Bayesian Inference for Mining Semiconductor Manufacturing Big Data for Yield Enhancement and Smart Production to Empower Industry 4.0

Article *in* Applied Soft Computing · November 2017

DOI: 10.1016/j.asoc.2017.11.034

3 authors, including:

Marzieh Khakifirooz
Tecnológico de Monterrey
37 PUBLICATIONS **93** CITATIONS

SEE PROFILE

Applied Soft Computing xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Bayesian inference for mining semiconductor manufacturing big data for yield enhancement and smart production to empower industry 4.0

Marzieh Khakifirooz, Chen Fu Chien*, Ying-Jen Chen

Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Hsinchu, Taiwan

ARTICLE INFO

Article history:
Received 30 September 2016
Received in revised form 4 June 2017
Accepted 22 November 2017
Available online xxx

Keywords:
Bayesian approach
Semiconductor manufacturing
Multi-collinearity
Yield enhancement
Big data analytics
Smart production

ABSTRACT

Big data analytics have been employed to extract useful information and derive effective manufacturing intelligence for yield management in semiconductor manufacturing that is one of the most complex manufacturing processes due to tightly constrained production processes, reentrant process flows, sophisticated equipment, volatile demands, and complicated product mix. Indeed, the increasing adoption of multimode sensors, intelligent equipment, and robotics have enabled the Internet of Things (IOT) and big data analytics for semiconductor manufacturing. Although the processing tool, chamber set, and recipe are selected according to product design and previous experiences, domain knowledge has become less efficient for defect diagnosis and fault detection. To fill the gaps, this study aims to develop a framework based on Bayesian inference and Gibbs sampling to investigate the intricate semiconductor manufacturing data for fault detection to empower intelligent manufacturing. In addition, Cohen's kappa coefficient was used to eliminate the influence of extraneous variables. The proposed approach was validated through an empirical study and simulation. The results have shown the practical viability of the proposed approach.

1. Introduction

Semiconductor fabrication facilities (fabs) are the most capital-intensive and fully automated manufacturing systems, in which similar equipment and processes are employed to produce integrated circuits via lengthy complicated processes with tightly constrained production processes, reentrant process flows, sophisticated equipment, and waiting time limits to fulfill the volatile demands of high product mix. The yield learning curve of semiconductor manufacturing [1–3] has demonstrated that data analytics, cumulative engineering training, and domain knowledge have significantly enhanced yield, and thus integrated yield enhancement methods [4] and [5] are widely employed. However, high dimensionality and multi-collinearity [6] among the operation variables cause difficulty in embracing the independent condition for statistical testing and conventional analysis. Furthermore, the increasing adoption of multimode sensors, intelligent equipment, and robotics have enabled the Internet of Things (IOT) and advanced analytics of automatically collected big data for predicting process behavior and identifying defective tools, chambers, and products to improve the yield and productivity [7–9] for smart production of semiconductor manufacturing. On the other hand, numerous fundamental demands for computational issues such as variable selection and data preparation are excessively dependent on domain knowledge. Therefore, substantial differences exist between research results and actual processes, and thus researchers have combined technological and psychological factors to improve the interaction between systems and behavior [10–12].

Indeed, building an accurate function of process data in industrial settings is difficult since the process variables are highly correlated [13] and [14], in which high collinearity occurs in high-dimensional modeling because of an increased probability of dependency among the parameters. Furthermore, generating a training set that can circumvent this phenomenon is difficult, since the variables cannot be dropped without understanding the interactions among parameters [15]. On the other hand, process engineers are interested in identifying a few essential variables to effectively identify root causes. Alternatively, a novel approach to hedge and compensate the critical dimension variation of the developed-and-etched circuit patterns via a short-loop processes is evolved for yield enhancement [16]. Also, advanced process control (APC) with dynamically adjusted proportional-integral run-to-run controller is developed to compensate overlay errors [5]. Since hundreds of factors must be considered simultaneously to accurately characterize the yield performance, a retrospective design of experiment (DOE) data mining that matches potential designs with a

* Corresponding author at: Department of Industrial Engineering & Engineering Management, National Tsing Hua University, 101 Section 2 Kuang Fu Road, Hsinchu 30013, Taiwan.
E-mail address: cfchien@mx.nthu.edu.tw (C.F. Chien).

https://doi.org/10.1016/j.asoc.2017.11.034
1568-4946/

huge amount of data automatically collected to enable effective and efficient data analytics [17].

To fill the gaps for dealing with multi-collinearity and empirical variable selection behavior, this study aims to develop a framework based on Bayesian inference and Gibbs sampling for analyzing semiconductor manufacturing big data with high volume of variables, where Cohen's kappa coefficient was used to eliminate the influence of extraneous variables. Indeed, the proposed Gibbs sampling methodology has been used for deep learning [17], high-dimensional linear regression [18], and prior knowledge learning [19]. To estimate the validity of the proposed approach, simulation and an empirical study were conducted with data collected in a world leading semiconductor manufacturing company.

The remainder of the paper is organized as follows: Section 2 introduces the fundamental material for application to semiconductor manufacturing. Section 3 presents a research framework with detailed procedures. Section 4 details the validation of the framework with simulation and subsequently empirical study in Section 5. Section 6 concludes with a discussion of results and further research directions.

2. Fundamental

The notation and terminologies used in this paper are as follows:

i Index of process steps
j Index of manufacturing tools
l Index of observations
k Number of process steps
s_i Number of manufacturing tools in the ith step
m_{s_i} Number of manufacturing chambers for each tool in the ith step
M Number of process variables
N Number of observations
X_{il} Nominal factor set of step-tool-chamber information (process variable)
$\boldsymbol{X} N \times M$ matrix of process information
$\boldsymbol{Y} N \times 1$ matrix of yield percentage
$\boldsymbol{Y}^c N \times 1$ matrix of yield category
$\hat{\boldsymbol{Y}}^c N \times 1$ matrix of predicted yield category
p Prior probability for the $i_{s_i - m_{s_i}}$ th binary variable
$\mathbb{I}$ Indicator function

Wafer fabrication is complex and lengthy that consists of segments of process steps including oxidation/deposition/metallization, lithography, etching, ion implantation, photo-resist strip, cleaning, inspection, and measurement. Fig. 1 illustrates a segment/short-loop of process steps, and at each step, the wafer is fabricated by a specific tool. A number of alternative tools and chambers may be qualified for performing the same process in a step. However, only one of the many tool-chamber features is applied to a wafer. In particular, since hundreds of factors must be considered simultaneously to accurately characterize the yield performance, a retrospective design of experiment (DOE) data mining that matches potential designs with a huge amount of data automatically collected to enable effective and efficient data analytics [20].

Suppose that k process steps exist for completing a product. s_i denotes the number of manufacturing tools in the ith step, where $i = 1, \ldots, k$. Similarly, m_{s_i} represents the number of manufacturing chambers for the jth corresponding tool in the ith step, when $1 \le j \le s_i$. Hence, the total set of process variables is estimated using

$$M = \sum_{i=1}^{k} \sum_{j=1}^{s_i} m_{s_i} = \sum_{i=1}^{k} s_i * m_{s_i}.$$

To achieve shorter production cycle time, faster process development, higher yield efficiency, and lower contamination risk, cluster tools that consist of specific processing, cleaning, or cooling chambers, with loading and unloading chambers are increasingly employed in semiconductor manufacturing. Without loss of generality, the tool compounds and process chambers are noted with singular nominal factors as the following explanation.

2.1. Approximate inference for distribution of nominal data

A categorical distribution, i.e., a multinomial distribution, is a generalization of the Bernoulli distribution for a categorical random variable with more than two possible outcomes [21]. In the present study, categorical distribution was used to construct the nonparametric Bayesian model for multivariate nominal data.

