CLASSIFICATION AND PREDICTION OF WAFER PROBE YIELD IN DRAM MANUFACTURING USING MAHALANOBIS-TAGUCHI SYSTEM AND NEURAL NETWORK

C.C. Wang[1*] & B.D. Wu[1]

ARTICLE INFO

Article details
Submitted by authors 1 Sep 2016
Accepted for publication 10 Apr 2019
Available online 29 May 2019

Contact details
* Corresponding author
ieccwang@mail.mcut.edu.tw

Author affiliations
1 Department of Industrial Engineering and Management, Ming Chi University of Technology, Taiwan

DOI
http://dx.doi.org/10.7166/30-1-1627

ABSTRACT

Wafer yield is a key indicator to pursuing excellence in semiconductor manufacturing. With the increased wafer size, the enhanced complexity and precision of wafer fabrication is possible. Using monitoring to improve the process by predicting the yield has become an important quality issue. Most research uses the number of wafer defects, the area of the wafer, and fixed statistical distribution to predict the yield. Such methods fail to establish a high yield model due to the random and system-wide distribution of wafer defects. This study proposes the Mahalanobis-Taguchi system (MTS) to determine the key variables from the wafer acceptance test (WAT), and establish a classification model of yield grade. The general regression neural network (GRNN) was used to build a predicted model of the wafer probe yield from selected common variables. A real case from a Taiwan manufacturer of dynamic random-access memory (DRAM) is used as an example. It can get the 82 key and significant sequence variables of the WAT, with classification precision of over 90% and the R^2 of the GRNN prediction model at 0.73. Through demonstration, the result can effectively increase the yield and reduce the quality cost in DRAM manufacturing.

OPSOMMING

Flinteropbrengs is 'n sleutel indikator vir halfgeleiervervaardiging. Die toename in flintergrootte maak dit moontlik om die kompleksiteit en presisie van flintervervaardiging te verbeter. Wanneer monitering gebruik word om die opbrengsskatting te verbeter word dit 'n gehalte kwessie. Die meeste navorsing gebruik die aantal defekte, die flinteroppervlakte, en vasgestelde statistiese verdelings om die opbrengs te skat. Sulke metodes faal egter om 'n hoë-opbrengs model te bewerkstellig as gevolg van die lukrake en stelselwye verspreiding van defekte. Hierdie studie gebruik die Mahalanobis-Taguchi stelsel om die sleutel veranderlikes van die flinter aanvaardingstoets te bepaal en skep dan 'n klassifikasie model van die opbrengsgehalte. 'n Veralgemeende regressie neurale netwerk is gebruik om 'n skatting van die opbrengs te maak van gekose gemeenskaplike veranderlikes. 'n Gevallestudie van 'n Taiwannese vervaardiger van dinamies ewetoeganklike geheue word as voorbeeld gebruik. Die benadering kan twee-en-tagtig van die sleutel en noemenswaardige veranderlikes van die aanvaringstoets identifiseer met 'n klassifikasiepresisie van meer as negentig persent. Die R^2-koëffisiënt van die neurale netwerk is 0.73. Die resultaat kan die opbrengs verhoog en die gehaltekoste van dinamies ewetoeganklike geheue vervaardiging verminder.

1 INTRODUCTION

With fierce competition in the original equipment manufacturer (OEM) market, and the fabrication of semiconductors, the yield of wafer fabrication is an essential factor that influences the cost and competitiveness of products. Yield is a major comprehensive indicator that measures the financial aspects, the ability to fabricate, and the stable supply of products. A high yield can reduce costs and increase the marginal benefits of companies. Fabrication with a steady yield can facilitate production arrangements and attract orders, which can enhance the competitiveness of an enterprise. In general, yield rate measurement is arranged in the three stages of wafer fabrication: the wafer processing yield, the wafer probe test yield, and the wafer package yield.

Previous research on yield models for wafer concentrated on defect clustering [1], productivity optimisation [2], and interconnect yield analysis [3]. Some emphasised yield prediction. Lee and Ha [4] pointed out that the key factors of yield in semiconductor fabrication are particles or contaminants on the wafer, substances in the manufacturing instruments, manufacturing process parameters, process engineers' attitudes, and the design of semiconductors. Their study focused on the wafer probe test yield. Macro-yield modelling and micro-yield modelling are frequently used to predict the wafer probe test yield. Macro-yield modelling predicts the yield rate according to the size and defect density of the wafer, and other relevant factors. Most previous researchers adopted macro-yield modelling to deduce a yield prediction model [5], such as Poisson's yield model, Murphy's yield model, Seed's yield model, the Bose-Einstein yield model, and the negative binomial yield model [6]. Micro-yield modelling predicts a yield according to the area and parameter sensitivity of circuit devices, and other factors of circuit design. Usually, it is used to evaluate the influence of different defects, fabrication deviation, and circuit distribution on the yield.

