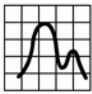


StatSoft®

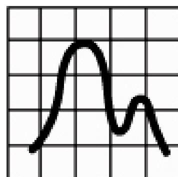
data analysis • data mining • quality control • web-based analytics

Semiconductor Industry

and

STATISTICA

Case Study: Silicon Wafer Manufacturing



STATISTICA
Solutions for Business Intelligence, Data Mining, Quality Control, and Web-based Analytics

U.S. Headquarters: StatSoft, Inc. • 2300 E. 14th St. • Tulsa, OK 74104 • USA • (918) 749-1119 • Fax: (918) 749-2217 • info@statsoft.com • www.statsoft.com

Australia: StatSoft Pacific Pty Ltd.
Brazil: StatSoft South America
Bulgaria: StatSoft Bulgaria Ltd.
Czech Rep.: StatSoft Czech Rep. s.r.o
China: StatSoft China
France: StatSoft France
Germany: StatSoft GmbH
Hungary: StatSoft Hungary Ltd.
India: StatSoft India Pvt. Ltd.
Israel: StatSoft Israel Ltd.
Italy: StatSoft Italia srl
Japan: StatSoft Japan Inc.
Korea: StatSoft Korea
Netherlands: StatSoft Benelux BV
Norway: StatSoft Norway AS
Poland: StatSoft Polska Sp. z.o.o.
Portugal: StatSoft Ibérica Lda
Russia: StatSoft Russia
Spain: StatSoft Ibérica Lda
S. Africa: StatSoft S. Africa (Pty) Ltd.
Sweden: StatSoft Scandinavia AB
Taiwan: StatSoft Taiwan
UK: StatSoft Ltd.

Table of Contents

SEMICONDUCTOR INDUSTRY AND *STATISTICA* **1**
CASE STUDY: SILICON WAFER MANUFACTURING **1**
Case Description 1
Understanding the Manufacturing Process 1
Problem Definition 2
DATA ANALYSIS WITH *STATISTICA* **3**
Feature Selection 3
Mean Plot 4
Scatterplot 4
Box Plot 5
Scatterplot 5
Classification Trees 5
Interactive C&RT/CHAID Algorithm 6
Result - *Interactive C&RT* 7
Result - *Interactive CHAID* 8
Other Applications of Data Mining 8
CONCLUSION **9**
References 9

Semiconductor Industry and *STATISTICA*

Data mining algorithms have played a crucial and successful role in a wide spectrum of advanced manufacturing processes. Yield management and process control in wafer manufacturing are other emerging and interesting areas where data mining methodologies find useful applications. The numerous steps and complex workflows during wafer manufacturing automatically generate large volumes of data and, hence, data mining technology is becoming increasingly important in semiconductor manufacturing.

The sophistication and complexities involved in chip manufacturing have always delayed the dream of creating a foolproof process to produce 100% yield. Although these manufacturing recipes (typically consisting of combinations of tools used in 300 – 500 steps) are carefully designed and revised to maximize yield, yield is still affected by errors that are inevitably introduced by systematic factors (e.g., defective tools or interactions between tools) as well as random factors (e.g., dust particles).

STATISTICA provides a suite of flexible analytic tools that can be used for different applications (e.g. for root-cause analysis, to identify systematic factors such as defective tools or combinations of tools that cause low yield problems). Predictive models, such as the ones described in this paper (*Stochastic Gradient Boosted Trees*, *Interactive C&RT*, *CHAID* algorithms, etc.) can be used to analyze the measurements recorded during the production process to identify the factors and their interaction that adversely affect yield and overall product quality.

Case Study: Silicon Wafer Manufacturing

Case Description

This case study provides an example of one of the various useful applications of data mining in the field of silicon wafer manufacturing. Note that this example illustrates only a small portion of the comprehensive capabilities of the *STATISTICA* toolkit.

Understanding the Manufacturing Process

Wafer manufacturing starts with a slice of crystal silicon that is subjected to a series of manufacturing steps (between 300 and 500), to yield microprocessor chips at the final stage. A batch of wafers (around 24 cut from a slice of silicon) is called a Lot. At the final stage, chips on the wafers are separated with a diamond saw to form individual integrated circuits. In between these elaborate processes, around 1,500 to 5,000 measurements are made on each chip.

