Computers & Operations Research 53 (2015) 328–337

Contents lists available at ScienceDirect

Computers & Operations Research

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

Multi-step virtual metrology for semiconductor manufacturing: A multilevel and regularization methods-based approach

Gian Antonio Susto [a,*], Simone Pampuri [b], Andrea Schirru [b], Alessandro Beghi [a], Giuseppe De Nicolao [b]

[a] Department of Information Engineering, University of Padova, via G. Gradenigo 6/B, 35131 Padova, Italy
[b] Department of Computer Engineering and Systems Science, University of Pavia, via Ferrata 1, 27100 Pavia, Italy

ARTICLE INFO

Available online 17 May 2014

Keywords:
Chemical vapor deposition
Etching
Industry automation
LASSO
Lithography
Regularization methods
Ridge regression
Semiconductor manufacturing
Statistical modeling
Virtual metrology

ABSTRACT

In semiconductor manufacturing, wafer quality control strongly relies on product monitoring and physical metrology. However, the involved metrology operations, generally performed by means of scanning electron microscopes, are particularly cost-intensive and time-consuming. For this reason, in common practice a small subset only of a productive lot is measured at the metrology stations and it is devoted to represent the entire lot. Virtual Metrology (VM) methodologies are used to obtain reliable predictions of metrology results at process time, without actually performing physical measurements. This goal is usually achieved by means of statistical models and by linking process data and context information to target measurements. Since semiconductor manufacturing processes involve a high number of sequential operations, it is reasonable to assume that the quality features of a given wafer (such as layer thickness and critical dimensions) depend on the whole processing and not on the last step before measurement only. In this paper, we investigate the possibilities to enhance VM prediction accuracy by exploiting the knowledge collected in the previous process steps. We present two different schemes of multi-step VM, along with dataset preparation indications. Special emphasis is placed on regression techniques capable of handling high-dimensional input spaces. The proposed multi-step approaches are tested on industrial production data.

1. Introduction

In recent years, Virtual Metrology (VM) techniques have received growing interest from semiconductor manufacturers, thanks to the prospective measurement cost reduction and improvements in production quality (by means of control schemes exploiting VM information) [1].

The goal of a VM module is that of defining the relationships between *process data* (*input*) and *metrology data* (*output*). Given the cost of metrology operations and the increasing availability of recorded data in modern equipment, reliable VM predictions are used in place of real metrology measurements [2,3]. The inputs of the VM algorithms are cost-free data like sensor data, logistic and recipe information, while the predicted output is generally critical dimensions (like layer thickness for Chemical Vapor Deposition, Etch depth of Etch Rate for the Etching) upon which the goodness of the performed process can be assessed. In this perspective, VM tools are seen as information providers, able to yield probabilistic information at process time on wafer quality features.

Thanks to the diffusion in the pasts years of VM modules and the improvement of their prediction accuracy, nowadays VM predictions are not only used to monitor process quality and to decrease the number of physical measures performed, but they are also exploited by intelligent tools like controllers [4,5], dispatching and sampling decision systems [6] that can take advantage of VM estimations to improve the overall process quality.

VM problems, and more in general, modeling of semiconductor manufacturing process quality features, pose a number of challenges, among which the most prominent are the following:

- *High-dimensionality*: The number of potential input process parameters is usually large, given the high number of process variables and even higher number of collected data/statistics and production information. This issue may lead to ill-conditioned problems and data over-fitting [7–9].
- *Data fragmentation*: The typical semiconductor manufacturing production is highly fragmented. Hundreds of different products are processed with different machine settings (*recipes*) on several tools that work in parallel, each one with different

* Corresponding author.
E-mail addresses: gianantonio.susto@dei.unipd.it (G.A. Susto), simone.pampuri@unipv.it (S. Pampuri), andrea.schirru@unipv.it (A. Schirru), beghi@dei.unipd.it (A. Beghi), giuseppe.denicolao@unipv.it (G. De Nicolao).

http://dx.doi.org/10.1016/j.cor.2014.05.008
0305-0548/

working stations (*chambers*) (see Fig. 1 for an example). A VM system is required to model the entire production, but separately modeling each *logistic path* (group of wafers with the same combination of recipe, tool and chamber) is unfeasible given the large amount of possible combinations versus the historical data available.

- *Multi-processes influence*: The information regarding the outcome of a process is related to both the process itself and previous steps along the production line that may contain information regarding the current state of the wafer or may physically affect the outcome of the wafer feature in exam. For instance, one physical reason of the superimposition of effects of multiple processes on wafer features is the fact that wafer fabrication is based on multiple layers build one on top of the previous, with a possible concatenation of effects due to layer surface disparities [10].

The first two issues are addressed in Section 2, where a brief review of modeling techniques for VM is given. The main focus of the paper is however on the last issue, namely, the influence of multiple processes on the wafer features predicted by the VM module (the *VM targets*), that has been only partly explored in the VM literature [11,12]. Classical VM modules typically consider the modeling of a single process only, that is, the last one before the physical metrology step, without taking into account the influence of the previous processes on the line may have on the physical/electrical parameters that the VM module aims to predict. If data regarding the previous processes can be retrieved and included in the input set, it is reasonable to expect that the VM systems prediction accuracy can be enhanced.

The resulting data collection problem is a difficult one. In fact, from the modeling point of view, the collected multi-step data more markedly present the aforementioned issues of high-dimensionality and fragmentation. The increase in dimensionality is clearly related to the inclusion of a larger number of parameters into the dataset, that are related to the previous processing steps. To illustrate the issue of data fragmentation, consider the example of Fig. 2 that regards three of the most important classes of semiconductor processes, namely, Chemical Vapor Deposition (CVD), Lithography (Litho), and Etching (Etch). The diagram represents a possible process flow in the case of 11 different work-stations for the aforementioned processes. Since different process tools can perform the same process step for a specific wafer, the number of possible paths grows exponentially with the number of steps. As a consequence, a homogeneous dataset referred to a specific path would comprise an insufficient number of observations.

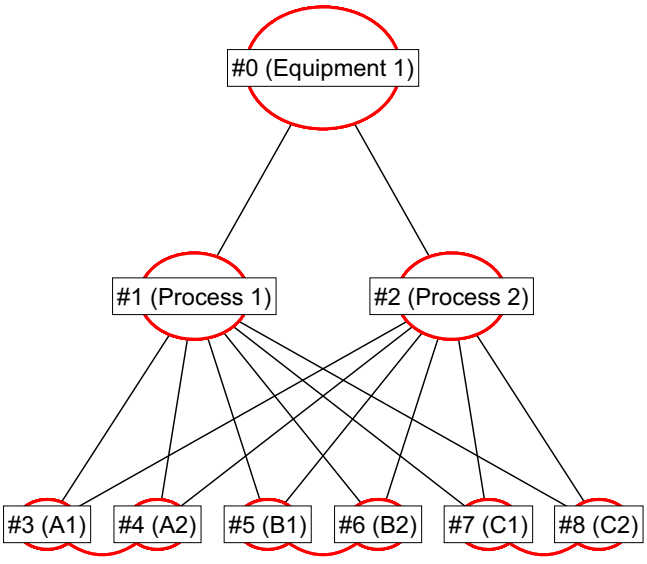

Fig. 1. Tree representation of a CVD (Chemical Vapor Deposition) tool with three chambers (A, B, C) with two subchambers each (1 and 2), involved in two processes (Process1 and Process 2). Therefore, for the processed wafers, twelve distinct logistic configurations (i.e., paths) are possible.