The wafer acceptance test (WAT) is the fabrication variable test after wafer manufacturing and before the wafer probe test [7]. With an electrical property measurement, it aims to check whether the circuit parameters of the wafer are in the acceptable range. Ke and Rao [8] proposed the back propagation artificial neural network (ANN) model, which is designed to infer electrical test parameters from the given list of parameters with the intention of reducing test time, enhancing throughput, and improving cycle time. It is rather time-consuming and destructive to undertake the chip probe with an integrated circuit (IC) [9]. Thus the electrical property parameters of the testing wafer are used to check whether there is any problem in wafer production. The most important function of the WAT is to ensure the normality of the chip and to avoid a low yield rate. The WAT can also reflect problems in the production line and judge the quality of the metal according to the measurement results. The present study obtained the WAT information of a cooperative manufacturer, and found that there were 250 variables, 164 of which were measurement variables. The manufacturer did not monitor all variables in the process analysis — only 31 controlled variables, which might significantly influence the yield rate according to the experience of the engineers. The results of monitoring show that there was a low yield rate and even a '0' yield rate. The reason for this abnormality was that the monitored WAT could not provide full information about fabrication. The industry demands an efficient WAT monitoring mechanism to solve the problem of a low yield rate in manufacturing.

The frequently used variable reduction methods must know the requirements of a probability distribution or statistical hypothesis. The analysis of the primary constituents can lead to collinearity of information; however, it is difficult to define new variables. The ANN takes time, and it is hard to interpret the results. Worse still, the determined key variables are not based on the correlation between the variables. Taguchi proposed the Mahalanobis-Taguchi system (MTS), with a combination of statistical and stable engineering principles [10]. Orthogonal arrays and signal-noise ratio are used in MTS to select important variables and reduce system dimensions. Featuring simple calculations, it does not require any statistical hypotheses of variables. MTS can be used for disease diagnosis, fire alarm monitoring, earthquake prediction, weather broadcasting, pass rate prediction, voice identification, and credit assessment [11,12]. ANN is a calculation system that uses a large number of simply connected neurons to stimulate biological neural networks, and frequently uses the method to improve the quality and predict the yield of wafers [13,14]. Featuring high learning accuracy, fast recall, and high non-linear mapping, the back-propagation ability of ANN is a prediction model of monitoring-based learning. Stapper [15] pointed out that defect clusters influence the yield rate of a wafer, and the defect clusters of different areas can create different influences on the yield rate of wafers. Hsu and Chien [16] suggested integrating statistics with ANN to analyse the wafer cluster in wafer bin map (WBM) to locate defective wafers and apply them to the yield rate tests of wafer factories.

This research is a case study of an industry-university cooperative research project. At present, the primary reason for the underestimated prediction results of the cooperation of manufacturers is direct yield prediction; however, manufacturers are most concerned about the level of the yield [17]. In the past, scholars used the number and area of wafer defects and defect distribution as the input variables of the yield rate prediction mode. However, such information is not obtained from the wafer probe test, and wafer manufacturers want to get the information about yield distribution as early as possible. This study therefore concentrated on establishing the classification of the yield grade of dynamic random-access memory (DRAM) wafers and predicting the yield through common variables.
This study adopted the WAT variables to establish a yield grade classification model, which can identify yields of different grades and the relationship among the WAT measurement variables. Moreover, it can select the common variables of the yield of different classes by the MTS classification mode to establish the generalised regression neural network (GRNN) yield prediction model. With the yield grade classification model and yield rate prediction model, it is possible to monitor quality in WAT fabrication, and to find the crucial new WAT variables related to yield. Achieving these two parts can eliminate the current bottleneck and significantly improve fabrication, which will help enhance the production of manufacturers and reduce quality costs.

2 METHODOLOGY

In this study, MTS and the GRNN were used to select key variables and to establish the classification and prediction models respectively. There were four stages in the analysis process. In Stage 1, the WAT data obtained from the engineering database was appropriately pre-processed. In Stage 2, the reasons for the different yield grades were considered, and the information was grouped. Meanwhile, the yield was divided into five groups, according to the suggestions of the engineers: ‘100%~85’, ‘85%~80%’, ‘80%~40%’, ‘40%~2%’, and ‘2%~0%’. In Stage 3, MTS was adopted to find the key variables from the 164 variables. In the calculation, the group with a yield of ‘100%~85%’ was taken as the benchmark group and used to determine the critical variable combinations and classification models in the different groups. In Stage 4, the results of MTS were adopted as the input variables, and the GRNN was used to establish the predictive model.