Bayesian models can represent dependency among variables, in which current knowledge about model parameters is expressed by prior distribution, denoted as $p(\theta)$, and will be updated with new evidence θ' with the likelihood $p(\theta'|\theta)$ to derive the posterior distribution. If posterior and prior probability distributions are from the same family of distributions, they are then called conjugate distributions, and the prior is called a conjugate prior. In particular, Dirichlet distribution is a conjugate prior for the categorical distribution of multinomial data.

However, quantifying the idea of a Bayesian model is difficult for multinomial data because of the complexity in estimating the parameters of the Dirichlet distribution. Nevertheless, one approach to facilitate this difficulty is to sample values from the distribution before computing the sample statistics.

Sampling from an arbitrary distribution can be extremely complicated. However, the Markov chain Monte Carlo (MCMC) method has facilitated Bayesian statistics [22] that can be applied widely. A Markov chain is a sequence of events with a distribution that depends only on the outcome of the previous event. One basis of Markov chain theory posits that, if the probabilities associated with different events are constructed in the correct manner and the chain has a sufficient length, then the event distribution can be made equal to any arbitrary distribution, including a posterior distribution.

Gibbs sampling is an MCMC technique that is suitable for this task. The concept of Gibbs sampling involves generating posterior samples by eliminating each variable for sampling from its conditional distribution, with the remaining variables fixed to their current values.

The Gibbs sampling-based searching algorithm [23] was originally proposed by George and McCulloch [24]. Garcia-Donato and Martinez-Beneito [25] showed that this simple sampling strategy combined with estimates based on the frequency of visits (the one implemented in the present study) provides extremely reliable results.

The Gibbs sampler is used to estimate the posterior distribution and assess the model parameters [26] and [27]. However, the sampler is particularly favorable for sampling from imbalanced class distribution [28]. Gibbs sampling was preferred since it functions efficiently in the presence of multi-collinearity and high dimensionality.

2.2. Cohen's kappa coefficient

Cohen's kappa is a statistical method that measures the levels of agreement between two raters, each of which is classified into several exclusive categories.

A brief overview of nonparametric techniques shows that kappa is most typically applied to predictive models devised using unbalanced data [29]. This study employed the kappa coefficient for data clearance and variable association.

Please cite this article in press as: M. Khakifirooz, et al., Bayesian inference for mining semiconductor manufacturing big data for yield enhancement and smart production to empower industry 4.0, Appl. Soft Comput. J. (2017), https://doi.org/10.1016/j.asoc.2017.11.034

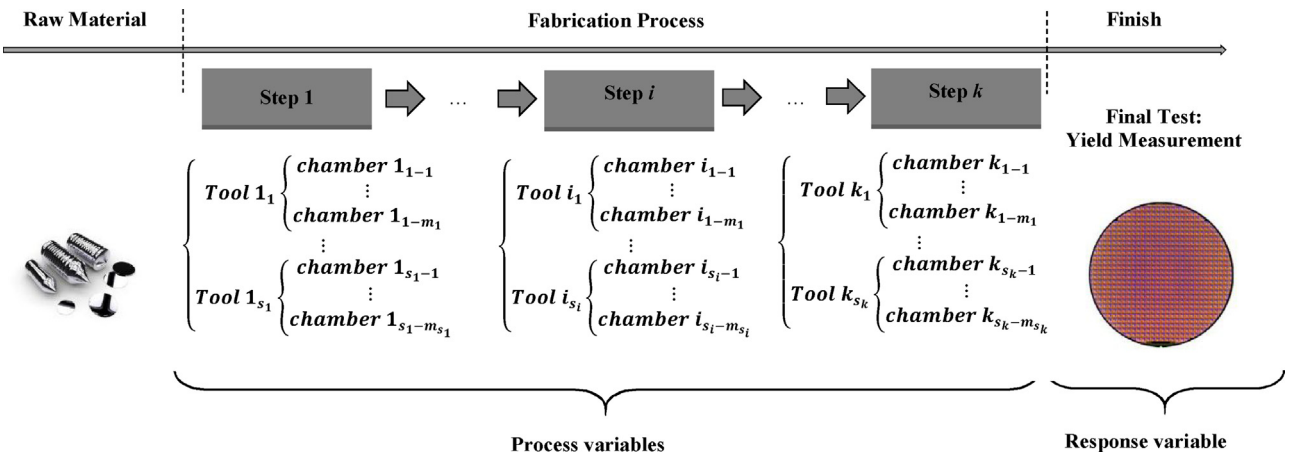

Fig. 1. Schematic of batch production of a wafer.

Class- or cluster-based methods split data into groups, but each variable can be incorporated into one group alone. Therefore, the rates of misclassification or miss-clustering exert a critical influence on the final decision. However, Cohen's kappa allows variables to be classified into more than one group, thereby eliminating any chance of misleading results.

2.3. Reliable k-fold repeated random subsampling validation

To estimate the performance accuracy of the predictive model in practice, the repeated random subsampling validation (repeated-CV) technique, was employed. The advantage of this method is when the responses are dichotomous with an unbalanced representation of the two response values in the data, then it is particularly useful if the random samples are generated in such a manner that the mean response values for the training and testing sets are equal. Furthermore, it can reduce the variance without increasing bias in comparison to the normal cross-validation method.

To lead to the best stability and performance of constructed model, a "one-standard error" rule [30] is used with cross-validation, in which we choose the most parsimonious model whose error is no more than one standard error above the error of the best model.

3. Proposed framework

In the present study, a data-mining framework was devised to examine large volumes of semiconductor manufacturing data for identifying inadequate tool-chamber at a determined production time. This framework comprises four major steps: 1) problem definition, 2) data preparation, 3) data mining and key factor screening, and (4) model construction, evaluation, and interpretation (Fig. 2).

3.1. Problem definition

In practice, both knowledge-based and data-driven inferences are used for diagnosing yield-loss factors [31], for which a rule-based expert system grounded in knowledge-based inference generates a prime opportunity for selecting the appropriate tool-chamber feature. In the present study, the extraordinary process variables were identified with respect to their prior probabilities.

3.2. Data preparation

As shown in Fig. 2, a simplified but comprehensive spreadsheet sufficed for the diagnostics to pair the steps and tool-chamber features when N represented the total sample size of a huge amount of information (Table 1). The actual values are binary (i.e., true [1] or false [0]), indicating whether the tool-chamber feature was used in a step. This approach is effective when the technical problem of missing information is involved.

3.3. Data mining and key factor screening

The present study employed various types of statistical tools to wrap the associated variables, filter out trivial factors, and conduct key factor screening by executing the following steps:

Step 1. Cohen's kappa statistics for each pair of input variables

Cohen's kappa was used as a measure of agreement between two individuals (true or false) for each pair of predicted binary variables.

Step 2. Wrap the associated variables

Variables with a high level of agreement were wrapped with their peers in the same group. A variable can appear in more than one group. If a specified variable is designated as a critical factor at the variable selection level, then the corresponding stack(s) can be discerned as a suspected stack for further investigation.

Step 3. Assign the cutting point and low-, medium-, or high-yield wafers

For computing Cohen's kappa, a new dummy variable must be created to be used as an indicator of the wafer level to individualize the low, medium, or high yield values. To maintain the normality condition for this step, the data structure is divided into two subsets: the high yield and top 50% of the medium yield and the low yield and bottom 50% of the medium yield. The binary system is defined in a manner by which the medium group always obtains a zero value. The remaining data mining and key factor screenings of this framework are organized separately for both subsets.

Step 4. Cohen's kappa statistics for each pair of X and Y

Cohen's kappa is again applied to eliminate trivial variables, but, at this step, it is applied to each set of response and predictor variables.