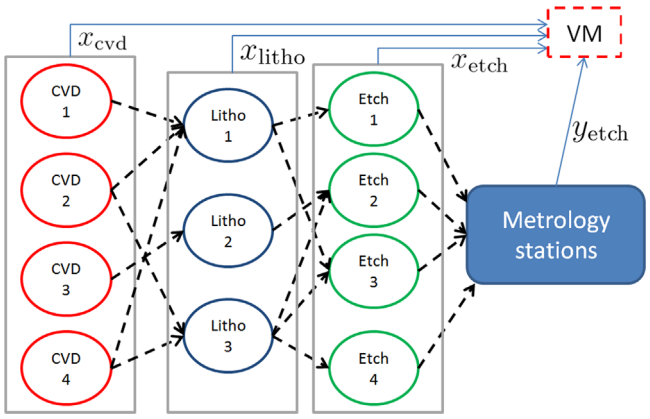

Fig. 2. Example of process flow in semiconductor manufacturing: the black dashed lines represent wafer dispatching events, while the solid blue lines represent information flows. The Virtual Metrology (VM) block collects process data (x) for several consecutive steps, and metrology data (y) for the latest step. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

In this paper we present a novel framework to address multi-step VM situations similar to the one shown in Fig. 2. The proposed approach relies on regularized machine learning methodologies to deal with high-dimensionality, and on a multilevel transformation of the input space to deal with data fragmentation. The goal is that of estimating quality indicators of wafers that have undergone several processes, by making use of data related to a subset of those processes that may have influenced the VM target (based on data availability and a priori physical knowledge).

The paper is organized as follows:

- Section 2.2 is devoted to review regularized machine learning techniques with focus on Regularization Methods.
- In Section 3 a brief description of Multi-Level techniques is provided and the proposed Multi-Step approaches are presented, in terms both of dataset preparation and model assumptions.
- In Section 4 a user case is presented and the proposed methodologies are validated exploiting a industrial manufacturing dataset.

Finally, in Section 5, final remarks and comments are provided. This paper extends the results presented in [13].

2. Modeling techniques for VM

In this section, the basic features of the modeling techniques employed for VM technologies are reviewed.

2.1. Literature review

Several features are required for a VM system to be successfully employed in a production environment (i.e. scalability to new production settings, fast computation, interpretability). Among them, prediction accuracy is the first and most important one, and consequently, the issue of modeling for VM has been at the heart of the debate in the scientific community in the past years.

Given the high complexity of the semiconductor processes under investigation and the difficulties of physical based models to capture all the relationships among the many parameters involved, black-box approaches are usually taken for modeling. Several statistical modeling and machine learning approaches, both linear [14] and non-linear [3], have been tested and compared in the literature in the past recent years.

There is a growing number of VM modules as more and more new machine learning techniques for regression become available, however, criticality's still exist, such as those associated with high-dimensionality and data fragmentation, that are particularly relevant when multiple processes are considered.

To deal with data fragmentation, few methods have been proposed in the literature with regard to semiconductor manufacturing applications, such as:

- smart clustering (information theory [15] or PCA-based [2] for example);
- multi-level techniques [16,17] (detailed in Section 3).

More attention has been given instead to the issue of high dimensionality, the main problem being how to handle VM problems with hundreds or thousands of variables while avoiding overfitting and ill-conditioning. A widely adopted approach for dealing with this issue is a 2-step procedure composed by the following:

1. *variable selection/model size reduction* – the number of variables is reduced by retaining those highlighted as important by process engineers, those expressing the maximum amount of variability in the input set [18], or by applying Principal Component Analysis (PCA) and retaining the first principal components only [3];
2. *modeling* – the regression algorithms are then applied to the reduced dataset.

This approach allows complex and non-linear techniques to be employed for modeling, like Artificial Neural Networks [19]. However, reduction of the input dataset may lead to suboptimal results and it is generally a time consuming procedure, thus in general preventing on-line recomputation, as often required in industrial VM systems.

Other techniques deal with high-dimensionality directly within the modeling step, like

- *regularization methods* (Ridge Regression (RR) [20], LASSO [14] and Elastic Nets [12]), that impose a penalty on model complexity to provide parsimonious models;
- *sparse methods* (Stepwise Selection [21], LARS [15] and LASSO, that belongs to both classes) that generate models using a subset only of the input variables.

In the present work, linear regularization methods (RR and LASSO) are adopted, given that they are simple and capable of computing models in a reasonable amount of time, while providing similar prediction performance of non-linear approaches [22]. Such techniques are briefly reviewed in the next subsection.

2.2. Linear methods for VM

The basic assumption in machine learning modeling is that the information needed to build an accurate predictive model can be learned from historical data. Given a training set of n examples

$$\{x_i, y_i,\ i = 1, \ldots, n\}$$

with $x_i \in \mathbb{R}^{1\times p}$ and $y_i \in \mathbb{R}$,

let $X \in \mathbb{R}^{n\times p}$ be a matrix of p-variate inputs, obtained by stacking the $\{x_i\}$, and $Y \in \mathbb{R}^n$ be the associated real-valued output vector. The modeling objective is to find a function $f(\cdot)$ such that, given a new observation $\{x_{\text{new}}, y_{\text{new}}\}$, a suitable norm of the difference between $f(x_{\text{new}})$ and y_{new} is small.

In the case in exam, linearity is assumed and $f(x)$ is a linear combination of the inputs x_i with coefficients vector $w \in \mathbb{R}^p$, that need to be estimated:

$$f(x_i) = \sum_{j=1}^{p} x_{i,j} w_j \tag{1}$$

where the coefficient w_j is associated to the j-th input variable.

When dealing with high-dimensional VM problems, classical approaches to linear modeling like Ordinary Least Squares suffer from two main drawbacks:

(i) when only few observations are available ($n \simeq p$), the estimated $f(x)$ might overfit or even interpolate the training examples;
(ii) the problem may be ill-conditioned or even singular, leading to an unstable solution.

To overcome these issues, *regularization* techniques have been developed. In general, such methodologies make additional assumptions on the complexity of $f(x)$, to improve prediction accuracy [23].

Ridge Regression is, perhaps, the most popular regularized machine learning algorithm. It consists in solving the following minimization problem:

$$J_{\text{RR}}(w) := \frac{1}{2}\|Y - Xw\|^2 + \frac{\lambda}{2} w'w = \text{RSS}(w) + \frac{\lambda}{2} w'w \tag{2}$$

where $\lambda \in \mathbb{R}^+$ is a regularization (hyper)parameter [24]. The larger the value of λ, the smaller the "complexity" of the selected model (the variance of the estimator) is, at the cost of worsening the performances on the training set $\{X, Y\}$ and introducing bias.