MTS uses Mahalanobis distance (MD) to make a comprehensive assessment of multiple variables, and then adopts orthogonal arrays and signal-noise ratio to select variables. MD is a measure based on correlations between variables and the different patterns that can be identified and analysed with respect to a reference point. MD is a discriminant analysis tool that can be used to check whether the multivariable information is ‘homogeneous’ or ‘heterogeneous’. A higher level of information homogeneity leads to a shorter MD and vice versa. Hence, MD can serve as the basis for quality classification. If there are *k* variables and *n* samples in the multivariable sample set, then the MD_j of the *j-th* sample, given by Hsiao and Su [18], is

$$MD_j = \frac{1}{k} Z_{ij}^T C^{-1} Z_{ij}, Z_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}, i = 1, 2, \ldots, k; j = 1, 2, \ldots, n \tag{1}$$

where, $Z_{ij} = (Z_{1j}, Z_{2j}, \ldots, Z_{kj})$ is the vector of the standardised value; x_{ij} is the *i-th* variable value of the *j-th* sample; $\overline{x_i}$ is the mean of the *i-th* variable; S_i is the standard deviation of the *i-th* variable; C^{-1} is the inverse of the correlation matrix *C*; *k* is the total number of variables; *i* is the number of variables; and *j* is the number of samples.

The benchmark group (100%~85%) was used to calculate and standardise the mean of the variables and the standard deviation. In the second step, the standardised results were used to calculate the inverse matrix of the relevant correlation coefficients to obtain all the Mahalanobis distances (MD) from the benchmark group. For the third step, information about the other groups was selected one by one, and then standardised with the mean and standard deviation of the variables in the benchmark group. Fourth, the inverse matrix of the relevant coefficients of the benchmark group was used to calculate the MD value of the other groups. When the number of variables was increased, MD was similar to a chi-square distribution. In terms of the classification threshold, this study adopted the maximised total precision of the training sample.

The configuration of orthogonal arrays (OA) and signal-noise (SN) ratios featuring 'the larger, the better' were then used to select key variables. The rank of effect gain selected the best number of variables. The computational formula for the larger-is-better SN ratio corresponding to the i^{th} run of OA is given by

$$\eta_i = -10\log\left(\frac{1}{b}\sum_{i=1}^{b}\frac{1}{MD_i}\right), i = 1, 2, \ldots, b \tag{2}$$

where *b* is the number of repetitions under each experimental combination.

After calculating all the SN ratios in the orthogonal arrays, the effect gain of the variables would be calculated. Those variables with a higher effect gain would be taken as the key variables of the classification mode. The formula is given by

$$Gain_j = \overline{SN_j^+} - \overline{SN_j^-} \tag{3}$$

where, $\overline{SN_j^+}$ denotes the mean of all the signal-noise ratios with the variable X_j; $\overline{SN_j^-}$ indicates the mean of all the signal-noise ratios without the variable X_j.

The selection of variables is a two-level issue. When the standard orthogonal arrays are at level-2, up to 63 variables can be processed. However, there were no appropriate orthogonal arrays in the 164 variables in the WAT data of this study. The paper thus proposed a modification that could be divided into the following steps:

1. All the variables are randomly grouped. Matlab® is used to equip each variable with a value ranging from '0' to '1', and the variables will be ranked in an ascending order. In this way, all the variables can be divided into three groups (55, 55, and 54).
2. MTS is undertaken in each group to calculate the effect gain of all the variables.
3. Repeat Step 1 and Step 2 many times until the mean of the effect gain of the variables becomes stable. In this way, the calculation of the effect gain of the 164 variables is completed.

After the MTS calculation of the other four groups ((100%~85% vs 85%~80%) (100%~85% vs 40%~20%), (100%~85% vs 20%~2%), and (100%~85% vs 2%~0%)), the common variables are selected to establish the prediction model by the GRNN.

