

Advanced Statistical Tools for Improving Yield and Reliability

Richard Kittler
Yield Dynamics, Inc.
Santa Clara, California

Contact Information:

Richard Kittler, Yield Dynamics, Inc., 2855 Kifer Road, Santa Clara, 95051, USA
Phone: 408.330.9320 x105, Fax: 408.330.9326, E-mail: rich@ydyn.com

Keywords: software statistics databases testing traceability yield

Significance: Failure analysts will be introduced to data mining and its relationship to other statistical techniques for improving yield and reliability.

Abstract:

Analysis of manufacturing data as a tool for failure analysts often meets with roadblocks due to the complex non-linear behaviors of the relationships between failure rates and explanatory variables drawn from process history. The current work describes how the use of a comprehensive engineering database and data mining technology overcomes some of these difficulties and enables new classes of problems to be solved. The characteristics of the database design necessary for adequate data coverage and unit traceability are discussed. Data mining technology is explained and contrasted with traditional statistical approaches as well as those of expert systems, neural nets, and signature analysis. Data mining is applied to a number of common problem scenarios. Finally, future trends in data mining technology relevant to failure analysis are discussed.

The current work describes how the use of a comprehensive engineering database and data mining technology overcomes some of these difficulties and enables new classes of problems to be solved. The characteristics of the database design necessary for adequate data coverage and unit traceability will be discussed. Data mining technology will be explained and contrasted with traditional statistical approaches as well as those of expert systems and neural nets. The relationships to signature analysis and classification trees will be explored. A number of common problem scenarios will be discussed.

The outlook for commercial availability of such capabilities will be surveyed and the system integration costs associated with their installation will be projected. Finally, the future of these new data mining and database technologies will be discussed as it applies to further improving the productivity of failure analysts.

Introduction:

The field of failure analysis plays a critical role in all areas of the yield ramp on a new technology or a new product introduction. With each succeeding generation process limits are being pushed into areas where it is increasingly difficult to forecast reliability and trace failures back to root cause [1]. Use of ad hoc statistical methods has often met with roadblocks due the complex non-linear behaviors of the relationships between failure rates and explanatory variables drawn from process history. Customer returns or failures during qualification typically involve small sample sizes from which little can be gleaned related to root cause.

Current Approaches and Limitations

Use of data analysis as a tool for failure analysis relies on ready access to broad classes of data as well as powerful analytical tools. Today most companies have islands of data. This data is a mixture of tool-based databases and file systems created by the analytical tools provided by suppliers, as well as in-house efforts at consolidating data from varied sources. The in-house efforts have evolved based on the scope of the analyses that were deemed necessary during succeeding generations of technology. While customers would like all data to be easily available to the analysis tools of their choosing,

suppliers are often most interested in keeping the data their tools generate accessible only by the add-on analytical packages they offer. There are as of yet very few standards for the interchange of data in defined formats even though the low-level SECS protocols have matured. Because these tools have evolved slowly it is rare that the database capabilities have kept up with analytical needs in an efficient fashion. For example, few databases support traceability of lot numbers across facilities and access to associated data in a transparent fashion. These limitations make it difficult to correlate final test or reliability results to fab tool or process data except via tedious manual methods.

Data analysis tools range from those provided on the analytical tools to ad hoc environments in third party packages such as MS/Excel, SAS, and SPLUS. In the interest of creating analysis capabilities that span categories of data, e.g. fab, sort, assembly, and final test, many larger semiconductor companies have created in-house applications written on top of some of these third party tools. Recently a base of independent software suppliers has begun to emerge to fill the need for powerful, flexible analysis capabilities that span this range of data. This contrasts with the ongoing evolution of tool-based analysis capabilities driven by the hardware suppliers. Due to the costs of developing analysis capabilities internally, and their reliance on the availability of data, such tools rarely extend beyond ad hoc manual charting, regressions, and ANOVA. These limitations make it difficult to search for relationships in data unless they are already anticipated, or to find relationships which involve interactions between factors, e.g. process tools at different steps.

Supplier-based offerings for data storage and analysis would have evolved faster relative to internal efforts at major companies had there been more well-defined industry standards for database design, data access methods, or analytical tool capabilities. Larger companies have been driven to create such standards internally while smaller companies have relied on piecemeal internal efforts coupled with what they perceived as the most flexible and affordable supplier based tools.