The RR problem (2) has a closed-form solution (see [24]) and it is therefore easy to be solved, moreover, it efficiently deals with the case of highly correlated dataset. On the other hand, the entries of w_{RR} (the coefficients of the RR model) related to irrelevant input variables are shrunk without reaching zero. This poses a potential threat when using an automatic prediction system, as an outlying value in any of the input parameters would result in completely wrong predictions.

To overcome such issue, a technique able to jointly select the order of the model in a sparse fashion and estimate the coefficients has to be used. One possible way to achieve this goal is to penalize the model coefficients w by using a ℓ_1 norm. In this way, the so-called *curse of dimensionality* is avoided, as well as the risk of obtaining over-parametrized models [25]. The most popular technique employing such regularization approach is the *Least Absolute Shrinkage and Selection Operator* (*LASSO*) [26], that solves the following optimization problem:

$$w = \arg\min_{w} \text{RSS}(w) \tag{3}$$

$$\text{with} \quad \sum_{j=1}^{p} |w_j| \le \lambda. \tag{4}$$

This formulation allows to obtain a sparse solution for w (that is, some entries of the selected w are 0) if λ is small enough. This extremely convenient property of the LASSO allows for the creation of low-order models even when the input space has high dimension. Intuitively, stability of the prediction is improved without sacrificing precision [27].

The hyper-parameter λ acts for both RR and LASSO as a tuning knob: by selecting small values of λ, the amount of complexity allowed in the model (and the number of selected variables in the LASSO)

decreases. Depending on the nature of the dataset, the 'optimal,' trade-off value of the regularization parameter is chosen as the one minimizing the prediction accuracy in cross-validation. In the case of LASSO, hard penalties provide models that are highly sparse, an interesting feature in terms of model interpretability. There is no closed-form solution to the minimization problem described in (3) and (4), therefore, to train a predictor by using the LASSO, it is necessary to resort to optimization techniques. Efficient implementations of the LASSO are nowadays available, some of the most popular algorithm are based on Sequential Minimization Optimization [28] and Least Angle Regression (LARS) [29].

In Fig. 3 a simple semiconductor manufacturing dataset with $p=10$ variables [15] is considered. The coefficients of RR and LASSO (implemented with LARS) vary with the level of complexity: for large values of λ (enhanced model simplicity) the RR coefficients are in fact shrunk close to each other, while at each iteration of the LARS (more and more complexity allowed in the model at each iteration) a new variable enters the LASSO solution.

In the present work, both RR and LASSO are taken into consideration, since they are both effective in the VM problems considered here. It is to be stressed that there is no way to state *a priori* which of the two approaches performs best in terms of prediction accuracy, computational time, and interpretability (see [30,31] to this regard,) actual results heavily depend on the dataset and requirements at hand.

3. Multistep virtual metrology

In this section, Multi-step VM problems are detailed and two approaches for VM input space definition are presented. To better clarify the Multi-Level paradigm that will be detailed, reference to the following Multi-Step setup will be made.

(i) A *production flow* is defined as a sequence of process steps; each step represents an operation to be performed on a wafer. For illustrative purposes, we refer again to the example of Fig. 2. The production flow in this case is defined by a CVD, a lithography, and an etching step.
(ii) Each step can be performed by different tools and the knowledge of which tool processed a specific wafer is available. Furthermore, each tool can be composed of different chambers.
(iii) Each tool provides information about the processed wafer, including sensor readings and recipe set points. It is assumed that all the tools that deal with a certain step (for instance, all the involved CVD equipment) provide homogeneous process information.
(iv) 'Single step', classical VM modules can be implemented on some tools to estimate key wafer features (for example, the thickness of the deposited layer for the CVD) that are generally measured only once in a lot (see Fig. 4).

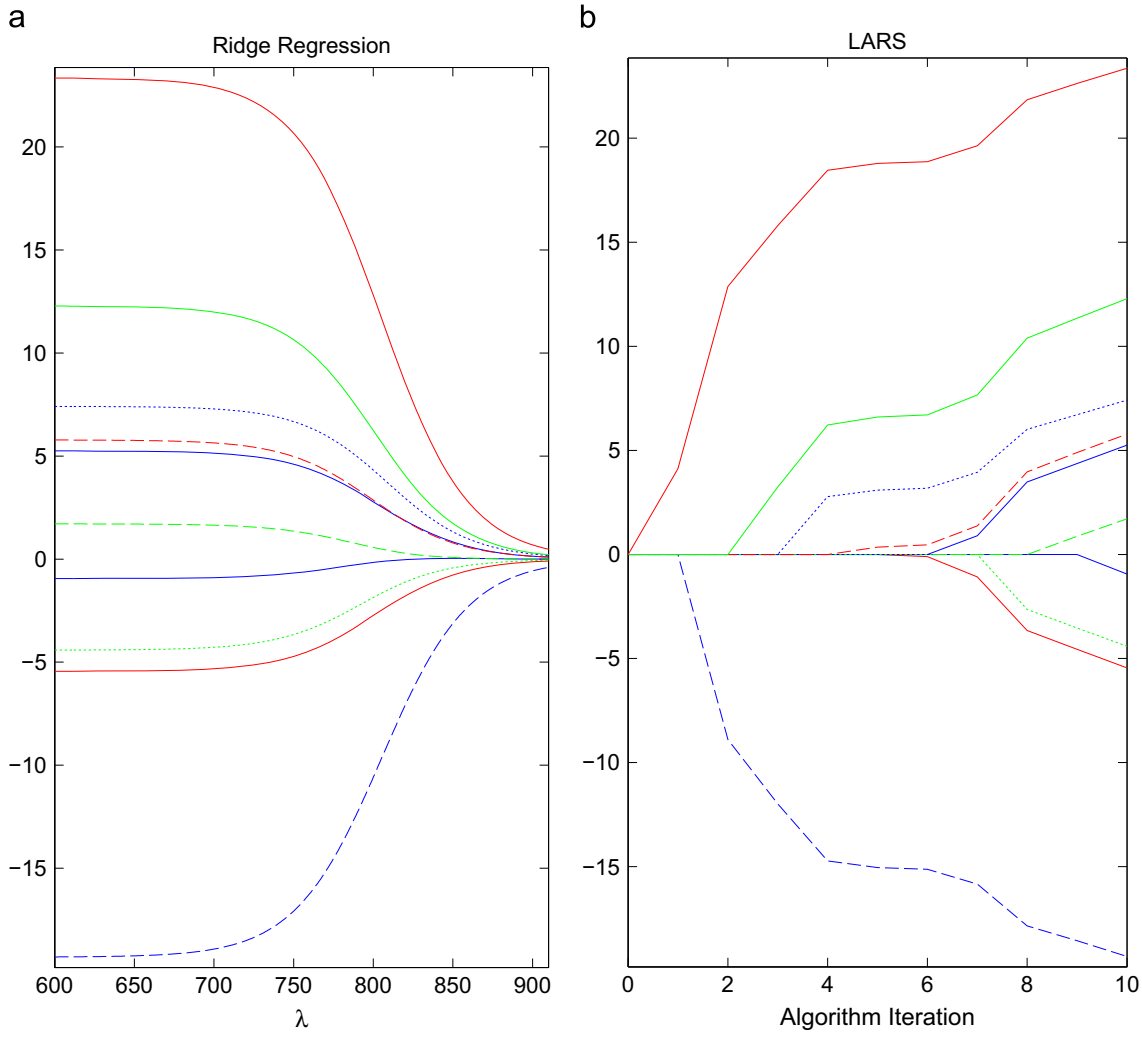

Fig. 3. Evolution of the RR and LARS coefficients with different complexity allowed in the model. Adapted from [15]. (a) RR coefficients' evolution and (b) LARS coefficients' evolution.