The GRNN is a monitoring-based learning network derived from the probability neural network. Its dynamic learning mode can be used for prediction and control, and it shows high prediction ability, regardless of whether the regression model is linear or non-linear. Meanwhile, it can predict the pass rate according to previous information, and without hypothetical distribution. The framework of the GRNN, where the input units are merely distribution units, provides all of the (scaled) measurement variables. It distributes all *x* measurement values among all the pattern units in the second layer, where each pattern unit represents a training example. If a new *x* vector enters the network, the squared value of the gap between the vector and the vector of the training sample will be aggregated and added into the non-linear function. The value obtained from this function will be the output value of the pattern unit. The output value will then be sent to the summation unit, which will aggregate the weight vectors and the result of all the observation values of *y*, multiplied by the value of the weight vectors. For the last step, the evaluated value of *y* can be obtained by the division between the two output values of the summation unit. The only thing in GRNN that must be determined is a smooth parameter. In general, the holdout method proposed by Specht [19] was adopted to define a smooth parameter. The steps of the methods are as follows:

1. Select specific σ value.
2. Remove a training sample each time and use the remaining samples to establish a network that will be adopted to evaluate the value of the removed sample.
3. Repeat Step 2 for *n* (the number of training samples) times and record the mean square error (MSE) between each evaluation value and the sample value. Then aggregate all the MSEs.
4. Select other values of σ and repeat Step 2 and Step 3.
5. The minimum value of σ of MSE is the best one.

3 CASE STUDY AND DATA ANALYSIS

The engineers selected 25,000 copies of information from the engineering information system to demonstrate the proposed method. The WAT parameters include all electrical measurements of devices or transistors. To avoid superfluous data and reduce analysis time, it is essential to select relevant and useful WAT parameters. In this study, each of the 164 WAT variables refers to an electrical or physical feature value such as Isat (saturation current), Ioff (lower leakage), Vt (threshold voltage), Rc (contract resistance), and BV (breakdown). During the first three weeks, the data was taken as the training sample; in the fourth week, it was taken as the testing sample. Given that the key variables that influence the yield are not necessarily the same, the WAT data was divided into five groups according to the yield grade, after discussion with the engineers: '100%~85%', '85%~80%', '80%~40%', '40%~2%', and '2%~0%'. Taking the group with the yield of '100%~85%' as the benchmark group, the researcher undertook MTS analysis of the remaining four groups.

First, the study conducted analysis where the group with the yield of '100%~85%' was taken as the benchmark group, and the group with the yield of '85%~80%' was regarded as an abnormal group. There were no appropriate orthogonal arrays in the WAT data. Thus the variables were randomly divided into three groups (55, 55, and 54). The effect gain was then calculated according to the calculation procedure of MTS. Taking the first group as an example; the first step was to establish the Mahalanobis space of the benchmark group. The MD of the abnormal group was then calculated. After that, the orthogonal arrays were configured, and the signal-noise ratio was calculated, the results of which are shown in Table 1. For the last step, the signal-noise ratio was used to calculate the effect gain of the computable variables. The second and third groups were calculated by the same procedure. In this way, the first stimulation of the effect gain of all the variables was completed.

Table 1: Design of orthogonal arrays and signal-noise ratio

Run	X_{76}	X_{75}	X_3	X_{137}	X_{56}	…	X_{138}	X_{55}	X_{142}	MD_1	…	MD_{1125}	SN ratio
1	1	1	1	1	1	…	1	1	1	6.2778	…	0.4473	-29.8352
2	1	1	1	1	1	…	2	2	2	10.1295	…	0.4242	-5.4724
3	1	1	1	1	1	…	2	2	2	8.148	…	0.5438	-3.2835
4	1	1	1	1	1	…	1	1	1	6.8302	…	0.4421	-31.4559
5	1	1	1	1	1	…	2	2	2	1.7009	…	0.4646	-6.6395
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
62	2	2	1	2	1	…	1	1	2	11.131	…	0.502	-32.1574
63	2	2	1	2	1	…	1	1	2	7.8756	…	0.5187	-31.937
64	2	2	1	2	1	…	1	2	1	7.731	…	0.3552	-2.5049

Second, the same procedure was adopted to calculate the mean variable effect gains on the basis of every ten times of stimulation, and the results were sorted, as in Table 2. It was found that the first 14 variables remained unchanged in the 220th and 230th times of stimulation, with the remaining results changing slightly. For that reason, the results of the 230th time of stimulation were taken as the basis for choosing important variable combinations. Likewise, the best numbers of stimulation for the remaining three groups — (100%~85% vs 40%~20%), (100%~85% vs 20%~2%), and (100%~85% vs 2%~0%) — were 260, 210, and 230 respectively.