Please cite this article in press as: M. Khakifirooz, et al., Bayesian inference for mining semiconductor manufacturing big data for yield enhancement and smart production to empower industry 4.0, Appl. Soft Comput. J. (2017), https://doi.org/10.1016/j.asoc.2017.11.034

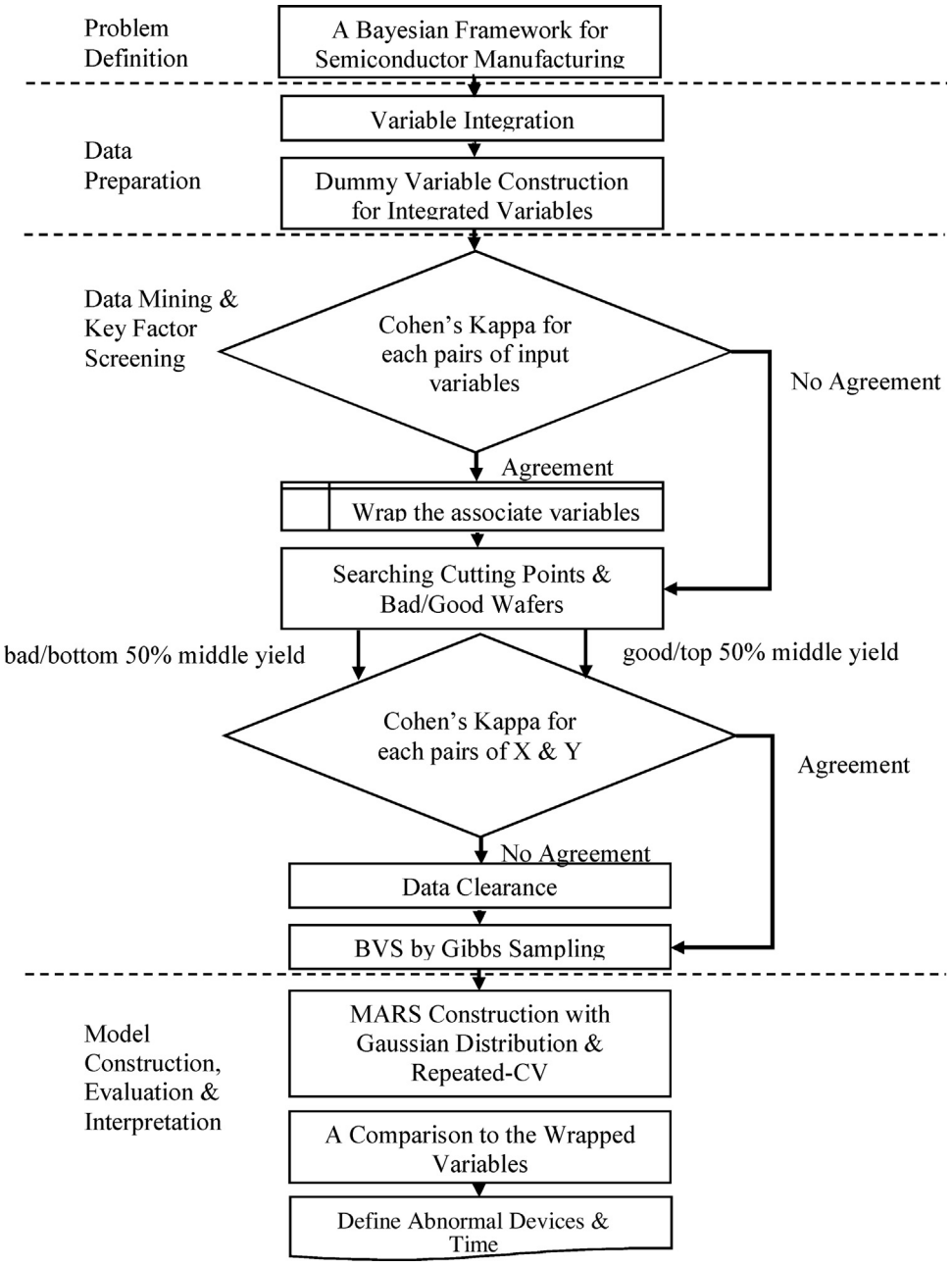

Fig. 2. Research Framework.

Table 1
Transformed sample data.

Wafer ID	step $1-\text{tool}1_1-\text{chamber}1_{1-1}$	...	step$i-\text{tool}i_{s_i}-\text{chamber}i_{s_i-ms_i}$	...	step$k-\text{tool}k_{s_k}-\text{chamber}k_{s_k-ms_k}$
w_1	0	...	1	...	1
...	...	...	...	...	...
w_N	1	...	0	...	1

Step 5 Data clearance

The previous step (Step 4) eliminates the predictor variables with a low level of agreement (point estimate of kappa statistic lower than 0.2).

Step 6. Bayesian variable selection through Gibbs sampler

To identify the variables that are included in the final model, Bayesian strategies were considered. Specifically, as mentioned in Section 2, for prior distribution, approaches based on Gibbs sampling were used.

3.4. Model construction, evaluation, and interpretation

In the final step, the multivariate adaptive regression spline (MARS) [32] algorithm with Gaussian distribution was employed to construct a proper model by using the selected key factors and cir-

Please cite this article in press as: M. Khakifirooz, et al., Bayesian inference for mining semiconductor manufacturing big data for yield enhancement and smart production to empower industry 4.0, Appl. Soft Comput. J. (2017), https://doi.org/10.1016/j.asoc.2017.11.034

cuit probing (CP) yield vector. Moreover, to evaluate the efficiency of the model, repeated-CV was conducted.

Next, to determine the level of accuracy of the proposed framework, the framework was assessed using the R-square and residual mean square error (RMSE) for the MARS method. The variables were selected using the random forest (RF) [33] or gradient boosting machine (GBM) [34]. Furthermore, the associated factors were recalled to obtain the impressionable group factors.

Finally, a decision table were constructed to analyze the outputs of the tool-chamber machines at each process step. The purpose of this phase was to explore the extensive process information to identify potential root causes for specifying the time cycle in the semiconductor manufacturing process.

By following the process framework, a simulation and empirical study was implemented to test the framework performance in the root cause detection of high and low yields. This approach reduced the cost and time required by the trial-and-error method.

4. Experimental design for the simulation study

The implementation of Gibbs sampling depends on the burn-in factor. Burn-in is defined as the number of sample generation iterations that are required for the samples to reach a stationary distribution [25]. Therefore, to develop the effect of prior probability and depict the performance of the Gibbs sampler for the proposed approach, a sequence of simulation studies was conducted. Conforming to the nature of a semiconductor fabrication data set, a small data set with a high dimensionality was considered. To facilitate the representation of the simulation setting and to mimic the real setting of batch-processing conditions in the manufacturing industry, the response variables and the observations of the categorical factors were simulated as follows.

Let $\boldsymbol{Y}$ denote a vector of continued response variable (i.e., yield percentage) and $\boldsymbol{Y}^c$ depict the categorical response variable (classified as high, medium, or low yield). In this case, $\boldsymbol{Y}^c$ is randomly produced by sampling three individuals (0, 1, 2) at random with a replacement, whereas $\boldsymbol{Y}$ is randomly generated from a uniform distribution based on the value of $\boldsymbol{Y}^c$, to set the distance [40,55), [55,70), and [70,85] as the lower and upper limits, respectively, of the uniform distribution for each low $\left(\boldsymbol{Y}^c = 0\right)$, medium $\left(\boldsymbol{Y}^c = 2\right)$, and high $\left(\boldsymbol{Y}^c = 1\right)$ class of the yield percentage.