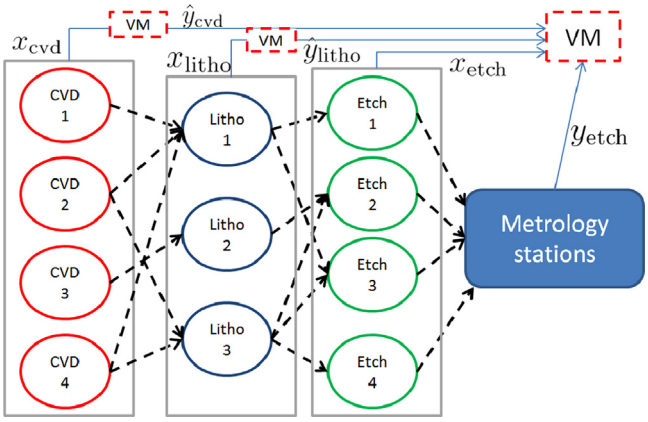

Fig. 4. In the 'cascade VM' scenario, single-step Virtual Metrology modules are producing information for some of the previous process steps. The predicted values are then incorporated in the input space.

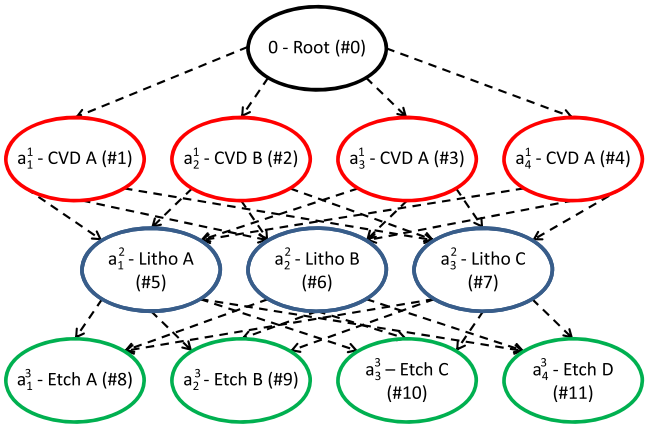

Fig. 5. Tree structure of the production flow in a Multi-Step problem with 4 CVD, 3 lithography and 4 etching tools.

In the following subsections, different multi-step approaches are grouped according to the type of information required for their application. The basic assumption is that, for the last step of the considered production flow, whose metrology values are the targets to be predicted (the VM targets), all relevant information is available.

In the following, we consider a process consisting of γ sequential steps. The i-th step can be performed by η_i different tools, while the total number of tools involved in the dataset is η:

$$\eta = \sum_{i=1}^{\gamma} \eta_i. \tag{5}$$

In Fig. 5 we refer again to the example of CVD, Lithography and Etching multi-step VM problem, where 4 CVD tools, 3 Lithography tools and 4 Etching tools are available.

Data fragmentation could be avoided by using a unique model for all the possible paths of the tree in Fig. 5, although suboptimal predictions would be obtained, since differences among the tools are not considered. On the other hand, having a model for every distinct path is cost intensive (in the settings of Fig. 5, 48 models should be computed, and a further degree of complexity would be introduced if taking into account the presence of different products and chambers).

A possible approach to the problem is provided by Generalized Additive Models (GAMs) [32], that are based on the idea of modeling each logistic entity (each node of the tree) instead of each different production flow (each path of the tree), to decrease the number of distinctive models and to increase the number of observations available for each model.

We introduce here the useful prediction flow related notation:

- γ is the production depth, i.e., the number of production steps considered in the modeling. With reference to the example of Fig. 5, $\gamma = 3$, since 3 processes in sequence are considered (CVD, Lithography and Etching);
- $\mathcal{A}_k = \{a_1^k, \ldots, a_{\eta_k}^k\}$ is the set of available tools for the k-th production step;
- $\eta = \sum_{k=1}^{\gamma} \eta_k$ is the total amount of different tools over the γ considered processes.

$\gamma$is then the number of processes that are considered in the multi-level VM problem, where the γ-th step is the one after which the VM target is measured.

In the following subsections two different multi-step approaches are presented.

3.1. Process-based multistep

The i-th wafer $\{x_i, y_i\}$ is associated with the *logistic path* $\mathcal{P}_i$, that is the sequence of tools on which the wafer has been processed:

$$\mathcal{P}_i = \{p_0, p_1^i, p_2^i, \ldots, p_\gamma^i\},$$
$$\text{where } p_k^i = \begin{cases} 0 & \text{if } k = 0 \text{ Root (common for all wafers)} \\ p_k^i \in \mathcal{A}_k & \text{elsewhere} \end{cases}.$$

For example, with reference to the example of Fig. 5, if the i-th wafer is processed on the tool CVD A, Litho C and Etch B, then $\mathcal{P}_i = \{0, 1, 7, 9\}$.

Following the Multilevel paradigm defined in [17], a GAM of the form

$$f(x_i) = \sum_{k \in \mathcal{P}_i} f_k(x_{i,k}) \tag{6}$$

is sought for, that is, the prediction is expressed as the sum of independent effects connected to all the logistic entities involved in the process. $x_{i,k}$ in (6) represents the input space for the wafer x_i, that is used to model $f_k(\cdot)$, and therefore, the effects of the k-th process performed on the specific tool p_k on the VM target. Let x_{i,p_k} be the part of x_i associated with the k-th process (sensor data collected in the tool p_k, logistic information associated, etc.). Then

$$x_{i,k} = [x_{i,0}\ x_{i,p_1}\ x_{i,p_2}\ \ldots\ x_{i,p_k}], \tag{7}$$

since only the process/sensor data related to p_k and the previous process steps must be taken into account, ignoring what happens in the future of the production flow.