Third, the effect gain was calculated according to the best number of stimulation. Table 3 shows that the significant variables obtained from MTS of the four groups were different from each other. For instance, when the pass rate declined to 2%~0%, the first five key WAT variables were X_{122}, X_{93}, X_{118}, X_{36}, and X_{98}. Some variables were important in groups with different pass rates. For example, X_{122} ranked second and first in the groups with a pass rate of '40%~2%' and '2%~0%' respectively. According to the ranking of the effect gain of variables, the corresponding improvement can be made in WAT variables with a high ranking to increase the yield.

Fourth, use MTS to classify the yield grade. Taking (100%~85% vs 85%~80%) as an example, the trial and error method was adopted to evaluate the threshold. Meanwhile the number of WAT variables gradually declined to 134, 130, 122, 109, 74, 43, 27, 16, and 5, according to the different effect gains (>0, >0.05, >0.1, >0.15, >0.2, >0.25, >0.3, >0.35, >0.4). These results are shown in Table 4. The highest precision (=0.7882) was chosen to determine the variable combination, and the original 164 WAT variables were reduced to 134 variables. Likewise, the thresholds of (100%~85% vs 40%~20%), (100%~85% vs 20%~2%), and (100%~85% vs 2%~0%) were 1.2, 1.2, and 1.7 respectively. The numbers of their WAT variables were 130, 128, and 114 respectively. Their precision was 82.29 per cent, 82.98 per cent, and 89.98 per cent respectively, as shown in Table 5. All the results were then evaluated on the basis of the test samples. The precision of the four groups (100%~85% vs 85%~80%), (100%~85% vs 40%~20%), (100%~85% vs 20%~2%), and (100%~85% vs 2%~0%) was 95.92 per cent, 95.24 per cent, 94.18 per cent, and 95.31 per cent respectively.

Table 2: Simulation number and the rank of effect gains of variable for (100%~85% vs 85%~80%)

	Simulation number								
Ranking of variable	10	20	…	180	190	200	210	220	230
1	X_{105}	X_{72}	…	X_{78}	X_{78}	X_{78}	X_{78}	X_{78}	X_{78}
2	X_{85}	X_{105}	…	X_{17}	X_{17}	X_{23}	X_{23}	X_{23}	X_{23}
3	X_{98}	X_{15}	…	X_{88}	X_{88}	X_{17}	X_{17}	X_{17}	X_{17}
4	X_{7}	X_{100}	…	X_{38}	X_{23}	X_{88}	X_{88}	X_{88}	X_{88}
5	X_{88}	X_{85}	…	X_{116}	X_{90}	X_{101}	X_{101}	X_{90}	X_{90}
6	X_{129}	X_{91}	…	X_{51}	X_{116}	X_{27}	X_{90}	X_{101}	X_{101}
7	X_{5}	X_{96}	…	X_{23}	X_{38}	X_{90}	X_{116}	X_{116}	X_{116}
8	X_{116}	X_{7}	…	X_{90}	X_{51}	X_{116}	X_{27}	X_{96}	X_{96}
9	X_{141}	X_{98}	…	X_{141}	X_{6}	X_{38}	X_{38}	X_{100}	X_{100}
10	X_{155}	X_{5}	…	X_{6}	X_{141}	X_{51}	X_{100}	X_{27}	X_{27}
11	X_{31}	X_{37}	…	X_{72}	X_{27}	X_{144}	X_{51}	X_{51}	X_{51}
12	X_{12}	X_{156}	…	X_{27}	X_{72}	X_{141}	X_{141}	X_{141}	X_{141}
13	X_{124}	X_{129}	…	X_{144}	X_{140}	X_{6}	X_{95}	X_{95}	X_{95}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
160	X_{104}	X_{104}	…	X_{81}	X_{81}	X_{81}	X_{81}	X_{81}	X_{81}
161	X_{113}	X_{113}	…	X_{113}	X_{113}	X_{113}	X_{113}	X_{113}	X_{113}
162	X_{74}	X_{74}	…	X_{74}	X_{74}	X_{74}	X_{74}	X_{74}	X_{74}
163	X_{43}	X_{43}	…	X_{123}	X_{123}	X_{123}	X_{123}	X_{123}	X_{123}
164	X_{123}	X_{123}	…	X_{43}	X_{43}	X_{43}	X_{43}	X_{43}	X_{43}

Table 3: The ranking of the effect gains of variables and for different yield groups

