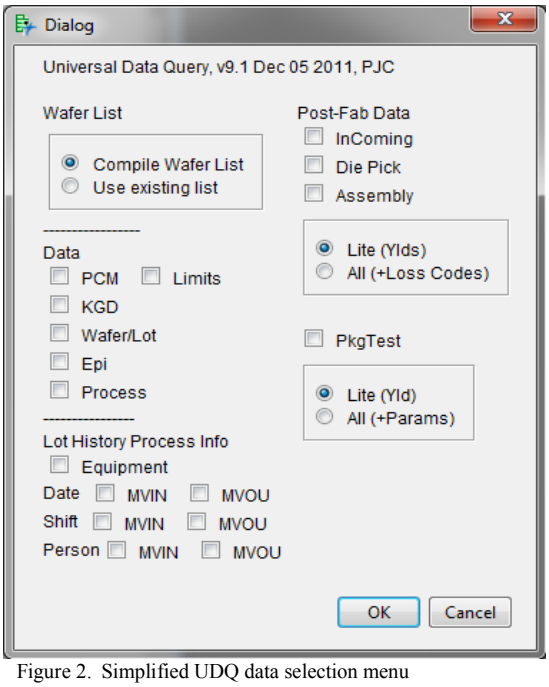


Figure 2. Simplified UDQ data selection menu

FIX

Once the data is obtained the accuracy of the data is an issue. Manufacturing data is notorious for being full of anomalies. On common problem is non-standard data entry. Fixing this usually requires specialized knowledge about the data or data source. An advantage of the UDQ is that this special knowledge can be built in. For example, the same parameter called by different names in two different fabrication sites can be standardized. Another problem is outliers. The UDQ can handle gross outliers, known to be caused by testing errors but there are also other data that lie clearly outside the usual distribution. The danger of filtering is that important information can be lost. There, however, is also danger in not filtering data, in that important relationships can be hidden.

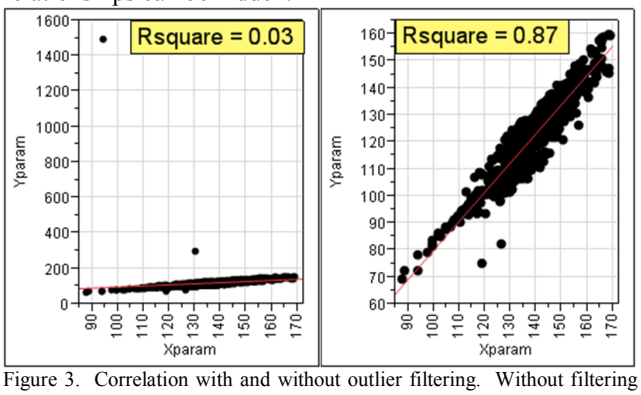


Figure 3. Correlation with and without outlier filtering. Without filtering this relationship would be missed by automated analysis.

If one wants to learn what most of the data is saying, and data exploration is to be done in semi-automated fashion, there must be outlier filtering that does not require manual term-by-term consideration. There are, of course, several methods that have been used for automated data filtering. The most common is probably the use of thresholds based on a number of standard deviations (sigma). The drawback of this is that tails will be "cut-off" of legitimate non-normal distributions and since outliers can have a great effect on sigma, some outliers can be missed. Use of inner percentiles of the data (e.g. 5th to 95th percentile) will eliminate date extremes from every term whether they are outlying or not. We therefore have developed a method of determining outliers based on normalized ordered differences (NODs). A full explanation of this method is beyond the scope of this paper. But to describe this briefly, the method attempts to determine larger than expected deviations that occur at the edges of the distribution. The data is ordered, a different vector is created and that is normalized by the mean absolute deviation from the median. Data outside a given threshold is then filtered.

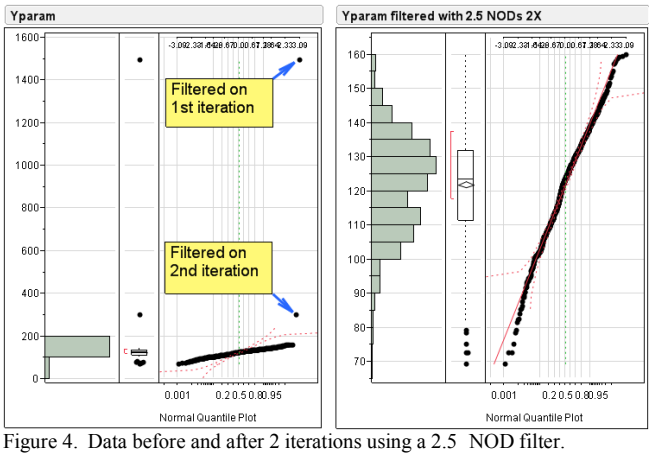


Figure 4. Data before and after 2 iterations using a 2.5 NOD filter.