The modeling goal is to estimate $\eta+1$ functions f_k (that is, as many functions as the number of different nodes of the tree, including the root one, by means of which commonalities between all the possible paths can be described) in a multitask learning approach [33]. It should be noted that the number of functions to be estimated may be reduced by exploiting data commonalities. In our example, the number of functions to be estimated is 12 (the number of nodes), instead of 48 (the number of paths). The underlying linear effect superposition assumption is a strong one, but it brings simplicity and reduction of problem dimensionality when performing VM in small datasets situations. In the case of Ridge Regression, for example, the computational cost of each node function f_k is $O(nm_k^2)$, where n is the overall number of observations and m_k is the number of variables of the k-th production step. Classical VM approaches have instead a computational cost of $O(n_i m_{TOT}^2)$ for each path, where n_i is the number of the observations associated with the path in exam and $m_{TOT} = \sum_{i=1}^{\gamma} m_k$. As can be seen, the Multistep approach increases complexity in the number of observations instead of the number of parameters, with a clear advantage in terms of

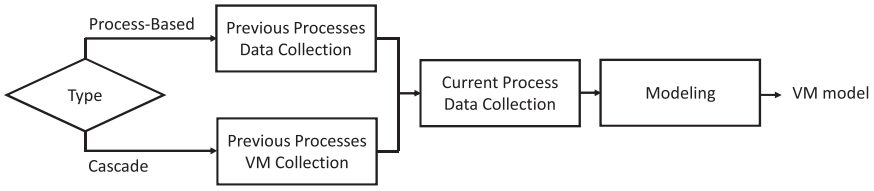

Fig. 6. Flowchart of the different multi-step methodologies.

computational effort. The multilevel approach introduced here can be integrated with the regularization methods described in Section 2.2 as shown in [17] (see [17] for further details).

The approach described above is named as *Process-based Multi-step*. Its main benefits are the following:

- it allows to include data from steps for which no measurements are available, or whose measurements are devoid of meaning with respect to the target step;
- it provides all the available information to the learning algorithms.

The main drawback of the approach is that the input space dimension is significantly increased, and a larger number of observations are likely to be required to estimate the predictive model.

In the next subsection we present a simplified version of the approach, where a different definition of the input matrices related to the previous steps in the production line is used.

3.2. Cascade multistep

We assume that for all the k tools in the production flow of interest before the last process, a 'single step' VM system (Fig. 4) is in place, that provides an estimation $v_{i,k}$ of an important wafer feature/parameter. We stress here that the various 'single step' VM systems in place have to be considered as an existing *a priori* condition in the design of the Multistep VM system, that is, there is no possibility of choosing the algorithms used to perform such VM steps.

We exploit the availability of the VM prediction $v_{i,k}$ to summarize all the information regarding the process performed in equipment k with a single parameter (or more than one in the case of multi-output VM systems). In this approach, the entries of the input vector x_i in (7) are defined as follows:

$$x_{i,k} = \begin{cases} v_{i,k} & \text{if } k < \gamma \\ \text{process and logistic data} & \text{if } k = \gamma \end{cases} \tag{8}$$

The input matrices are therefore populated only with the previous Virtual Metrology predictions for equipments that do not belong to the target step, and with all the available information (sensor and logistic data) for the last step.

We call this approach *Cascade Virtual Metrology* as it allows to build a pipe system in which the predictive information is forward propagated to concur to further model estimations. The main advantage of this methodology is the small overhead appended to the input space, an useful feature to ease the model selection process and reduce the computational burden. Conversely, the two main drawbacks of this approach are the following:

- Virtual Metrology systems must already be in place for steps that precede the target step.
- Using VM information as an input (usually, some weighted combination of process parameters) may lead to information loss between two or more steps.

A problem may arise if the prediction performance of the VM models providing the estimations used as input deteriorates after the Multi-step VM model has been computed. A solution to this issue may be to periodically recompute the Multi-step VM model (accordingly to computational capabilities and production requirements) once the solution is implemented on-line.

In Fig. 6 it is reported a flowchart of the proposed methodologies.

4. Experimental results

To validate the proposed Multistep VM approaches, a dataset provided by a semiconductor manufacturing industry[1] regarding production wafers is employed as a benchmark. Such dataset has been collected considering the following production flow, that consists of 3 deposition steps and 1 lithography step:

(i) *Chemical Vapor Deposition* (*CVD*): A process in which a thin film of solid material is produced on the surface of a wafer. The quality of the deposited layer is usually evaluated by measuring its thickness (THK) and uniformity (typically the standard deviation of various measurements performed at different coordinates on the wafer).
(ii) *Thermal Oxidation*: A process in which multiple wafers are heated (usually in a furnace) to produce a thin layer of oxide by forcing an oxidizing agent to react with the wafer materials.
(iii) *Coating*: A process in which the wafer is covered by a viscous solution of photoresist that is rapidly removed in order to produce a thin layer.
(iv) *Lithography*: This process allows to remove predefined parts of the wafer substrate by means of photomasks. In this way, geometric patterns are transferred on the photoresist. The results of this operation are evaluated by measuring geometric features (e.g. height, width, depth) of the created pattern that are named Critical Dimensions (CDs) [34].

The VM target is a particular CD after lithography. The available dataset of $n=583$ samples consists of data from 4 CVD tools (*A*, *B*, *C* and *D*), 2 Thermal Oxidation tools (*E* and *F*) and a single coating and lithography machine. The samples distribution is detailed in Table 1.

We evaluate the performance of the two different Multistep (MS) VM approaches (process based and cascade) and two Regularization algorithms (Ridge Regression and LASSO, described in Section 2.2). It should be noted that the only tuning knob for RR and LASSO is the hyperparameter λ: the value of the regularization parameter has to be chosen to maximize the prediction performances.

To evaluate the importance of considering multiple processes to enhance VM predictions, we consider several combinations of the available processes:

- *Lithography*: Data regarding the lithography process only (that is, the classical 'single step' VM approach).

[1] Courtesy of Infineon Technology AG, Austria, Villach facility.

Table 1
Sample sizes of the considered datasets.

CVD tool	*A*	*A*	*B*	*B*	*C*	*C*	*D*	*D*
Thermal oxidation tool	*E*	*F*	*E*	*F*	*E*	*F*	*E*	*F*
# of samples	91	125	149	14	31	26	86	61

Table 2
Regressors sizes of the considered datasets.

Dataset name	**# regressors**
Lithography	827
CVD, Lithography	1034
CVD, Oxidation, coating	305
Full Information	1132

- *CVD, Lithography*: Data regarding CVD and Lithography.
- *CVD, Oxidation, Coating*: Data regarding CVD, Thermal Oxidation and Coating.
- *Full information*: Data regarding CVD, Thermal Oxidation, Coating and Lithography.

The sizes of the aforementioned datasets are summarized in Table 2.

Cross-validation is then employed, with a twofold purpose, namely, to tune the hyperparameter λ and to evaluate the performance of the proposed models. Performance is evaluated in terms of Root Mean Squared Error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{(1-q)n}\sum_{i=1}^{(1-q)n}(\hat{y}_i - y_i)^2}; \tag{9}$$

in this work we chose $q=0.7$. Assessing the quality of the model on a set of observations of the phenomena that has not being used for model construction is essential to have a fair evaluation of the prediction performance, especially when only a limited number of observations are available; for this reason the dataset of n samples is split in two parts:

- a training dataset (qn samples, where $0<q<1$), that is used to build the model;
- a validation dataset ($(1-q)n$ samples), that is used to assess the prediction performance of the model.