Data File

The data file *wafer_yield.sta* consists of "Lot level" data from a wafer manufacturing fabrication. This data file contains 2,858 variables and 2,062 cases. Most of the variables in the data file contain "Tools" and "Log time" information pertaining to 1,283 pieces of equipment, used at different steps ("Log-points") of the production process. There are also a few other quality measurements (around 40 of them), of which the most important dependent variable is the MULTIPROBE yield, the final yield measure for a lot (or percentage of functioning chips). A new variable LowHiYield was calculated to categorize the final yield measurement as High/Low using the median value of 59.025 (calculated from MULTIPROBE yield).

Variable Information

• Predictors: EQUIPMENT - 1,406 variables; LOG TIME – 1,406 variables; OTHERS – 46 variables

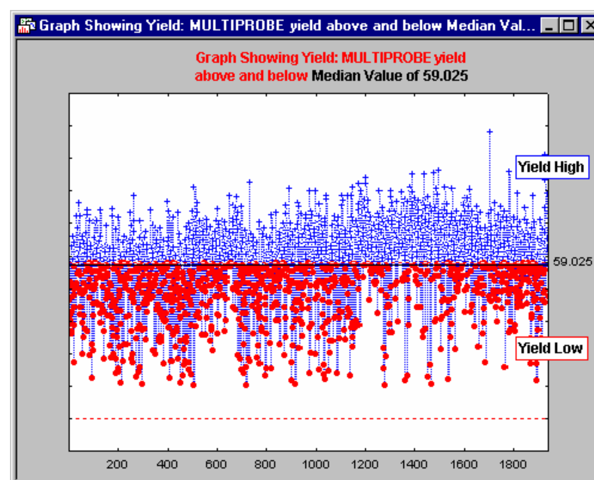
• Outcome / Dependent Variable (or Variable of interest) – MULTIPROBE yield

• New Dependent Variable – LowHiYield (based on median value - of 59.025)

Problem Definition

The problem can be defined by two goals: 1) Identification of tools and interaction between tools that causes low yield; and 2) Identification of the log-points when low yield problems exist (i.e., to examine any confounding or correlation between specific tools used at particular times).

STATISTICA includes a comprehensive selection of graphical methods for both data analysis and presentation of results. The following illustration is a visual presentation of MULTIPROBE yield observations.



Quality Control Chart

Such simplified quality control charts can be used to visually observe the distribution of MULTIPROBE yield falling above and below standard cutoff levels (in this case, yield was categorized as high or low based on a median value of 59.025). Any specific patterns of yield over consecutive lots can usually quickly be detected in this chart.

Data Analysis with *STATISTICA*

Feature Selection

This demonstration used the *STATISTICA* Feature Selection tool to identify the best predictors (in this case, equipment) that clearly discriminated between High/Low yield (dependent categorical) or cause resultant variability in MULTIPROBE yield (continuous dependent variable). The *Feature Selection* tool is extremely useful for reducing the dimensionality of analytic problem, i.e., to select the specific predictors (out of 1283 in this case) that are likely "candidates" causing yield problems. Thus, engineers can quickly focus their efforts on only the few specific types of equipment that are the likely root causes of yield problems.

Data: Best predictors for categorical dep...

	Best predictors for cate	
	Chi-square	p-value
Equipment NW00/5722	137.8842	0.000000
Equipment NW00/5922	137.5644	0.000000
Equipment NW00/5822	137.8248	0.000000
Equipment NW60/8751	84.1724	0.000000
Equipment NW30/9002	61.1891	0.000000
Equipment NW51/11223	37.8189	0.000001
Equipment NW80/11223	39.6548	0.000001
Equipment NW55/4672	31.9599	0.000002
Equipment NW80/8799	41.2171	0.000005
Equipment NW50/11223	31.9504	0.000006

LowHiYield (Categorical)

Data: Best predictors for continuous dep...