Ranking of variable	(100%~85% vs 85%~80%)		(100%~85% vs 40%~20%)		(100%~85% vs 20%~2%)		(100%~85% vs 2%~0%)	
	variable	effect gains	variable	effect gains	variable	effect gains	variable	effect gains
1	X_{78}	0.5924	X_{116}	0.6236	X_{121}	0.2278	X_{122}	0.2729
2	X_{23}	0.4641	X_{78}	0.5351	X_{122}	0.2047	X_{93}	0.2694
3	X_{17}	0.4551	X_{51}	0.5105	X_{21}	0.1994	X_{118}	0.2688
4	X_{88}	0.4343	X_{70}	0.4573	X_{70}	0.1994	X_{36}	0.2674
5	X_{90}	0.4039	X_{145}	0.4504	X_{142}	0.1971	X_{98}	0.253
6	X_{101}	0.3967	X_{54}	0.4375	X_{62}	0.1908	X_{52}	0.2529
7	X_{116}	0.3921	X_{110}	0.4219	X_{28}	0.189	X_{96}	0.2519
8	X_{96}	0.3901	X_{57}	0.4208	X_{125}	0.1867	X_{110}	0.2445
9	X_{100}	0.3738	X_{39}	0.3961	X_{1}	0.1865	X_{162}	0.2423
10	X_{27}	0.3713	X_{162}	0.3939	X_{86}	0.1808	X_{43}	0.2397
11	X_{51}	0.3599	X_{129}	0.3894	X_{163}	0.179	X_{100}	0.2366
12	X_{141}	0.3589	X_{38}	0.3806	X_{98}	0.1789	X_{123}	0.2324
13	X_{95}	0.3564	X_{36}	0.3801	X_{96}	0.1778	X_{137}	0.2316
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
160	X_{81}	-2.0701	X_{126}	-7.3133	X_{109}	-2.5953	X_{158}	-4.6049
161	X_{113}	-10.2477	X_{97}	-10.6336	X_{113}	-2.8346	X_{15}	-7.3073
162	X_{74}	-13.6918	X_{113}	-11.6341	X_{126}	-8.2348	X_{65}	-9.2578
163	X_{123}	-36.489	X_{43}	-30.0052	X_{35}	-13.142	X_{125}	-12.5813
164	X_{43}	-37.6084	X_{123}	-37.827	X_{127}	-13.2393	X_{113}	-20.2627

Table 4: （100%~85%）vs（85%~80%）of the correct classification rate

Effect gains	Number of variables	Threshold	Total correct rate	Correct rate of benchmark group	Correct rate of abnormal group
>0	134	1.125	0.7822	0.7697	0.7858
>0.05	130	1.15	0.7663	0.7697	0.7653
>0.1	122	1.125	0.7635	0.7634	0.7636
>0.15	109	1.05	0.7316	0.7129	0.7369
>0.2	74	1	0.6976	0.7003	0.6969
>0.25	43	0.95	0.6637	0.6656	0.6631
>0.3	27	0.825	0.5922	0.5931	0.592
>0.35	16	0.75	0.5624	0.5615	0.5627
>0.4	5	0.7	0.5312	0.5426	0.528

Table 5: The optimal for the correct rate for the different benchmark groups and the abnormal groups in the study

Effect gains	Benchmark group	Abnormal group	Number of variables	Threshold	Total correct rate	Correct rate of benchmark group	Correct rate of abnormal group
>0	100%~85%	85%~80%	134	1.125	0.7822	0.7697	0.7858
>0	100%~85%	80%~40%	130	1.2	0.8229	0.8107	0.825
>0	100%~85%	40%~2%	128	1.2	0.8298	0.8297	0.8333
>0	100%~85%	2%~0%	114	1.7	0.8997	0.9054	0.75

The 31 WAT variables currently used by the manufacturer were used to calculate the MTS procedure, as proposed by the researcher. According to Table 6, the precision of the four groups, (100%~85% vs 85%~80%), (100%~85% vs 40%~20%), (100%~85% vs 20%~2%), and (100%~85% vs 2%~0%), was 55.62 per cent, 59.39 per cent, 61.87 per cent, and 66.96 per cent respectively. These precisions were far lower than the results of this study. Therefore, the established MTS procedure is effective in seeking and monitoring the WAT variables of different yield grades, and shows a high level of classification rate.