The splitting between training and validation sets can strongly affect the performance and the outcome of the tuning procedure. This issue can be dealt with by resorting to Repeated Random Sub-Sampling Validation [35], also known as *Monte Carlo cross-validation* (*MCCV*). A set of K MCCV simulations consists of performing an analysis on K different random splits of the available observations into training and validation datasets: K different models are built accordingly and the performance of the proposed methodology is assessed as the average model performance over K simulations. In this way, the MCCV procedure yields an evaluation of modeling performance that is not affected by the particular partition of the dataset.[2]

To achieve stable MCCV estimates, the number K of simulations needs to be relatively large – in the order of hundreds/thousands. In this work, $K=1000$ MC simulations have been performed. The hyperparameter λ is tuned for both Ridge Regression and LASSO to minimize the RMSE over the K MC simulations on the validation data set.

[2] It has been shown [36] that MCCV is asymptotically consistent resulting in more pessimistic predictions of the test data compared with full cross-validation.

Before analyzing the performance of the proposed multi-level approaches, we show how classical VM approaches behave when multiple sources are considered. The single step VM (Lithography) has been compared with two multi-sources approaches that considered also the previous steps (CVD, Oxidation, Coating):

- 1-*path*: where all the observations, independently from the logistic path, have been modeled together;
- *all-paths*: where all the logistic paths (8 paths) have been modeled separately and then the performance is averaged.

The average results over the MC simulations are reported in Table 3, while in Fig. 7 (for RR) and Fig. 8 (for LASSO) are reported the boxplots of the RMSE at the value λ^*:

$$\lambda^* = \arg\min_{\lambda}\frac{1}{K}\sum_{i=1}^{K}\mathrm{RMSE}_i(\lambda),$$

where $\mathrm{RMSE}_i(\cdot)$ is the RMSE obtained at the i-th MC simulation. In the case of Cascade multi-step, it is assumed that early stage VM modules are already in place, as is typical of present day fab environments. We remark here that such VM modules are individually designed to achieve a specific goal, without considering the possibility of using their estimates as inputs of other VM modules that may be placed downstream the process flow, or added afterwards. As a consequence, their output $v_{i,k}$ (whatever the employed modeling approach) has to be fed to both multi-sources VM schemes, employing LASSO and RR. In the specific situation at hand, the $v_{i,k}$ have been computed via RR (Fig. 8).

Observe that multi-sources modeling exhibit worse performance than one source VM, in particular in the all-paths approach, where the availability of few samples for the singular logistic paths translates in poor prediction capabilities of the models.

We consider now the performance of multi-level techniques: the results are summarized in Table 4. Some remarks are in order:

- the proposed Multistep VM strategies allow to improve the performances of classical VM approaches;
- the process data of the target process alone still performs fairly well: intuitively, excluding data from the target step yields the worst results;
- Ridge Regression outperforms LASSO in most cases for the dataset at hand; this might be due to the presence of important collinearities in the input space, where the natural averaging properties of the Ridge Regression can act as a noise-mitigating filter;
- for both algorithms (Ridge Regression and LASSO), the best overall performances were obtained considering all process steps (CVD, Coating, Oxidation and Lithography);
- the *Cascade* strategy obtains the best results, compared to the most complete approach: this peculiar outcome might be attributed to the small sample size with respect to the input space dimensionality.

To assess the statistical significance of the results reported in Table 4 the test mean on the errors obtained with single step and multi-step (the best between cascade and process-based) with full information has been computed.[3] In the case of LASSO the resulting p-value was equal to 0.0002 while for RR $p=0.0001$, therefore indicating that the obtained results are statistically significant.

The experimental/tuning simulations are reported in Fig. 9 (Process-based with RR), in Fig. 10 (cascade with RR), in Fig. 11

[3] *Welch's t Test* has been employed in the computation of the p-values.

Table 3
Minimum RMSE [10^{-3}] for different modeling approaches. In bold the minimum values for each experiment.

Method	MS approach	Lithography	1-path	all-paths
RR	Process-based	**91.59**	92.01	100.81
RR	Cascade	**91.59**	91.63	99.476
LASSO	Process-based	**99.53**	100.85	101.11
LASSO	Cascade	**99.53**	99.54	100.87

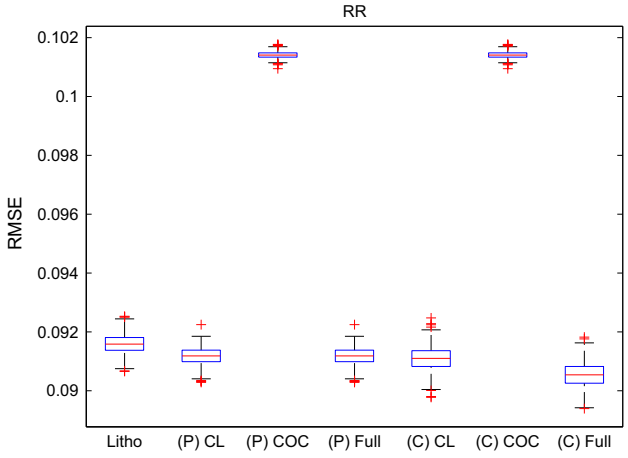

Fig. 7. Boxplot of the RMSEs at λ^* with RR. *Notation*: Process (P) and Cascade (C) based – Lithography (Litho) – CVD, Lithography (CL) – CVD, Oxidation, Coating (COC) – Full Information (Full).

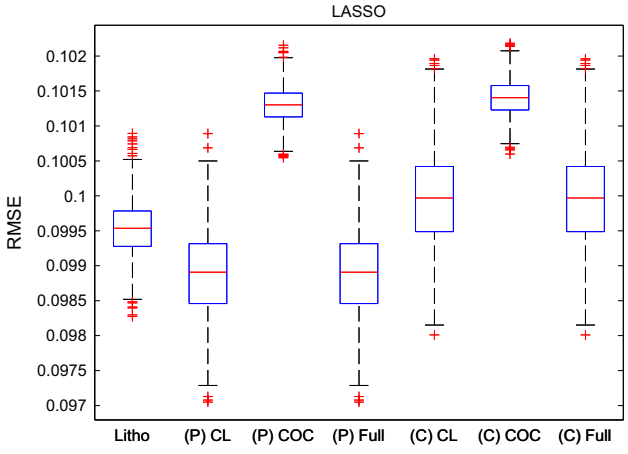

Fig. 8. Boxplot of the RMSEs at λ^* with LASSO. Notation: Process (P) and Cascade (C) based - Lithography (Litho) - CVD, Lithography (CL) - CVD, Oxidation, Coating (COC) - Full Information (Full).

Table 4
RMSE [10^{-3}] of Multi-Level approaches. In bold the minimum values for each experiment.