In general, the nominal factors express the tool or chamber information used at each step (Fig. 1 and Table 1),

$$X_{il} = X_{i_{s_i}l} + X_{i_{s_i-ms_i}l}, \tag{1}$$

where X_{il} represents the nominal factors set as the tool-chamber information for each step; $X_{i_{s_i}l}$ and $X_{i_{s_i-ms_i}l}$ denote the tool and chamber information, respectively; and i and l indicate the ith process step and lth sample, respectively. Similarly, s represents the number of manufacturing tools in the ith step and m_{s_i} denotes the number of manufacturing chambers for the corresponding tools.

To assess the unrepeated set of manufacturing variables, duplicated nominal factors were eliminated.

For each process step, s and m were generated independently from a random integer in the range of 2–3, 3–4 or, 4–5 to enable the tool-chamber combination in the system to represent high to low level of multi-collinearity. Next, $\boldsymbol{X}_{i_{s_i}.}$ and $\boldsymbol{X}_{i_{s_i-ms_i}.}$ vectors were randomly and independently produced by sampling random integers in the range of s or m with a replacement. Finally, Phases 2 and 3 as well as Step 1 of Phase 4 of the proposed framework were executed using the simulated data to investigate the effect of prior probabilities and the performance of the Gibbs sampler.

The simulations were repeated for sample sizes of 50, 250, and 1000. A total of 100 process steps were executed when the total number of iterations performed after the burn-in process for the Gibbs sampler was equal to 1, 3, or 10 times of the variable size, and when the samples from the beginning of the chain (burn-in process) amounted to 10% of the total number of iterations.

To investigate the equality/inequality of the proposed approaches from the simulated data, the accuracy and inference efficiency were determined using the classified predicted and simulated yield categories of the models, with the data as the essential provender for accuracy. The simulated accuracy was measured by averaging more than 100 replications. A theoretically measured accuracy was produced using (2), which was applied to six methods. The results are presented in Fig. 3.

$$\text{Accuracy (ACC)} = \Sigma_{l=1}^{N} \mathbb{I}\left(\hat{\boldsymbol{Y}}_l^c = \boldsymbol{Y}_l^c\right)/N \tag{2}$$

Fig. 3 depicts the sample size and replication cycle and reveals that the accuracy of the Gibbs sampler with a lower iteration is higher than that of the others. A comparison of the three large-sample Gibbs methods shows that they had similar accuracy. However, the GBM had a higher accuracy compared with the other models when the number of observations was 1000. In other words, high dimensionality had a lesser effect on a large sample. Therefore, the Gibbs sampler requires a greater number of iteration cycles to obtain a convergence result. In a large sample, the difference among the performances of the five advanced methods was almost negligible. Furthermore, the accuracy of each method always improved in the absence of multi-collinearity (a greater number of components).

Because of multi-collinearity and high dimensionality, the Gibbs sampler is an effective tool for variable selection in a small sample with few iterations. Conversely, in a large sample, numerous iteration cycles are recommended. However, it does not guarantee superior results.

Moreover, the simulation study demonstrated that, in the case of collinearity, the GBM is typically more effective compared with RF that has been validated by several studies [35–38].

5. Empirical study

5.1. Problem definition

The problem in the present study involved 20 lots of 500 wafers, each with a yield vector as a response variable, and 100 process stages as predictor variables. Each lot passed through all the stages. As shown in Fig. 4, this problem induces both high and low CP yield in a fab. The engineers were required to retrieve related fabrication data, which contained considerable variety and complexity, to identify the root causes, and make inadequate tool-chamber being aligned with the golden ones.

5.2. Data preparation

To conform the framework and manage nominal factors, the 100 process stages transfer to dummy variables. The transformed data included 1988 factors, each of which comprised the history of the stages, tools, and chambers.

Because the raw data contained numerous missing elements, data was prepared by including the imputation of missing elements, until 1460 factors remained.

5.3. Data mining and key factor screening

After defining the problem and preparing the data, Cohen's kappa statistics and the Gibbs sampler were used to identify abnormal process stages and machines, which could serve as a reference for engineers to troubleshoot and diagnose defects.

Please cite this article in press as: M. Khakifirooz, et al., Bayesian inference for mining semiconductor manufacturing big data for yield enhancement and smart production to empower industry 4.0, Appl. Soft Comput. J. (2017), https://doi.org/10.1016/j.asoc.2017.11.034

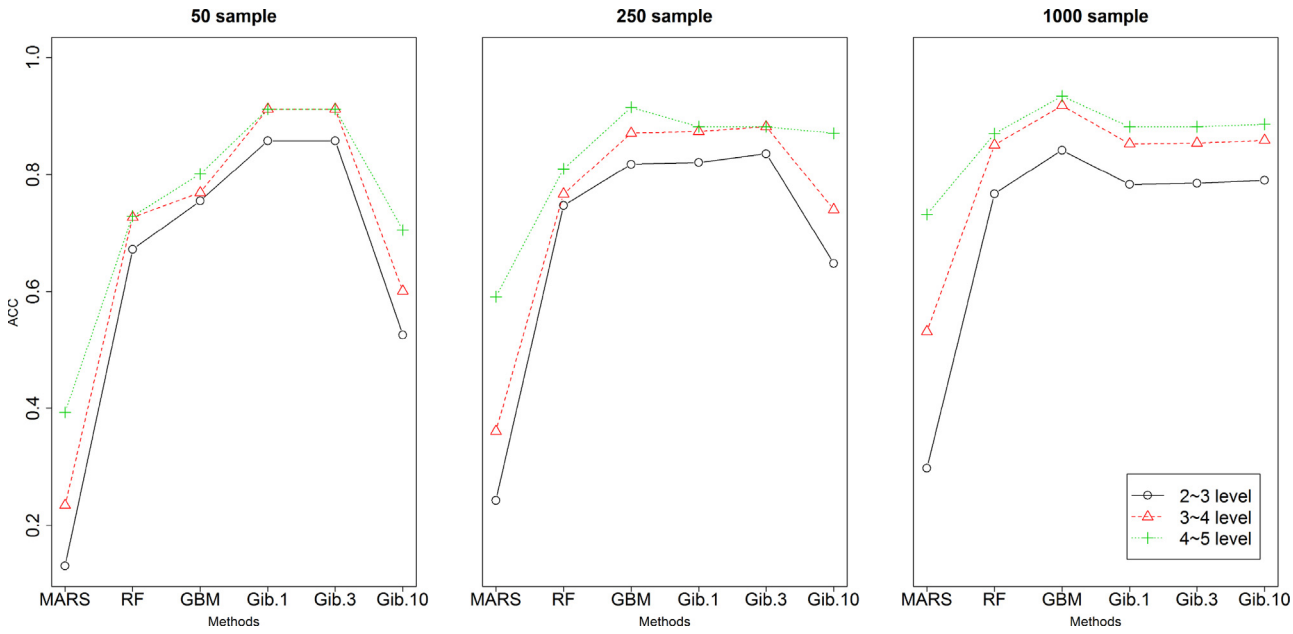

Fig. 3. Accuracy of simulation cases. Three notation of x-axis "Gib.1", "Gib.3" and "Gib.10" correspond to results of different iteration times for Gibbs sampler.