Table 6: The result of the correct rate using the current variables of the manufacturer

Benchmark group	Abnormal group	Number of variables	Threshold	Total correct rate	Correct rate of benchmark group	Correct rate of abnormal group
100%~85%	85%~80%	31	0.85	0.5562	0.5552	0.5564
100%~85%	80%~40%	31	0.88	0.5939	0.5931	0.5941
100%~85%	40%~2%	31	0.95	0.6187	0.6188	0.6167
100%~85%	2%~0%	31	1.05	0.6696	0.6697	0.6692

Fifth, the selected WAT variables currently used by the manufacturer were compared. According to the different levels of pass rates, 82 common WAT variables were selected from the four groups, (100%~85% vs 85%~80%), (100%~85% vs 40%~20%), (100%~85% vs 20%~2%), and (100%~85% vs2%~0%). The manufacturer currently monitors 20 variables (those in italics), shown in Table 7, in the common WAT variables from MTS. The results show that the selected WAT variables have practical representativeness. In the current situation, the variables monitored by engineers, according to their experience, are inconsistent with the actual situation, which explains the poor monitoring effect.

Finally, the 82 selected common WAT variables and the GRNN were used to establish the pass rate prediction model. The researcher wrote a Matlab® program with an initial smooth parameter of 0.5. The program could automatically modify the smooth parameter to reduce the error to a minimum level. In this way, a smooth parameter value with the minimum error could be obtained, as shown in Table 8. Through comparison, the prediction model (R^{2}=73.12%), as established with common variables, is better than the current prediction model (R^2=33.84%).

Table 7: The common variables from MTS analysis

X_{1}	X_{21}	X_{39}	X_{60}	X_{83}	X_{100}	X_{137}	X_{148}	X_{163}
X_{2}	X_{23}	X_{40}	X_{61}	X_{85}	X_{101}	X_{138}	X_{149}	X_{164}
X_{4}	X_{24}	X_{41}	X_{62}	X_{86}	X_{110}	X_{139}	X_{150}	
X_{6}	X_{25}	X_{45}	X_{63}	X_{89}	X_{115}	X_{141}	X_{151}	
X_{8}	X_{27}	X_{46}	X_{64}	X_{90}	X_{116}	X_{142}	X_{152}	
X_{12}	X_{28}	X_{49}	X_{70}	X_{91}	X_{117}	X_{143}	X_{153}	
X_{13}	X_{30}	X_{51}	X_{75}	X_{93}	X_{118}	X_{144}	X_{154}	
X_{16}	X_{32}	X_{52}	X_{79}	X_{95}	X_{122}	X_{145}	X_{159}	
X_{18}	X_{33}	X_{55}	X_{80}	X_{98}	X_{132}	X_{146}	X_{161}	
X_{20}	X_{37}	X_{59}	X_{82}	X_{99}	X_{133}	X_{147}	X_{162}	

Table 8: The result of the GRNN between the variables by MTS and current variables

		Training sample		Testing sample	
	Smooth parameter α	R^2	MSE	R^2	MSE
Variables by MTS	0.3943529	0.7509	0.002	0.7312	0.002
Current variables	0.1847059	0.4212	0.008	0.3384	0.008

4 CONCLUSION

The production of semiconductors involves thousands of complicated steps that are closely related to each other. Any error in a process step will result in a lower wafer yield rate. With the WAT data offered by a DRAM manufacturer in Taiwan as the research subject, the MTS and the GRNN are proposed to establish a classification and prediction yield model of a wafer probe test in this paper.

First, the variables were grouped according to the yield rate grades, and MTS was adopted to seek the WAT variables that monitored different yield rates. According to the experience of the engineers, previous practitioners selected a fixed 31 of the 164 variables for quality monitoring. However, the results showed a slightly lower wafer yield, and even a '0' yield rate. This is because the monitored WAT could not provide sufficient information about fabrication. There should be different WAT variables to control the yield of different grades. The results of this study obtained four kinds of WAT variables, corresponding to different yield grades. The number of variables were 134, 130, 128, and 114 respectively. Meanwhile, a yield grade classification model was established, with a precise classification rate of over 90 per cent for the testing sample. A summary of the classification models with yield rates of different grades led to 82 common WAT variables and the establishment of the GRNN yield rate prediction model. The R2 of the model is over 0.7, which is the result expected by the industry. The research results are better than those of Chang, Chen and Wang [20] used for the backpropagation neural network (BPNN) and the group method of data handling (GMDH) for WAT yield prediction.

This study gives specific suggestions for practitioners to improve their WAT monitoring mechanism. Through a demonstration, the result can increase the wafer yield rate and reduce quality cost in DRAM manufacturing. Improving yield would significantly reduce the manufacturing cycle time. An accurate yield prediction model would help to prevent the production of non-conforming wafers before a malfunction is detected in the process. Practically, it is rather time-consuming and destructive to undertake a chip probe (C/P) with an IC. Thus effective monitoring of the WAT variables can check whether there is any problem in wafer fabrication, and reduce or gradually replace the C/P test. This would ensure the normal functions of the chip and avoid a low pass rate. For future research and application, it is suggested that a multivariable control chart be adopted to monitor the WAT variables. Developing big data analysis for quality problems in the production line or machine parameters can provide more efficient quality monitoring in fabrication.