Method	MS approach	Fig.	Litho	CVD, Litho	CVD, Oxid., coating	Full Info
RR	Process-based	9	91.59	91.17	101.41	**91.01**
RR	Cascade	10	91.59	91.11	101.41	**90.55**
LASSO	Process-based	11	99.53	98.85	101.31	**98.85**
LASSO	Cascade	12	**99.53**	99.94	101.41	99.94

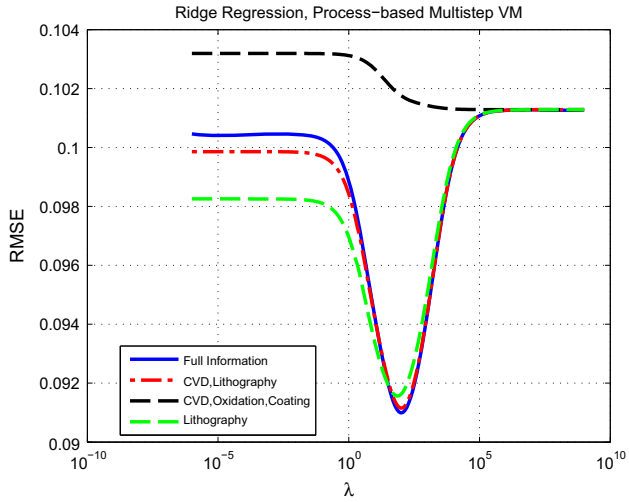

Fig. 9. Averaged (over $K=1000$ MC simulations) RMSE as a function of λ for Process-based Multistep VM with RR.

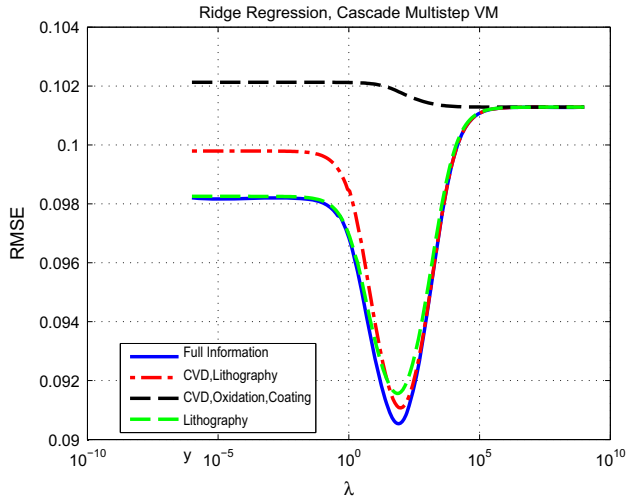

Fig. 10. Averaged (over $K=1000$ MC simulations) RMSE as a function of λ for Cascade Multistep VM with RR.

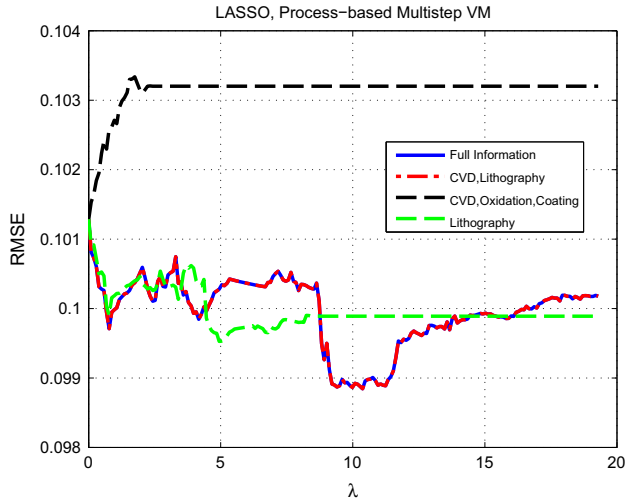

Fig. 11. Averaged (over $K=1000$ MC simulations) RMSE as a function of λ for Process-based Multistep VM with LASSO.

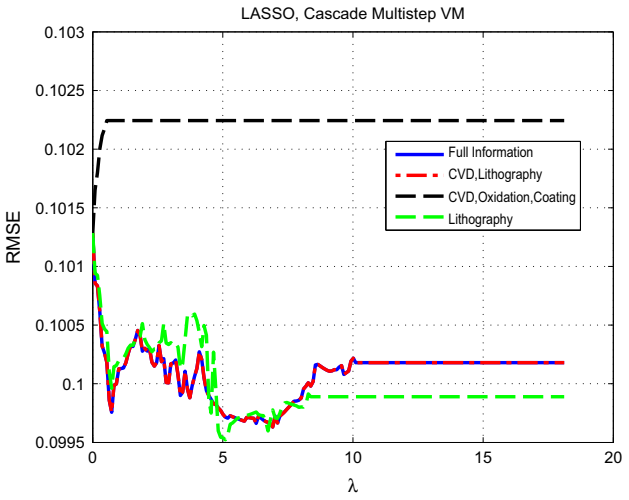

Fig. 12. Averaged (over $K=1000$ MC simulations) RMSE as a function of λ for Cascade Multistep VM with LASSO.

(Process-based with LASSO) and in Fig. 12 (cascade with LASSO), where the behavior of RMSE as a function of λ is shown.

5. Conclusions

In this paper, a novel strategy for Virtual Metrology in semiconductor manufacturing has been proposed. The proposed *Multi-Step Virtual Metrology* approach consists in using information about previous process step(s), as process data, logistic data, and virtual and actual measurement values, jointly to the current process information, to improve the precision and the accuracy of the VM system.

The proposed approach is based on regularization methods (Ridge Regression and LASSO) and multi-task learning techniques to deal with the two most prominent issues in VM modeling:

- high dimensionality;
- data fragmentation.

The aforementioned issues are amplified if we try to gather information from multiple processes instead of considering just the VM target related process.

To cope with the previous issues, two different Multi-Level strategies have been proposed, namely

(i) Process-Based Multi-Step, where the overall collected information during all process steps are used as input of the VM module;
(ii) Cascade Multi-Step, where intermediate VM estimations are used to summarize the information related to previous processes before the last one in the line.

These methods have been tested on a production dataset associated with a four-step process flow (CVD, Thermal Oxidation, Coating, Lithography) that are in series in the wafer production, generally considered to influence variability of the Critical Dimension in the Lithography, that represents the VM target. Four different combinations of data sources have been considered to evaluate the performances of the multi-step approach with increased complexity and dimensionality of the problem.

Results show that VM performance can be improved by enriching the dataset with information related to past processes. The evaluation of VM system quality is mainly provided by the accuracy of the predictions [19,37,38]. If the predictions are considered accurate enough, then they can be used instead of real measurements in cost reduction policies. For this reason, even relatively 'small' improvements in VM prediction quality are considered highly valuable in this application setup.

The drawbacks of the proposed framework w.r.t. classical VM solutions are related to the availability of previous step data and the increased modeling complexity. However, both issues are usually manageable in modern semiconductor manufacturing environments. It should be also remarked that care must be taken when designing multi-step strategies. The inclusion of information related to more process steps can be usually associated to an enrichment of the dataset, however, over complicated modeling settings may lead to poor prediction accuracy. For this reason, sample size and a priori knowledge (if available) on the relevance of a process in the past to the VM target under exam should always be considered before implementing a multi-step VM strategy.