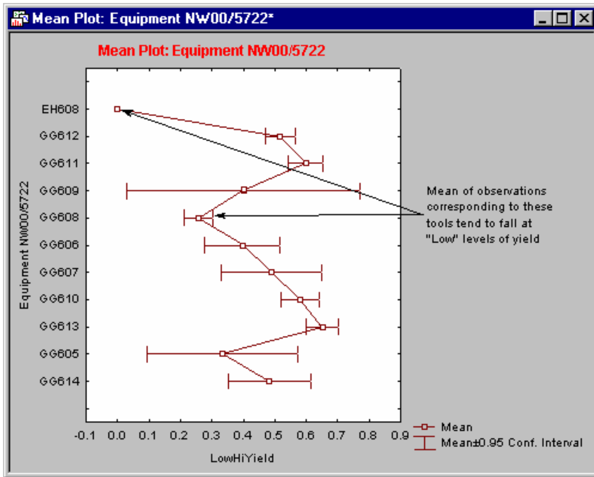
	Best predictors for con	
	F-value	p-value
Equipment NW00/5822	16.07817	0.000000
Equipment NW00/5722	14.53163	0.000000
Equipment NW00/5922	14.45699	0.000000
Equipment NW60/8751	12.57838	0.000000
Equipment NW30/9002	6.58303	0.000000
Equipment NW50/11223	7.91513	0.000000
Equipment NW80/8799	4.96985	0.000001
Equipment NW60/11223	5.64243	0.000002
Equipment NW80/11223	5.51984	0.000003
Equipment NW50/3210	4.84602	0.000006

MULTIPROBE yield (Continuous)

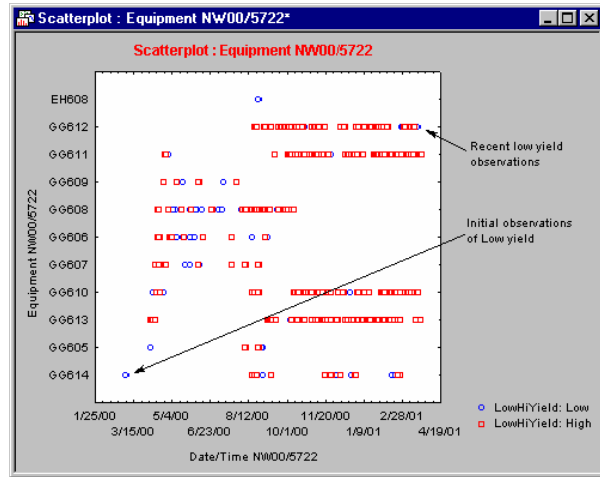
Feature selection identified the same eight pieces of equipment (highlighted in red) as best predictors for the dependent variable "MULTIPROBE yield" and the derived dependent variable "LowHiYield." This validates the use of categorical variable "LowHiYield" to better understand and define the problem.

Exploratory Analyses: The next logical step would be to drill down into the pieces of equipment that were identified as the best predictors to pinpoint specific tools within these pieces causing "Low" yield.