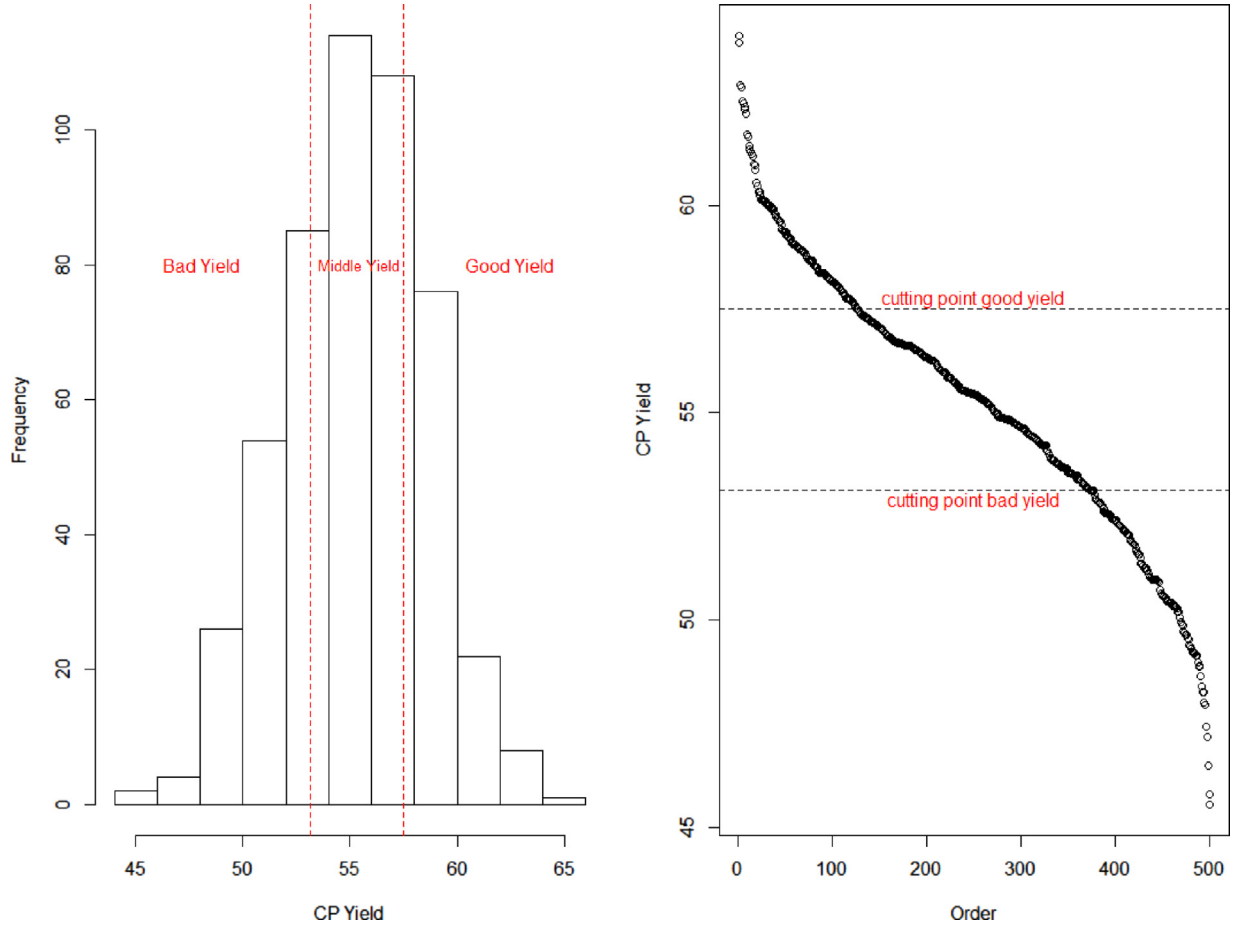

Fig. 4. Scatter and histogram plot of sorted CP yield.

Table 2
The class distribution for the Kappa test for each pairs of input variables.

Almost perfect agreement	Substantial agreement	Moderate agreement	Fair agreement	Slight agreement	No agreement
3[a]	109	1764	24,539	280,081	758,574

[a] Number of pairs at each level of agreements.

Please cite this article in press as: M. Khakifirooz, et al., Bayesian inference for mining semiconductor manufacturing big data for yield enhancement and smart production to empower industry 4.0, Appl. Soft Comput. J. (2017), https://doi.org/10.1016/j.asoc.2017.11.034

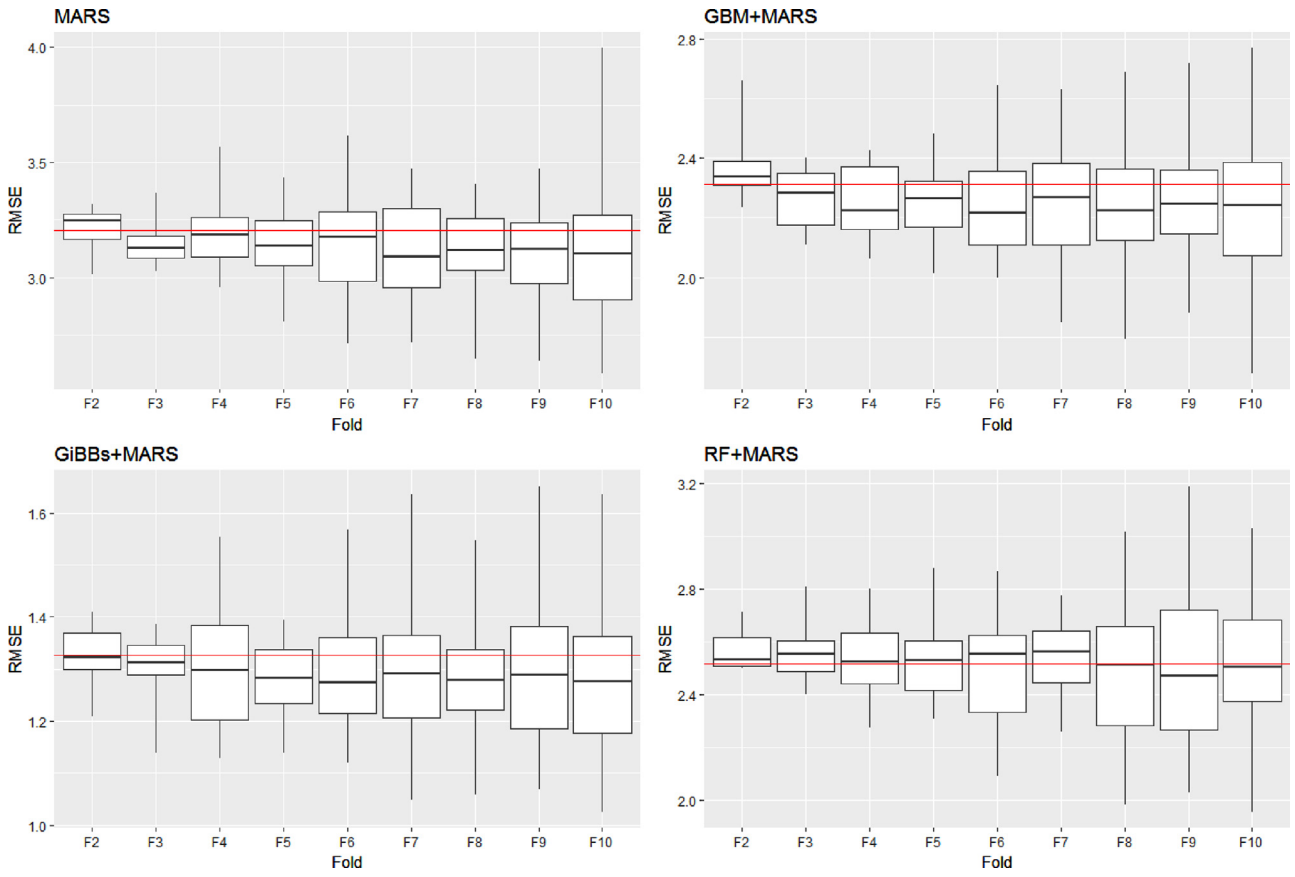

Fig. 5. Total prediction error (red) and cross-validation curve for comparing the distribution of RMSE for constructed models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Phase 1: In total, 1,065,070 Cohen's kappa statistics were computed for each pair of input variables. The distribution of the kappa attributes is listed in Table 2.

Phase 2: For the normal distribution of data (Fig. 4), the 25th and 75th quartiles of the CP yield were exploited as cutting points to classify the yields. The medium, low, and high groups contained 250, 125, and 125 wafers, respectively.