Our experience is that a threshold of 2 to 3 NODs works well to eliminate data that intuitively appears to outlying. The filtering procedure may be iterated several times until outliers are no longer found in any term in the data set. Because the mean absolute deviation will be appropriately larger for multi-modal distributions, data in these modes will not be screened out if enough observations exist in the mode.

EXPLORE

Exploration or analysis of the data once obtained should be performed such that significant relationships to a particular issue are not missed. JMP®, of course has many built in data analysis techniques that can be used to reveal relationships in the data set. We however, have developed additional techniques using JMP® scripting that find the most statistically significant relationships to any issue in the data set in a semi-automated fashion. One such script was developed to perform a linear or polynomial least squares fit of all numeric parameters to any other particular parameter. Graphs are then displayed in order of significance for all relationships found above a specified threshold. We find that it is important to see the relationships graphically and not just rely on the statistics.

Whereas the statistics for correlation work best for normal continuous parameters we also sometimes use it to correlate to yield. Since yield is generally a highly non-normal distribution we sometimes find it best to transform the yield distribution using the logit function, defined as log[y/(1-y)] where the natural logarithm is used and y is yield. This function equals plus and minus infinity for yields of identically 1 and 0, but a work-around can be used to redefine those to a maximum and minimum value, respectively. Figure 5 shows a common example where most yields are well behaved but a deviation in the process has caused some very poor yielding product. The transformed data is much closer to a normal distribution. The converted values may take a little getting used to.

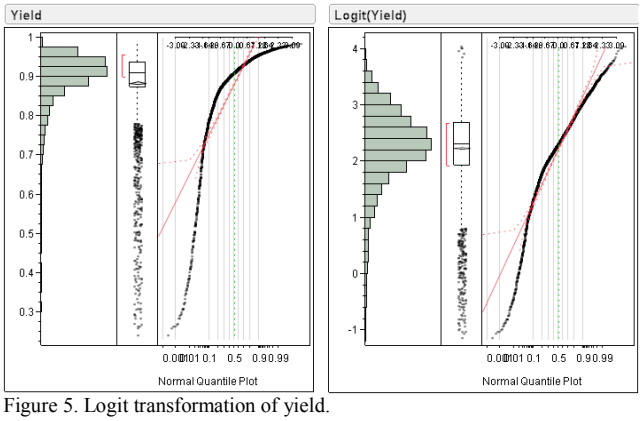
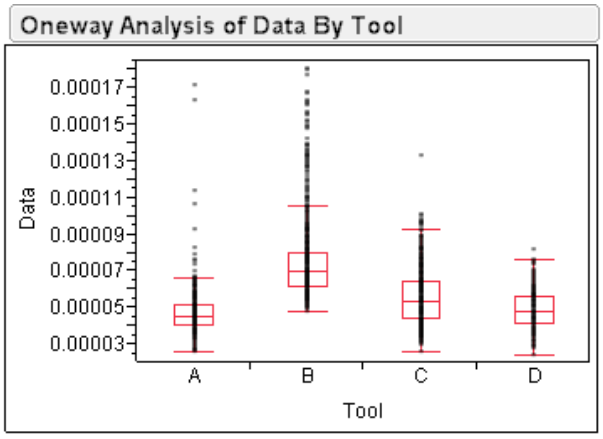


Figure 5. Logit transformation of yield.

Another script we have developed performs one-way analysis for a given numerical parameter on all categorical parameters. The user has an option of choosing the analysis type (e.g. ANOVA, Rank sum test). We often use Kruskal-Wallis rank sum test which is more robust to non-normal distributions. Also, since the data are converted to the rank in the population, values which are well outside the bulk of the data are de-weighted. The statistics of the Kruskal-Wallis analysis will not be shown here but can be found at various sites on the web and in the original work for those so inclined [2]. Chi squared is determined based on the mean rank for each level compared to an expected mean rank if the ranks were random by level. Only graphs of the data that show probability of non-random rankings are plotted and those are ordered by significance. This is very convenient way of determining, for example, whether particular process tools are leading to parametric or yield variation.



Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Score Mean	(Mean-Mean0)/Std0
A	1034	1161448	1123.26	-30.363
B	1548	4957231	3202.35	42.481
C	1404	2610865	1859.59	-10.314
D	289	410406	1420.09	-10.240

1-way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
2019.3408	3	0.0000*

Figure 6. Sample output from JMP® using the Kruskal-Wallis test.

Some of these types of analyses are performed on the data daily with scheduled scripts to make evident trends and correlation that need to be recognized.

VALIDATE

At this point in the process it is not uncommon to have revealed one or more important relationships concerning the issue being studied. If applicable, steps to remediate an issue may be possible to implement immediately (e.g. shutting down a poorly performing tool). The production results themselves may provide validation. In any event, it

is likely, at least, that the path towards a true solution has now become much clearer. Even so, the knowledge gained is one of correlations and statistical significance. The work to prove causation must still be done. But now, some hypotheses can be eliminated and the chance of testing the true hypothesis is increased.