REFERENCES

[1] **Michalka, T.L., Varshney, R.C. & Meindl, J.D.** 1990. A discussion of yield modeling with defect clustering, circuit repair, and circuit redundancy. *IEEE Transactions on Semiconductor Manufacturing*, 3(3), 116-127.

[2] **Stapper, C.H., McLaren, A.N. & Dreckmann, M.** 1980. Yield model for productivity optimization of VLSI memory chips with redundancy and partially good product. *IBM Journal of Research and Development*, 24(3), 398-409.

[3] **Ghaida, R.S. & Zarkesh-Ha, P.** 2009. A layout sensitivity model for estimating electromigration-vulnerable narrow interconnects. *Journal of Electronic Testing*, 25(1), 67-77.

[4] **Lee, J.H. & Ha, S.H.** 2009. Recognizing yield patterns through hybrid applications of machine learning techniques. *Information Sciences*, 179(6), 844-850.

[5] **Tyagi, A. & Bayoumi, M.A.** (1992). Defect clustering viewed through generalized Poisson distribution. *IEEE Transactions on Semiconductor Manufacturing*, 5(3), 196-206.

[6] **Berglund, C.N.** 1996. A unified yield model incorporating both defect and parametric effects. *IEEE Transactions on Semiconductor Manufacturing*, 9(3), 447-454.

[7] **Chien, C.F., Hsu, C.Y. & Hsiao, C.W.** 2012. Manufacturing intelligence to forecast and reduce semiconductor cycle time. *Journal of Intelligent Manufacturing*, 23(6), 2281-2294.

[8] **Ke, J. & Rao, M.V.C.** 2006. Extraction of electrical test parameters by artificial neural network. *International Journal of Industrial Engineering: Theory, Applications and Practice*, 13(1), 71-80.

[9] **Liu, S.F., Chen, F.L., Lin, S.C., Chang, K.J. & Yu, S.M.** 2011. A clustering analysis model for golden die extractions based on wafer acceptance test at semiconductor R&D stage. *Journal of Testing and Evaluation*, 39(4), 522-528.

[10] **Taguchi, G. & Jugulum, R.** 2002. *The Mahalanobis-Taguchi strategy: A pattern technology system*. New York:John Wiley & Sons.
[11] **Teshima, S., Hasegawa, Y. & Tatebayashi, K.** 2012. *Quality recognition & prediction: Smarter pattern technology with the Mahalanobis-Taguchi system*. New York: Momentum Press.
[12] **Wang, J. & Duan, C.** 2009. Structural health monitoring using Mahalanobis-Taguchi system. International Conference on Information Engineering and Computer Science, Wuhan, China.
[13] **Sinanoglu, C.** 2008. Design of an artificial neural network for assembly sequence planning system. *International Journal of Industrial Engineering: Theory, Applications and Practice*, 15(1), 92-103.
[14] **Tsai, T.N. & Tsai, C.W.** 2014. Development of a closed-loop diagnosis system for reflow soldering using neural networks and support vector regression. *International Journal of Industrial Engineering: Theory, Applications and Practice*. 21(1), 19-33.
[15] **Stapper, C.H.** 1986. On yield, fault distributions, and clustering of particles. *IBM Journal of Research and Development*, 30(3), 326-338.
[16] **Hsu, S.C. & Chien, C.F.** 2007. Hybrid data mining approach for pattern extraction from wafer bin map to improve yield in semiconductor manufacturing. *International Journal of Production Economics*, 107(1), 88-103.
[17] **Li, T.S., Huang, C.L. & Wu, Z.Y.** 2006. Data mining using genetic programming for construction of a semiconductor manufacturing yield rate prediction system. *Journal of Intelligent Manufacturing*, 17(3), 355-361.
[18] **Hsiao, Y.H. & Su, C.T.** 2009. Multiclass MTS for saxophone timbre quality inspection using waveform-shape-based features. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 39(3), 690-704.
[19] **Specht, D.F.** 1991. A general regression neural network. *IEEE Transactions on Neural Networks*, 2(6), 568-576.
[20] **Chang, P.Y., Chen, K.T. & Wang, C.Y.** 2011. Modeling of wafer die yield by WAT parameters. *Journal of Quality*, 18(6), 519-538.