To avoid a class imbalance problem and address hidden skewed data without any loss of generality, the data set was itemized into two binary groups: Group 1 comprised all high-yield wafers and the top 50% of the medium yield and Group 2 comprised all low-yield wafers and the remainder of the medium yield. Phases 3 and 4 of this step were applied simultaneously to the two groups.

For convenience in analysis, a new variable containing the yield groups was created. The wafers in the medium group were marked as 0, and those in the low and high groups were marked as 1. The cutting points are shown in Fig. 4. The cutting points were at 53.12% and 57.51% of the yield rate. These marks enabled to more clearly distinguish if the same process stage fabricated the low-yield and high-yield wafers.

Phase 3: Because an excessive number of process factors existed, kappa statistics were used to narrow the number of factors in this step. The kappa was applied to identify potential process factors with an appropriate measure of reliability. Similar to Phase 1, the kappa statistic was used to compare the rating of the grouped yield with each dummy variable, thus eliminating the influence of the root-cause factors. The process variables with a moderate and low level of agreement were removed. At the end of this phase, 198 and 175 predictor variables were identified as inputs for the next step from each high-yield and low-yield group, respectively, of the sub-dataset.

Phase 4: To apply Bayesian inference by using the Gibbs sampler, the well-crafted BayesVarSel R package [39] was used. Subsequently, prior probabilities were estimated from the sample frequencies for each variable, as follows:

$$p_{i_{s_i-ms_i}} = \sum_{l=1}^{N} I\left(X_{i_{s_i-ms_i}l} = 1\right)/N, \tag{3}$$

where N denotes the sample size and $\sum_{s,m} p_{i_{s_i-ms_i}} = 1$ for $i = 1, \ldots, k$. Implementing a Gibbs sampler in each sub-data set from the binary expression of the most probable model yielded 10 predictor variables with a higher estimated inclusion probability for model construction.

Eventually, before the overall model was constructed, the high-yield and low-yield groups were merged again.

5.4. Model construction, evaluation, and interpretation

The MARS was employed for identifying any suspicious nonlinear relationship between the predictors and the response variable, in which the Gaussian family accepted the identification of the response variable. To evaluate the effectiveness and practical viability of the proposed approach, three other conventional approaches, namely the GBM, RF, and simple MARS, were selected for comparison. The number of critical variables of the GBM and RF was chosen to be equal to the selected variable by using the Gibbs sampler. Through the adopted repeated-CV method, RMSE and adjusted R-squared were used as the evaluation criteria for the training data. A "one-standard error" rule is used with cross-validation, in which we choose the most parsimonious model whose error is no more

Please cite this article in press as: M. Khakifirooz, et al., Bayesian inference for mining semiconductor manufacturing big data for yield enhancement and smart production to empower industry 4.0, Appl. Soft Comput. J. (2017), https://doi.org/10.1016/j.asoc.2017.11.034

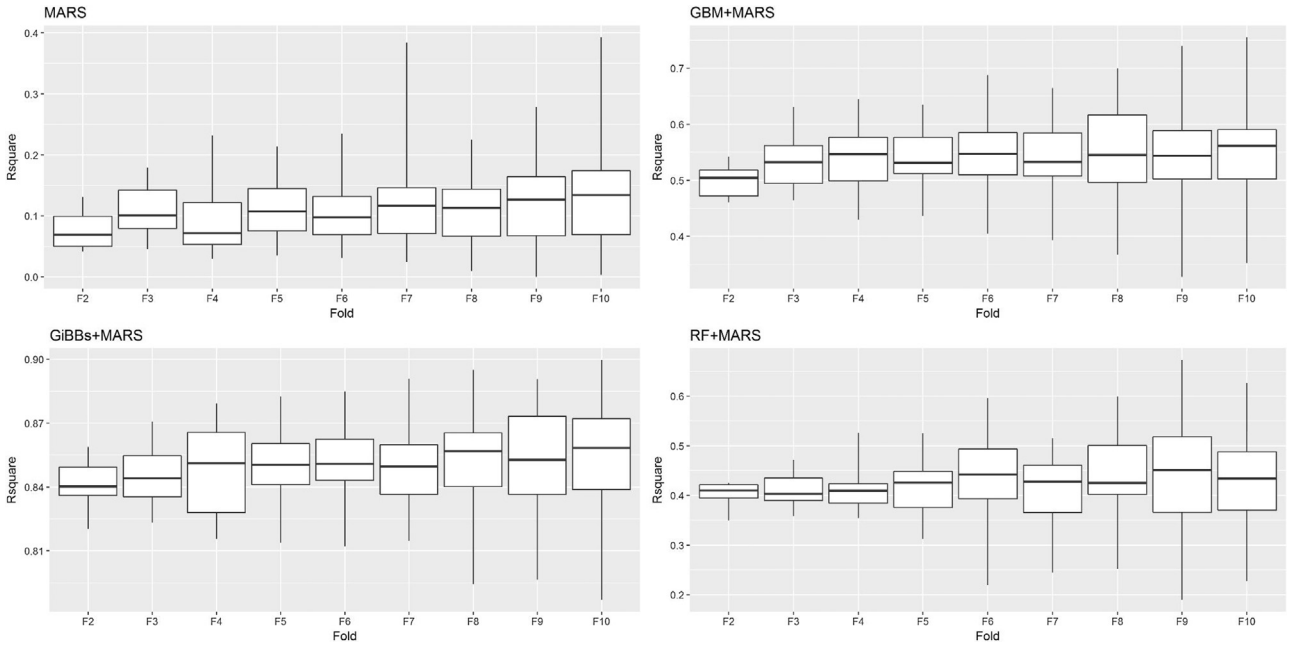

Fig. 6. Total prediction error (red) and cross-validation curve for comparing R-squares of constructed models.

Table 3
Telic decision table, core structure.

Factors	Date	
	Bad	Good
Stage10 – Tool2 – Chamber3	before 8/29 2:32	after 8/29 12:50
Stage12 – Tool2 – Chamber1	after 8/30 3:26 till 8/30 3:43	before 8/29 10:55
Stage12 – Tool2 – Chamber4	after 8/29 7:36 till 8/30 3:44	before 8/29 7:36
Stage13 – Tool5 – Chamber2	–	generally effected the high yield
Stage17 – Tool2 – Chamber2	after 8/30 12:21	before 8/30 10:37
Stage23-Tool3-Chamber2	–	generally effected the high yield
Stage44 – Tool7.- Chamber2 and Chamber3	at 9/3	at 9/1
Stage49 – Tool1.- Chamber4	at 9/3	at 9/2
Stage57 – Tool1.- Chamber3	–	generally effected the high yield

than one standard error above the error of the complete model. The results are summarized in Fig. 5 and Fig. 6. The evidence proves that at least 5-fold cross-validation will overestimate the true prediction error for the proposed method. A comparison of the sampling distributions for the four models revealed that, in this case, the MARS combined with a Bayesian variable selection technique was favorable.

To validate the proposed framework, identify its advantages, and evaluate the reliability of the selected variables, the variables were reverted to the associated predictor variables by using Cohen's kappa and compared, which revealed three pairs of substantial agreements, 74 pairs of moderate agreements, 500 pairs of fair agreements, and 6528 pairs of slight agreements between the selected and other predictor variables.

All factors with moderate and higher kappa levels were considered as potential critical factors that could influence the yield enhancement.

The final stage of the framework involved mining the critical time spots. Table 3 shows a portion of the final critical rules was summarized as a core structure for decision making; the supporting time series for the rules selected using the Gibbs sampler as well as the critical factors selected using Cohen's kappa. The problem or perfection was detected at a certain cycle time when a relatively large number of wafers deviated significantly at the high- or low-yield scope.