References

[1] Susto G, Pampuri S, Schirru A, Nicolao GD, McLoone S, Beghi A. Automatic control and machine learning for semiconductor manufacturing: review and challenges. In: 10th European workshop on advanced control and diagnosis; 2012.
[2] Lynn S, Ringwood J, Ragnoli E, McLoone S, MacGearailty N. Virtual metrology for plasma etch using tool variables. In: Advanced semiconductor manufacturing conference, 2009. ASMC'09. IEEE/SEMI. Berlin, Germany: IEEE; 2009. p. 143–8.
[3] Susto G, Beghi A, DeLuca C. A virtual metrology system for predicting cvd thickness with equipment variables and qualitative clustering. In: IEEE conference on emerging technologies and factory automation; 2011. p. 1–4.
[4] Chen P, Wu S, Lin J, Ko F, Lo H, Wang J, et al. Virtual metrology: a solution for wafer to wafer advanced process control. In: IEEE international symposium on semiconductor manufacturing, 2005, ISSM 2005. San Jose CA: IEEE; 2005. p. 155–7.
[5] Susto G, Schirru A, Pampuri S, DeNicolao G, Beghi A. An information-theory and virtual metrology-based approach to run-to-run semiconductor manufacturing control. In: 8th IEEE international conference on automation science and engineering; 2012. p. 358–63.
[6] Kurz D, Kaspar J, Pilz J. Dynamic maintenance in semiconductor manufacturing using Bayesian networks. In: IEEE conference on automation science and engineering (CASE). Trieste, Italy: IEEE; 2011. p. 238–43.
[7] Friedman J. On bias, variance, 0/1 loss, and the curse-of-dimensionality. Data Min Knowl Discov 1997;1(1):55–77.
[8] Schirru A, Susto G, Pampuri S, McLoone S. Learning from time series: supervised aggregative feature extraction. In: IEEE 51st annual conference on decision and control (CDC); 2012. p. 5254–9.
[9] Susto G, Beghi A, De Luca C. A predictive maintenance system for epitaxy processes based on filtering and prediction techniques. IEEE Trans Semicond Manuf 2012;25(4):638–49.
[10] Xiao H. Introduction to semiconductor manufacturing technology, vol. 16. Upper Saddle River, NJ: Prentice Hall; 2001.
[11] Khan A, Moyne J, Tilbury D. An approach for factory-wide control utilizing virtual metrology. IEEE Trans Semicond Manuf 2007;20(4):364–75.
[12] Susto G, Johnston A, O'Hara P, McLoone S. Virtual metrology enabled early stage prediction for enhanced control of multi-stage fabrication processes. In: IEEE international conference on automation science and engineering (CASE). Madison, WI: IEEE; 2013. p. 201–6.
[13] Pampuri S, Schirru A, Susto G, DeNicolao G, Beghi A, DeLuca C. Multistep virtual metrology approaches for semiconductor manufacturing processes. In: 8th IEEE international conference on automation science and engineering; 2012.
[14] Pampuri S, Schirru A, Fazio G, De Nicolao G. Multilevel lasso applied to virtual metrology in semiconductor manufacturing. In: IEEE conference on automation science and engineering (CASE). Trieste, Italy: IEEE; 2011. p. 244–9.
[15] Susto G, Beghi A. A virtual metrology system based on least angle regression and statistical clustering. Appl Stoch Models Bus Ind 2013;29:362–76.
[16] He J, Zhu Y. Hierarchical multi-task learning with application to wafer quality prediction. In: IEEE 12th international conference on data mining; 2012. p. 290–8.
[17] Schirru A, Pampuri S, De Luca C, De Nicolao G. Multilevel kernel methods for virtual metrology in semiconductor manufacturing. In: IFAC world congress, vol. 18; 2011. p. 11614–21.
[18] Prakash P, Johnston A, Honari B, McLoone S. Optimal wafer site selection using forward selection component analysis. In: 23rd IEEE/SEMI advanced semiconductor manufacturing conference (ASMC); 2012.
[19] Cheng F, Huang H, Kao C. Dual-phase virtual metrology scheme. IEEE Trans Semicond Manuf 2007;20(4):566–71.
[20] Purwins H, Nagi A, Barak B, Hockele U, Kyek A, Lenz B, et al. Regression methods for prediction of pecvd silicon nitride layer thickness. In: IEEE conference on automation science and engineering (CASE); 2011. p. 387–92.

[21] Lin T-H, Cheng F-T, Wu W-M, Kao C-A, Ye A-J, Chang F-C. NN-based key-variable selection method for enhancing virtual metrology accuracy. IEEE Trans Semicond Manuf 2009;22:204–11.
[22] Susto G, Beghi A. Least angle regression for semiconductor manufacturing modeling. In: IEEE international conference on control applications (CCA); 2012. p. 658–63.
[23] Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: data mining, inference and prediction. Math Intell 2005;27(2):83–5.
[24] Hoerl A, Kennard R. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 2000;42(1):55–67.
[25] Bellman R, Lee E. History and development of dynamic programming. IEEE Control Syst Mag 2002;4(4):24–8.
[26] Susto G, Schirru A, Pampuri S, Beghi A. A predictive maintenance system based on regularization methods for ion-implantation. In: Proceedings of the 23rd IEEE/SEMI advanced semiconductor manufacturing conference; 2012. p. 175–80.
[27] Ramirez I, Lecumberry F, Sapiro G. Universal priors for sparse modeling. In: 3rd IEEE international workshop on computational advances in multi-sensor adaptive processing (CAMSAP). Aruba, Dutch Antilles: IEEE; 2010. p. 197–200.
[28] Platt JC. Fast training of support vector machines using sequential minimal optimization. In: Advance in kernel methods—support vector learning. Cambridge, MA: MIT Press; 1999.
[29] Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Ann Stat 2004;32:407–99.
[30] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc: Ser B (Methodol) 1996;58:267–88.
[31] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc: Ser B (Stat Methodol) 2005;67(2):301–20.
[32] Hastie T, Tibshirani R. Generalized additive models. Stat Sci 1986;1:297–310.
[33] Caruana R. Multitask Learning. MMachine Learning 1997;28:41–75.
[34] Ito R, Okazaki S. Pushing the limits of lithography. Nature 2000;406(6799):1027–31.
[35] Picard RR, Cook RD. Cross-validation of regression models. J Am Stat Assoc 1984;79(387):575–83.
[36] Shao J. Linear model selection by cross-validation. J Am Stat Assoc 1993;88(422):486–494.
[37] Kang P, Lee H-J, Cho S, Kim D, Park J, Park C-K, et al. A virtual metrology system for semiconductor manufacturing. Expert Syst Appl 2009;36(10):12554–61.
[38] Su Y-C, Lin T-H, Cheng F-T, Wu W-M. Accuracy and real-time considerations for implementing various virtual metrology algorithms. IEEE Trans Semicond Manuf 2008;21(3):426–34.