Extending Database Designs

Data from a broad array of data sources is needed to support root cause identification for reliability failures or customer returns. For analysis of reliability failures the database should

store and index data from the reliability lab back through final test, assembly, wafer sort, and fab. In addition, for customer returns, the database needs to also be capable of mapping a package mark code to one or more backend lot numbers. Given a set of lot numbers from which the part(s) originated, the system needs to know the process flows those lots followed and be able to retrieve key data for all or some of the steps within these flows. For each lot and processing step, this should include the begin- and end-processing timestamps, equipment used, and any data collected. If lot numbers change due to use of sub-lots in the back-end then the database needs to maintain traceability over the split process. In the ideal case, the reliability or failure analysis engineer could use the system to retrieve and analyze all lot history and engineering data to establish a root cause. If traced to a processing excursion, they could then determine what other lots might be in jeopardy, trace these lots forward to customer orders, and notify other effected customers.

Today very few companies have such comprehensive data environments in place with the applications capable of supporting these analyses.

Use of Data Mining

Once the data is brought together through whatever means are necessary, the analysis tools must then be capable of finding patterns in the data to explain observed behaviors. The process of finding hidden patterns in data to help explain the behavior of one or more response variables is called data mining [2-4]. It differs from other methods of modeling in that there is no preconceived model to be tested. Rather, a model needs to be found using a pre-set range of explanatory variables. In general these variables may have a variety of different data-types and include outliers and missing data. Some of the variables to be included in the model may be highly correlated and the underlying relationships may be non-linear and include interactions.

In all cases we are seeking methods to automatically search for structure in the data. Without such capabilities the engineer is doomed to viewing endless lists of trend charts and scatter plots to look for relationships. Success in this process relies on being able to page quickly through the plots, having sufficient time to view all of the charts, and maintaining lots of patience. With luck a relationship is found, it is shared with

the cognizant process engineers, and evaluated as either root cause or a dead end.

Defined as above in this broad fashion, data mining can be performed using a number of analytical techniques. Some of these techniques involve the use of traditional statistics, while others employ more exotic techniques such as neural nets, association rules, and decision trees. We'll look at each of these techniques individually. In all cases we are seeking to establish a model to explain the variation of a response across a set of observations together with the ability to accurately predict the response for new data we may encounter.

Statistical Methods employ such tried and true techniques as ANOVA, regressions, and contingency analyses. In using these techniques you must have a model or class of models in mind and create programs that evaluate each model across one or more ranges of the explanatory variables. The success of such efforts depends on picking the right class of models to fit. For instance, if the true behavior is quadratic and you develop an automated routine to search for a linear model then the fit and predictive power could be poor.

Neural nets offer the opportunity to create a model using technology similar to the learning patterns of the human brain. A neural net model consists of a network of nodes on input and output layers separated by one or more hidden layers. Each input and output layer node is associated with a variable in the dataset. Nodes have connections to all nodes in adjacent layers. Response functions on each hidden layer node determine how a signal from the input direction is propagated to nodes in the output direction. The network is trained by adjusting the weights on the hidden layer nodes to minimize the error in the outputs across a set of training data. The trained network can then be used to make predictions for new data. This method of use is called supervised learning. It has been shown that under certain conditions a model with a single hidden layer is equivalent to multiple linear regression. Neural net models are useful when large amounts of data need to be modeled and a physical model is not known well enough to use statistical methods. One of the downsides of the approach is that it is difficult to make a physical interpretation of the model parameters. Also, the predicted outcomes of the model are limited to the scope of the training set used. In that sense it is not able to discover any new relationships in the data and hence is not really data mining.

Association Rules is a tool used to look for patterns of coincidence in data. For instance, how often do failing lots go through various combinations of deposition and etcher combinations. In its simplest form this is the same as a contingency table. In more advanced forms the order with which events occur is also taken into account. For example, how often do failing lots go through two reworks at first metal and are then etched on a certain etcher. Association rule analysis is useful in discovering patterns of behavior but does not produce a predictive model.