6. Conclusion

The detailed historical data recorded online during the manufacturing of chips necessitates the application of data mining and big data analytics [7]. Many researchers have reported their efforts with notable advancements mostly in detection research for identifying the cause of an unresolved problem and certain specialized samples have been collected for identifying root causes [7,40–43].

In addition to data preparation, data clearance and variable selection are critical steps in the data-mining process, although they are extremely time consuming. However, they cannot be ignored and require a considerable amount of patience.

The proposed framework combined a Bayesian approach with statistical inferences and the perspective of data-mining, which enabled this approach to explore complex semiconductor manufacturing data. On the basis of the empirical results, the feasibility of the proposed approach was validated, indicating that integrating domain knowledge and experience into a system can improve the results. Furthermore, using domain knowledge alone can result in restricted conjunctions in the rules for tools, chambers, and steps that are related or that occur within a reasonable time frame.

The proposed approach is flexible for managing numerous factors of collinearity and high dimensionality. In addition, the simulation results validated its effectiveness.

For a large sample, the Gibbs sampler is potentially time consuming, and its convergence to a true distribution is extremely

Please cite this article in press as: M. Khakifirooz, et al., Bayesian inference for mining semiconductor manufacturing big data for yield enhancement and smart production to empower industry 4.0, Appl. Soft Comput. J. (2017), https://doi.org/10.1016/j.asoc.2017.11.034

slow. However, this phenomenon has a negligible effect on semiconductor manufacturing data because of the cost and expense associated with developing and maintaining highly complex production systems that prevent the engineer from being overwhelmed by excess data. A recent study reported the development of an enhanced Gibbs sampler algorithm for faster interactions [44].

Moreover, this technical analysis revealed that the Gibbs sampler is a method with a higher explanatory power and fewer critical explanatory variables compared with the RF and GBM.

For the purpose of model construction, regards to our consideration about accuracy, consistency and processing time of model construction, methods with a heavy cost for computation time have not been selected. To pay the heed on the bad and good yield, forwarded the feature selection step, the class of medium yield has not been tightened. Therefore, due to a lack of homogeneity and similarity in this non-tight class, the discriminant techniques such as principle component analysis (PCA) or neural networks have not been utilized. The major outliers (very good and very bad yield) were in the center of attention in this study, ignoring, skipping or excluding them from the final model, has been reduced the consistency of our model, hereupon, to make an unbiased model for outliers, techniques like, support vector machines (SVM) or *k*-nearest neighbor (*k*-NN) have been avoided. Indeed, the consistency of the proposed models is enhanced by dealing with either mixed- or non-linearity issues, while all methods constructed on the linear characteristics such as generalized linear model (GLM) and partial least square (PLS) methods have been declined.

6.1. Effect modification and interactions

The present study addressed multi-collinearity diversity within each process step. However, association or collinearity also arises between the flow steps in semiconductor manufacturing from the set-up time or internal layout. In addition, the effect of one covariate on that of another covariate warrants further investigation.

6.2. Noise variables in manufacturing

Process engineers can freely select from process variables to effectively satisfy performance requirements. However, noise variables are uncontrollable effects that process engineers cannot directly control (e.g., material property variations and variable costs resulting in certain limitations). Because engineers cannot select the values of these uncontrollable random or operational effects, all values must be carefully considered in design calculations, thus appropriately accounting for the effect of noise in decisions [45]. Therefore, to determine the impact of uncontrollable effects on design and manufacturing processes, investigating fuzzy measures is warranted in the future to identify the effect of noise variables. This technique is comparable to a stochastic process with random effects.

Acknowledgements

This research is supported by Ministry of Science and Technology, Taiwan (MOST 105-2218-E-007-027; MOST 105-2622-8-007-002-TM1). The authors appreciate the constructive comments and advice from the anonymous reviewers.

References

[1] L. Argote, D. Epple, Learning curves in manufacturing, Science 247 (4945) (1990) 920–924, http://dx.doi.org/10.1126/science.247.4945.920.
[2] N.W. Hatch, D.C. Mowery, Process innovation and learning by doing in semiconductor manufacturing, Manage. Sci. 44 (11) (1998) 1461–1477, http://dx.doi.org/10.1287/mnsc.44.11.1461.
[3] F. Zoua, L. Wanga, X. Hei, D. Chen, Teaching–learning-based optimization with learning experience of other learners and its application, Appl. Soft Comput. 37 (2015) 725–736, http://dx.doi.org/10.1016/j.asoc.2015.08.047.
[4] C.-F. Chien, W.-C. Wang, J.-C. Cheng, Data mining for yield enhancement in semiconductor manufacturing and an empirical study, Expert Syst. Appl. 33 (1) (2007) 192–198, http://dx.doi.org/10.1016/j.eswa.2006.04.014.
[5] C.-F. Chien, Y.-J. Chen, C.-Y. Hsu, H.-K. Wang, Overlay error compensation using advanced process control with dynamically adjusted proportional-integral R2R controller, IEEE Trans. Autom. Sci. Eng. 11 (2) (2014) 473–484, http://dx.doi.org/10.1109/TASE.2013.2280618.
[6] Y.-C. Chang, C. Mastrangelo, Addressing multicollinearity in semiconductor manufacturing, Qual. Reliab. Eng. Int. 27 (6) (2011) 843–854, http://dx.doi.org/10.1002/qre.1173.
[7] C.-F. Chien, S.-C. Chuang, A framework for root cause detection of sub-batch processing system for semiconductor manufacturing big data analytics, IEEE Trans. Semicond. Manuf. 27 (4) (2014) 475–488, http://dx.doi.org/10.1109/TSM.2014.2356555.
[8] S.-C. Hsu, C.-F. Chien, Hybrid data mining approach for pattern extraction from wafer bin map to improve yield in semiconductor manufacturing, Int. J. Prod. Econ. 107 (1) (2007) 88–103, http://dx.doi.org/10.1016/j.ijpe.2006.05.015.
[9] C.-F. Chien, C.-Y. Hsu, Data mining for optimizing IC feature designs to enhance overall wafer effectiveness, IEEE Trans. Semicond. Manuf. 27 (1) (2014) 71–82, http://dx.doi.org/10.1109/TSM.2013.2291838.
[10] T. Chen, Forecasting the yield of a semiconductor product with a collaborative intelligence approach, Appl. Soft Comput. 13 (3) (2013) 1552–1560, http://dx.doi.org/10.1016/j.asoc.2012.01.003.
[11] C.-W. Liu, C.-F. Chien, An intelligent system for wafer bin map defect diagnosis: an empirical study for semiconductor manufacturing, Eng. Appl. Artif. Intell. 26 (5–6) (2013) 1479–1486, http://dx.doi.org/10.1016/j.engappai.2012.11.009.
[12] C.-F. Chien, C.-W. Liu, S.-C. Chuang, Analysing semiconductor manufacturing big data for root cause detection of excursion for yield enhancement, Int. J. Prod. Res. 55 (17) (2017) 5095–5107, http://dx.doi.org/10.1080/00207543.2015.1109153.
[13] I.L.-G. Chong, S.L. Albin, C.-H. Jun, A data mining approach to process optimization without an explicit quality function, IIE Trans. 39 (8) (2007) 795–804, http://dx.doi.org/10.1080/07408170601142668.
[14] I.G. Chong, C.H. Jun, Performance of some variable selection methods when multicollinearity is present, Chemom. Intell. Lab. Syst. 78 (1) (2005) 103–112, http://dx.doi.org/10.1016/j.chemolab.2004.12.011.
[15] A.B. Kahng, B. Lin, S. Nath, Enhanced metamodeling techniques for high-dimensional IC design estimation problems, in: Proc. DATE, Grenoble, France, 2013, pp. 1861–1866, http://dx.doi.org/10.7873/DATE.2013.371.
[16] C.-F. Chien, Y.-J. Chen, C.-Y. Hsu, A novel approach to hedge and compensate the critical dimension variation of the developed-and-etched circuit patterns for yield enhancement in semiconductor manufacturing, Comput. Oper. Res. 53 (2015) 309–318, http://dx.doi.org/10.1016/j.cor.2014.05.009.
[17] X.W. Chen, X. Lin, Big data deep learning: challenges and perspectives, IEEE Access. 2 (2014) 514–525, http://dx.doi.org/10.1109/ACCESS.2014.2325029.
[18] A. Tan, J. Huang, Bayesian inference for high-dimensional linear regression under mnet priors, Can. J. Stat. 44 (2) (2016) 180–197, http://dx.doi.org/10.1002/cjs.11283.
[19] Z. Chen, A. Mukherjee, B. Liu, Aspect extraction with automated prior knowledge learning, in: Proc. 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, USA, 2014, pp. 347–358, http://dx.doi.org/10.3115/v1/P14-1033.
[20] C.-F. Chien, K.-H. Chang, W.-C. Wang, An empirical study of design-of-experiment data mining for yield-loss diagnosis for semiconductor manufacturing, J. Intell. Manuf. 25 (5) (2014) 961–972, http://dx.doi.org/10.1007/s10845-013-0791-5.
[21] K.P. Murphy, Machine Learning: A Probabilistic Perspective, MIT press, London, 2012, pp. 35–36, http://dx.doi.org/10.1080/09332480.2014.914768.
[22] C. Andrieu, N. de Freitas, A. Doucet, M.I. Jordan, An introduction to MCMC for machine learning, Mach. Learn. 50 (1) (2003) 5–43, http://dx.doi.org/10.1023/A:1020281327116.
[23] W.R. Gilks, S. Richardson, D.J. Spiegelhalter, Markov Chain Monte Carlo in Practice, Chapman and Hall, London, 1996, pp. 204–211, http://dx.doi.org/10.1002/(SICI)1097-0258(19980615)17:11<1301:AID-SIM882>3.0.CO;2-9.
[24] E.I. George, R.E. McCulloch, Variable selection via Gibbs sampling, J. Am. Stat. Assoc. 88 (423) (1993) 881–889, http://dx.doi.org/10.1080/01621459.1993.10476353.
[25] G. García-donato, M.A. Martínez-beneito, On sampling strategies in Bayesian variable selection problems with large model spaces, J. Am. Stat. Assoc. 108 (501) (2013) 340–352, http://dx.doi.org/10.1080/01621459.2012.742443.
[26] C. Hu, H. Cao, Aspect-level influence discovery from graphs, IEEE Trans. Knowl. Data Eng. 28 (9) (2016) 1635–1649, http://dx.doi.org/10.1109/TKDE.2016.2538223.
[27] O. Amayri, N. Bouguila, A Bayesian analysis of spherical pattern based on finite Langevin mixture, Appl. Soft Comput. 38 (2016) 373–383, http://dx.doi.org/10.1016/j.asoc.2015.10.024.
[28] B. Das, N.C. Krishnan, D.J. Cook, RACOG and wRACOG: two probabilistic oversampling techniques, IEEE Trans. Knowl. Data Eng. 27 (1) (2015) 222–234, http://dx.doi.org/10.1109/TKDE.2014.2324567.
[29] L.I. Kuncheva, A bound on Kappa-error diagrams for analysis of classifier ensembles, IEEE Trans. Knowl. Data Eng. 25 (3) (2013) 494–501, http://dx.doi.org/10.1109/TKDE.2011.234.