Dependent Variable (Categorical) – LowHiYield



Mean Plot



Scatterplot

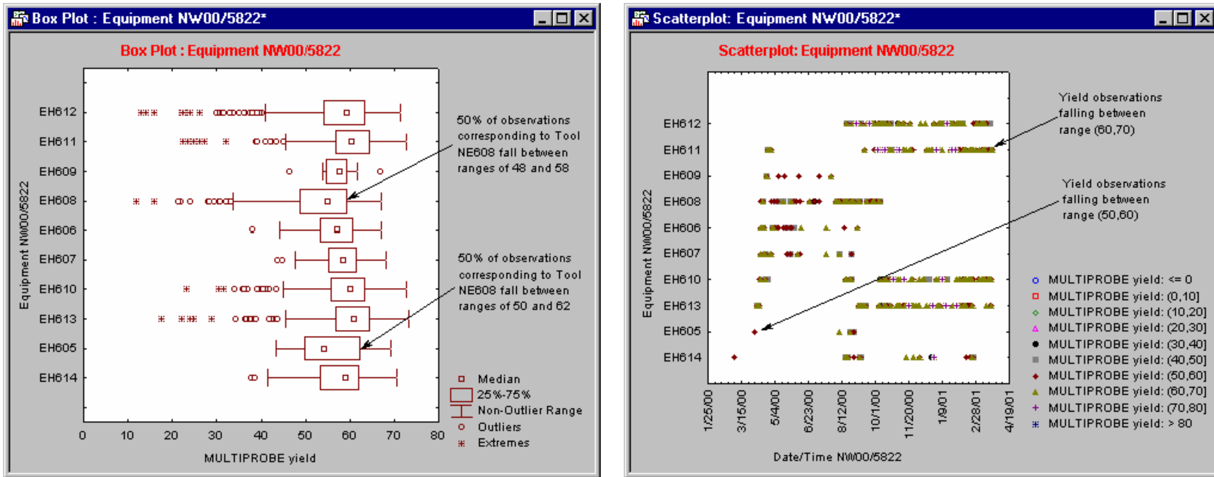
Mean Plot

This plot visualizes the means for the dependent variable (LowHiYield) broken down by the tools within the equipment NW00/5722 (the feature selection tool identified this equipment as Best predictor for LowHiYield). From this graph, we say that Low yield was associated with tools EH608 and GG608. Each mean marker is accompanied by a set of whiskers |——| representing error bars, which are the confidence intervals around the mean (hence, related to the variability observed for the respective tools). These plots are typically used to compare marginal means across groups and visualize the reliability of these means.

Scatterplot

These categorized graphs (in this case scatterplots) show High/Low yield observations corresponding to each category (different tools), for particular equipment. These "component" graphs are placed sequentially in one display, allowing for comparisons between the patterns of data shown in the graphs for each of the requested groups (e.g., tools). The categorized scatterplot shows which tools caused Low/High yield and at which log-times. Apparently, the tools identified by the Mean plot that caused Low yield problems (EH608 and GG608) were primarily used earlier in the time described in the data. The more recent problems seem to occur when tool GG612 is used.

Dependent Variable (Continuous) – MULTIPROBEyld



Box Plot **Scatter Plot**

Box Plot

In box plots, ranges of values are plotted separately for groups of cases defined by values of a categorical variable. The central tendency (e.g., median or mean) and range or variation statistics (e.g., quartiles, standard errors, or standard deviations) are computed for each group of observations, and the selected values are presented in the style specified in the *Graph Type*. Influential and outlying data points can also be plotted (see the Outliers and Extremes legend above).

The results for the untransformed dependent (outcome) variable MULTIPROBE yield selected another piece of equipment, NW00/5822, as the best predictor using *Feature Selection*. The box plot shows that the yield corresponding to tools EH608 and EH605 is low compared to other tools for the same equipment. The rectangle drawn around the median value indicates the range within which 50% of its yield observations (around the median) fall. Also note that many outliers were observed corresponding to most tools in use.

Scatterplot

In this case, each observation of MULTIPROBE yield was categorized into 10 different categories (representing different ranges of yield) to plot them against a time scale. The different symbols and colors (see legend) for different ranges were used to depict the yield distribution.

Classification Trees

Classification trees are used to predict membership of cases or objects into classes of a categorical dependent variable from their measurements on one or more predictor variables. Classification tree

analysis has traditionally been one of the main techniques used in data mining. The *Classification Trees* module in *STATISTICA Data Miner* is a full-featured implementation of techniques for computing binary classification trees based on univariate splits for categorical predictor variables, ordered predictor variables (measured on at least an ordinal scale), or a mix of both types of predictors. It also has options for computing classification trees based on linear combination splits for interval scale predictor variables.

The goal of classification trees is to predict or explain responses on a categorical dependent variable, and as such, the techniques in this module have much in common with the techniques used in the more traditional methods of *Discriminant Analysis*, *Cluster Analysis*, and *Logistic Regression*. The flexibility of classification trees makes it a very attractive analysis option, but this is not to say that its use is recommended to the exclusion of other methods. As an exploratory technique, classification trees are, in the opinion of many researchers, unsurpassed.