Decision Trees are analytical tools for developing hierarchical models of behavior. The tree is built by iteratively evaluating which variable explains the most variability of the response based on the best rule involving the variable. Classes of rules include linear models, binary partitions, and classification groups. Other classes of models could also be considered. A binary partition of a continuous variable would be of the form: 'If e-test variable $T1234 < 1.23e-5$ '. Such a rule would partition the data into two groups, one group for which the rule is true and another for which it is false. The root node of the tree involves a rule that explains the most variability of the response. The child nodes are built by finding other variables that explain the most variability of the data subsetted by the first node, etc. The process stops when a point of diminishing returns is reached in a manner similar to automated step-wise regression procedures. Decision trees are useful when the relationships are not known and when you need to make broad categorical classifications or predictions. They are less useful when you need to make precise predictions for a continuous variable.

Overall, decision trees offer the most power in detecting hidden trends in data, in being most physical, and in offering the predictive capability needed to understand patterns of behavior. For this reason we proceed further in the current discussion using decision trees as the reference data mining method.

Relation to Knowledge-Based Systems

We will now look at the relationship between decision trees and knowledge-based systems such as expert systems, signature analysis, and classification trees.

Expert systems provide the capability to create hierarchical knowledge systems given a set of rules. The systems generally guide a user

through a decision making or diagnostic process. Data mining using decision trees provides another method of deriving rules upon which expert systems are built.

Signature analysis is a form of expert system that is specifically designed to assimilate clues associated with diagnostic data to fingerprint a process failure. Here again, decision trees can help to discover patterns that associate a given failure with a set of process conditions. Once associations are known they can be applied to new data through signature analysis to implicate likely process conditions that led to the failure. In a similar manner, historical data could be used to derive rules for suggesting next steps in a failure analysis given the results obtained thus far.

Classification trees are a special case of decision trees when the response variable is categorical. They can be built with or without the use of data mining technology if the knowledge can be obtained through other means.

Data Mining Applications

In this section we will explore how data mining using decision trees and the use of a comprehensive engineering database can be used to help solve several yield and failure analysis problems. In each case availability of a comprehensive engineering database and advanced analytical tools, such as data mining using decision trees, is crucial to the solution of a complex yield or reliability problem.

~~///~~ Scenario #1: a customer has returned some parts from an order due to early failure. The failure is traced to a process problem at one of the diffusion steps based on signature analysis using rules derived from a decision tree analysis of historical data. The engineering database is then used to:

1. trace the mark date code on the failed parts to a single fab lot number,
2. provide the date-time that this lot went through the suspect diffusion step and the specific furnace and tube that was used,
3. interrogate the tool event history to discover that a problem was found and fixed 2 days after that lot went through the tube,
4. provide the list of lots that went through the same tube after the failed lot but before the fix was put in place,
5. trace this list of lots forward to mark date codes and other customer orders,

This information was then communicated to the product line who drafted a communication to the other effected customers notifying them of a possible problem with their order and suggesting a recall of the effected parts.

~~///~~ Scenario #2: a rise in the number of early life failures across several products on the same fab technology prompts the reliability engineer to do a commonality analysis of the processing history of the failed parts. The first steps in the analysis are to use the engineering database to:

1. look up the fab lot numbers associated with the all of the parts tested on this same technology in the last 3 weeks,
2. retrieve the fab lot history and in-line data for this set of fab lots

Once this data is available data mining is performed on the dataset to discover that the rate of failure on the lots is tied to use of a specific deposition tool at either of two metalization layers within a window of process dates. The decision tree results are supported by trend charts of the failure rates versus the date through the deposition system. The decision tree results and the supporting charts were then shared with the fab engineering group for follow-up. The fab engineers use the engineering database to investigate events and comments logged to the tool history during the suspect time period in an attempt to discover clues for the window of failures. After retrieving the data and performing a brief review, they begin the analysis by using data mining to compare the tool history of the failed tool to that of other similar tools used at the same time.

~~///~~ Scenario #3: yield through a bond strength test is beginning to drop. The responsible engineer uses the engineering database to retrieve lot history information through assembly for all of the lots tested in the last three weeks. This includes material batch numbers and subcontractors used at various steps. Data mining is then used to derive a model relating the measured bond strength results on each lot to this lot history information. The decision tree shows that one of the sub-contractors is the culprit but that they are only having a problem with one of the package types. As part of an integrated analysis environment, the engineer plots out the supporting box plots and trend charts that allow the relationships shown in

the decision tree to be more clearly visualized. The data and charts are then communicated to management and the sub-contractor is contacted as the first step in fixing the problem.