Please cite this article in press as: M. Khakifirooz, et al., Bayesian inference for mining semiconductor manufacturing big data for yield enhancement and smart production to empower industry 4.0, Appl. Soft Comput. J. (2017), https://doi.org/10.1016/j.asoc.2017.11.034

[30] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed., Springer, Verlag, New York, 2009, pp. 241–249, http://dx.doi.org/10.1007/978-0-387-84858-7.
[31] C.-M. Fan, Y.-P. Lu, A Bayesian framework to integrate knowledge-based and data-driven inference tools for reliable diagnoses, in: Proc. WSC, Austin, TX, 2008, pp. 2323–2329, http://dx.doi.org/10.1109/WSC.2008.4736337.
[32] J. Friedman, Multivariate adaptive regression spline, Ann. Stat. 19 (1) (1991) 1–141, http://dx.doi.org/10.1214/aos/1176347963.
[33] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, http://dx.doi.org/10.1023/A:1010933404324.
[34] J. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 29 (5) (2001) 1189–1232, http://dx.doi.org/10.1214/aos/1013203451.
[35] A. Knudby, A. Brenning, E. LeDrew, New approaches to modelling fish–habitat relationships, Ecol. Modell. 221 (3) (2010) 503–511, http://dx.doi.org/10.1016/j.ecolmodel.2009.11.008.
[36] G.L. Simpson, H. John, B. Birks, Statistical learning in palaeolimnology, Dev. Paleoenviron. Res. 5 (2012) 249–327, http://dx.doi.org/10.1007/978-94-007-2745-8_9.
[37] A. Knudby, E. LeDrew, A. Brenning, Predictive mapping of reef fish species richness, diversity and biomass in Zanzibar using IKONOS imagery and machine-learning techniques, Remote Sens. Environ. 114 (6) (2010) 1230–1241, http://dx.doi.org/10.1016/j.rse.2010.01.007.
[38] R.A. Viscarra Rossela, T. Behrens, Using data mining to model and interpret soil diffuse reflectance spectra, Geoderma 158 (1–2) (2010) 46–54, http://dx.doi.org/10.1016/j.geoderma.2009.12.025.
[39] G. Garcia-Donato, A. Forte, BayesVarSel: Bayes Factors, Model Choice and Variable Selection in Linear Models, R Package Version 1.7.0, 2016 https://CRAN.R-project.org/package=BayesVarSel.
[40] D.-H. Baek, I.-J. Jeong, C.H. Han, Application of data mining for improving yield in wafer fabrication system, in: Proc. ICCSA, Singapore, 2005, pp. 222–231, http://dx.doi.org/10.1007/11424925_25.
[41] S.M. Weiss, R.J. Baseman, F. Tipu, C.N. Collins, W.A. Davies, R. Singh, J.W. Hopkins, Rule-based data mining for yield improvement in semiconductor manufacturing, Appl. Intell. 33 (1) (2010) 318–329, http://dx.doi.org/10.1007/s10489-009-0168-9.
[42] W. Yamwong, T. Achalakul, Yield improvement analysis with parameter-screening factorials, Appl. Soft Comput. 12 (2012) 1021–1040, http://dx.doi.org/10.1016/j.asoc.2011.11.021.
[43] C.-F. Chien, C.-Y. Hsu, P.-L. Chen, Semiconductor fault detection and classification for yield enhancement and manufacturing intelligence, Flexible Serv. Manuf. J. 25 (3) (2013) 367–388, http://dx.doi.org/10.1007/s10696-012-9161-4.
[44] R.Y. Rubinstein, Ad Ridder, R. Vaisman, Fast Sequential Monte Carlo Methods for Counting and Optimization, Wiley, Hoboken, New Jersey, 2013, pp. 104–105, http://dx.doi.org/10.1002/9781118612323.
[45] H.-J. Sebastian, E.K. Antonsson, Fuzzy Sets in Engineering Design and Configuration, Springer, Pasadena, CA, 2012, pp. 2–36, http://dx.doi.org/10.1007/978-1-4613-1459-2.

Please cite this article in press as: M. Khakifirooz, et al., Bayesian inference for mining semiconductor manufacturing big data for yield enhancement and smart production to empower industry 4.0, Appl. Soft Comput. J. (2017), https://doi.org/10.1016/j.asoc.2017.11.034