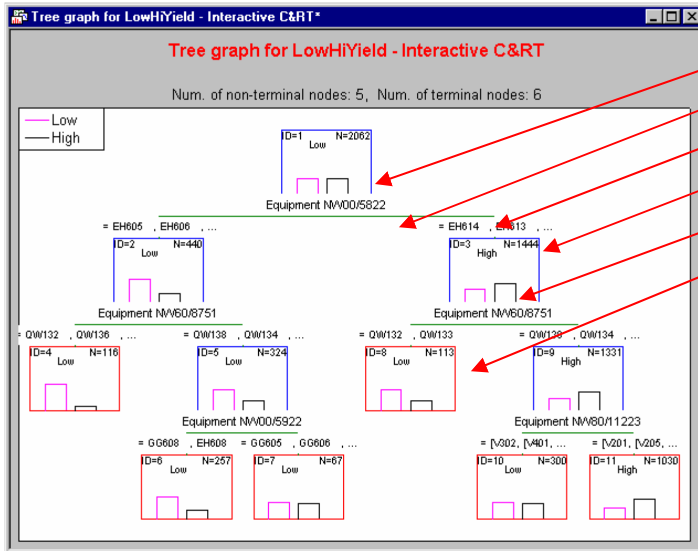
The study and use of classification trees are not widespread in the fields of probability and statistical pattern recognition (Ripley, 1996), but classification trees are widely used in applied fields as diverse as manufacturing (as in this case), medicine (diagnosis), computer science (data structures), botany (classification), and psychology (decision theory). Classification trees readily lend themselves to being displayed graphically, helping to make them easier to interpret than they would be if only a strict numerical interpretation were possible.

Interactive C&RT/CHAID Algorithm

The *STATISTICA Interactive Trees (C&RT, CHAID)* module builds ("grows") classification and regression trees as well as *CHAID* trees based on automatic (algorithmic) methods, user-defined rules and criteria specified via a highly interactive graphical user interface (brushing tools), or combinations of both. The purpose of the module is to provide a highly interactive environment for building classification or regression trees (via classic *C&RT* methods or *CHAID*) to enable users to try various predictors and split criteria in combination with almost all functionality for automatic tree building.

Note: The capability of *Interactive C&RT* and *CHAID* algorithms to handle missing data is one reason to use these modules for exploratory analysis.

Result - *Interactive C&RT*



Interpreting these trees is quite straightforward. The *C&RT* algorithm identified interactions between equipment 43NW/3600, 746NW/6529, and 428NW/2225 to identify specific or combinations of tools that led to “High” or “Low” Yield.

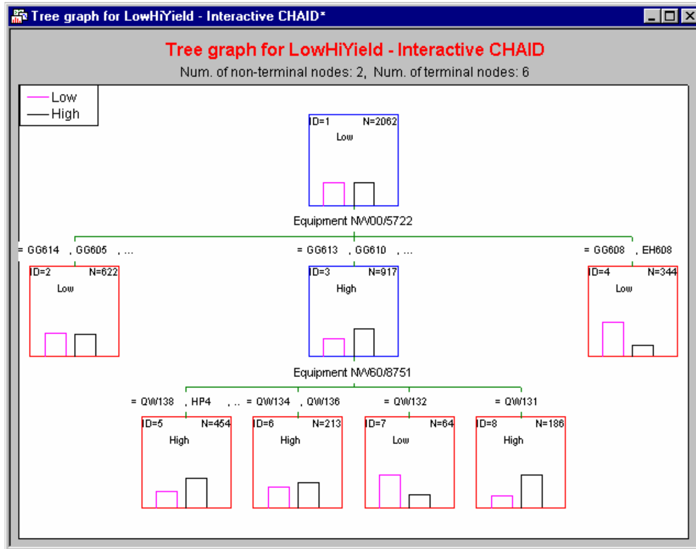
The rules generated by these trees (also available from a tree structure table) can help engineers to pin-point equipment (or specific/combination of tools) that cause “Low-yield” problems. As you can see from the above graph, the *C&RT* algorithm had distinguished 5 decision outcomes (contained in 6 terminal nodes highlighted in red) built on 5 “if then” conditions to predict the category of yield. By following the path from the root node (ID=1) to terminal node (ID=4), we can derive a rule for “Low” yield. If tools NE 605, NE 606, etc., related to equipment 43NW/3600 are used with tools NT132, NT136, etc. of equipment 746NW/6529, then there were 116 cases (observations) recorded out of which most of them fall into the “Low” yield category. Similarly, we can analyze the other branches and draw further conclusions. The legend identifying which bars in the node histograms correspond to the two categories of yield is located in the top-left corner of the graph.