Commercial Availability

Many companies have built portions of the capabilities described above. To be effective the engineering database must be comprehensive and integrated tightly with the engineering analysis environment. It should be able to be distributed across separate servers and physical locations. The engineering analysis environment must offer a full suite of graphics and statistics tools including such advanced technologies such as data mining. It should integrate with the engineering database and offer the user the capability of transparently retrieving data regardless of its location. It should also offer the capability of integrating with legacy systems as if they were part of the main database environment.

A wide variety of standalone data mining tools are available on the commercial market. Some are general-purpose tools and others are tailored to use in specific industries such as finance or retail. To be truly effective as a tool to improve yield and reliability, such software technology needs to be available as part of an integrated yield management software environment. Otherwise, the engineer will be faced with importing and exporting data from one tool to another depending on where they are in the analysis cycle. To achieve this ease of use, data mining needs to be available as part of a commercially acquired yield management system, or it needs to be integrated in a seem-less fashion with an in-house system. The latter case requires both a fully functional programmatic interface as well as an adequate internal staff to set-up and maintain the integration.

Today there are few commercial products that provide both the breadth of database capabilities and the advanced data analysis features we have described, however such product offerings are beginning to emerge. Among others, Yield Dynamics has recently introduced their Genesis yield management software that incorporates a proprietary data mining module within their full product suite. In evaluating this and other offerings, care should be taken to insure that purchased products can be integrated with existing systems and there is flexibility in the hardware that can be used. Ideally, a supplier will provide a product that can be installed in such a

manner so as to complement existing capabilities and provide a path for a phased migration to more complete use. Preference should also be given to those solutions that allow you to leverage hardware and software systems already in place.

Another component of the cost for new systems capabilities is that of setup and integration. Setup includes installation and configuration on the target hardware platform. Integration involves interfacing to existing data sources. Integration costs can be large if the legacy databases are varied and complex. Availability of a high-level data integration language is desirable for maximum productivity in this effort as well as lower future maintenance costs.

Future Trends

Future work in the development of data mining software will extend options for visualization, use of alternate algorithms, and efficiency. Data mining tools will become more prevalent in yield management software. In addition, the leading yield management companies will continue to integrate in new statistical technologies and intelligent data access routines to support automated searches for more complex patterns of process variation. Scripting languages will be extended to support more sophisticated monitoring capabilities. Database offerings will provide more extensive schemas for the efficient storage and retrieval of new forms of data and images. Pre-computed summary information within the database will be more comprehensive to make certain forms of analysis both easier and more efficient. Some standards will begin to evolve driven by SEMATECH and other consortia that seek to promote open plug-and-play type architectures for yield management software. As a result of these trends, the tools available for yield and reliability analysis will improve further in capability and efficiency.

Summary and Conclusions

Today's systems for the analysis of yield and reliability are emerging from an embryonic phase. Islands of data and analysis capability are growing together through the efforts of in-house developers and emerging third party suppliers. Driving forces of more complex processes and programs that span larger regions of the process make such efforts a requirement for competitive I.C. manufacturing. Key to the development of integrated capabilities to meet these needs will be the use of a comprehensive engineering

database, intelligent data access routines, and advanced statistical technology such as data mining.

Acknowledgements:

Special thanks to Weidong Wang and Jon Buckheit of Yield Dynamics, Inc. for their careful reading of this manuscript and to Nick Atchison of Silicon Systems, Inc. for his encouragement on the topic.

References:

- 1) W. Maly, "Testing-Based Failure Analysis: A Critical Component of the SIA Roadmap Vision", *Proc. of ISTFA 1996*, pp. 3-6 (1996)
- 2) C. Westphal and T. Blaxton, *Data Mining Solutions*, John Wiley and Sons, New York (1998)
- 3) J. Friedman, "Data Mining and Statistics: What's the Connection", www-stat.stanford.edu/reports/friedman (1997)
- 4) G. Shapiro *et. al.*, "An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications", *Proceedings of KDD-96*, pp. 89-95 (1996)