EXAMPLE

A brief example will be presented showing the value of these techniques and the advantage of the vertical integration in RFMD. For an important new product, yields were varying widely from lot-to-lot due to limited capability to achieve customer linearity specifications. Using the UDQ all available data from package test to substrate were assembled and grouped by package test lot and wafer. Routines to find the best relationships between package linearity and all wafer and die related data resulted in several revelations. Among them, a parameter which I will call Xquotient, was found to correlate well with linearity. Particular MBE reactors were shown to be associated with better linearity yields and lower Xquotient distributions.

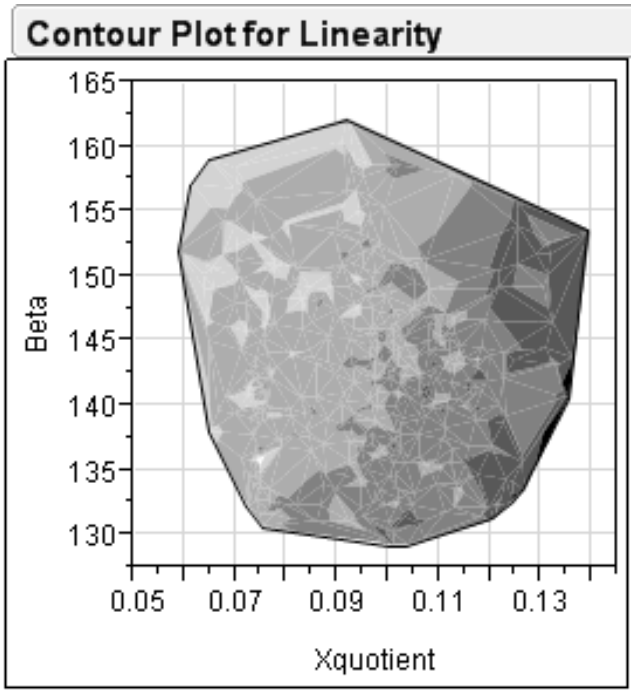


Figure 7. Contour graph of linearity ranging from good (light) to poor (dark).

An immediate containment action was obvious. Use only wafers from the 'best' reactors for this product. By doing this and applying specifications based on the revealed relationships, scrap rate for this product was reduced by about a factor of three. In addition, through further investigation of the MBE process, the variables that affect Xquotient have become better understood and this has eventually resulted in higher quality, more consistent epitaxial wafers from tool-to-tool and run-to-run.

CONCLUSION

By efficiently uncovering knowledge from the data and then guiding and tracking projects using a DMAIC approach, many true improvements have been obtained in our manufacturing process that have led to better sustained yields. The result of using such a process in the RFMD fab is illustrated in the temporal chart (Figure 8) of composite Known Good Die (KGD) scrap rate of all technologies.

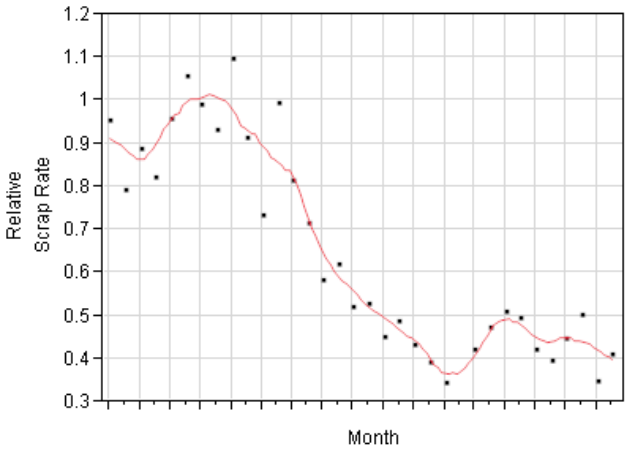


Figure 8. Month by month relative scrap rate for all product die tested.

ACKNOWLEDGEMENTS

The author would like to thank the wafer fab staff at RFMD for their support and contributions to yield improvement. Special thanks to Yik Yang, Dain Miller and Nathan Conrad for contributions to the Universal Data Query.

REFERENCES

[1] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth. "From Data Mining to Knowledge Discovery in Databases", AI Magazine, Vol 17, No 3, 1996, p 37.

[2] W. Kruskal and W. Wallis "Use of ranks in one-criterion variance analysis", Journal of the American Statistical Association Vol 47 No 260, 1952, p 583

ACRONYMS

KDD: Knowledge Discovery in Databases,
Knowledge Discovery and Data Mining
UDQ: Universal Data Query
NOD: Normalized Ordered Differences
DMAIC: Define, Measure, Analyze, Improve, Control
KGD: Known Good Die Test