Subsequent analyses with other data mining algorithms lend further support for these results.

U.S. Headquarters: StatSoft, Inc. • 2300 E. 14th St. • Tulsa, OK 74104 • USA • (918) 749-1119 • Fax: (918) 749-2217 • info@statsoft.com • www.statsoft.com

Australia: StatSoft Pacific Pty Ltd.
Brazil: StatSoft South America
Bulgaria: StatSoft Bulgaria Ltd.
Czech Rep.: StatSoft Czech Rep. s.r.o
China: StatSoft China
France: StatSoft France
Germany: StatSoft GmbH
Hungary: StatSoft Hungary Ltd.
India: StatSoft India Pvt. Ltd.
Israel: StatSoft Israel Ltd.
Italy: StatSoft Italia srl
Japan: StatSoft Japan Inc.
Korea: StatSoft Korea
Netherlands: StatSoft Benelux BV
Norway: StatSoft Norway AS
Poland: StatSoft Polska Sp. z.o.o.
Portugal: StatSoft Ibérica Lda
Russia: StatSoft Russia
Spain: StatSoft Ibérica Lda
S. Africa: StatSoft S. Africa (Pty) Ltd.
Sweden: StatSoft Scandinavia AB
Taiwan: StatSoft Taiwan
UK: StatSoft Ltd.

Result - *Interactive CHAID*



The *CHAID* algorithm identified interaction between NW00/5722 and NW60/8751, which is consistent with those from the initial *Feature Selection* analysis. As you can see from the above graph, the *CHAID* algorithm distinguished 6 decision outcomes (contained in 6 terminal nodes highlighted in red) built on 6 if-then conditions to predict the category of yield. One can examine the splits in this classification tree exactly as was done with the *C&RT* decision tree. For instance, in the first split, the *CHAID* algorithm identified that the use of specific tools (GC608 and EC608) for equipment NW00/5722 adversely affected yield (observations contained in the terminal node for this decision rule show more Low yield observations). However, *CHAID* results are sometimes more sensitive to particular data patterns, and the *C&RT* results should be examined and verified first.

Other Applications of Data Mining

Data mining techniques can also be used for several other applications in this industry:

1. **Automated methods that can identify and classify defective clusters of memory chips**. (Spatial Signature Analysis) - Quality control in the semiconductor industry has been traditionally based on overall summary data (for instance by measuring the ratio of good chips to the total number of chips produced). The use of these aggregate measures would be acceptable if the defective chips are randomly distributed across the surface of the wafers and across the wafers in a lot. In reality, defects occur in clusters, or exhibit systematic patterns that can be used to trace the factors that cause these problems. The identification and classification of defective chips can be fully automated by using advanced clustering techniques and other data mining algorithms instead of manually inspecting each piece, which is the current practice prevalent in the industry. This can help production units to cut down on inspection time and cost, with more reliable results (defect classification).

2. **Roll-off during wafer polishing:** One of the final stages in chip manufacturing involves polishing the wafers before turning them into integrated circuits. During this step, it's often seen that the wafer's edge becomes rounded rather than flat resulting in so called "roll-off." The roll-off (or rounded edges) reduces the area for creation of chips, which is highly undesirable. Parametric factors (such as temperature, pressure, type of polishing pads, etc.) causing this problem can be easily detected by data mining techniques, and feedback from advanced predictive models can be automatically linked to a control system to adjust the settings before quality starts deteriorating.

Conclusion

The example(s) described here provide a "glimpse" at the many possible solutions that can be achieved by using *STATISTICA Data Miner* software. The unique analytic and graphics capabilities and features in *STATISTICA* can be applied to solve a variety of manufacturing problems, and in particular, to address yield issues as discussed in this example. Advanced data mining tools are extremely useful to augment the effectiveness of existing technologies and processes to achieve quicker yield ramps and higher yield, to optimize capacity and productivity levels.

References

White, K., Mastrangelo, C., Modeling, Analysis, and Information Technologies for Semiconductor Manufacturing, from http://www.sys.virginia.edu/research/semi.asp

Kusiak, A., Decomposition in Data Mining: An Industrial Case, from http://www.icaen.uiowa.edu/~ankusiak/Journal-papers/Decomp.pdf