Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

DYNAMIC SAMPLING FOR RISK MINIMIZATION IN SEMICONDUCTOR MANUFACTURING

Étienne Le Quéré
Stéphane Dauzère-Pérès
Karim Tamssaouet

Mines Saint-Étienne, Univ Clermont Auvergne
CNRS, UMR 6158 LIMOS
Department of Manufacturing Sciences and Logistics
880 route de Mimet
F-13541 Gardanne, FRANCE

Cédric Maufront
Stéphane Astie

Yield and Statistic Process Control Department
Soitec
922, Parc technologique des Fontaines
F-38190 Bernin, FRANCE

ABSTRACT

To control the quality of their processes, manufacturers perform measurement operations on their products. In semiconductor manufacturing, measurement capacity is limited because metrology tools are expensive, thus only a limited number of products can be measured. Selecting the set of lots to control to minimize risk is called sampling. In this paper, the objective is to minimize the number of wafers at risk, i.e. the number of wafers processed on a machine between two lots that are controlled. The problem can be modeled as the maximization of a submodular set function subject to various capacity constraints. The resulting problems, which are NP-hard, can be modeled as integer linear programs. Greedy heuristics and an exchange procedure are also presented. Computational experiments on industrial and randomly generated instances show that the integer linear programs solve the problems optimally, and that the heuristics have sufficiently good approximation ratios for industrial implementation.

1 INTRODUCTION

In this paper, we study the problem of optimally sampling a set of lots to measure to minimize the risk on production machines in semiconductor manufacturing. Providing efficient approaches for this problem is critical in the industry because of the operational constraints on the measurement tools. Control operations are becoming more and more costly as technology improves. We formulate the sampling problem in semiconductor manufacturing as the budgeted maximization of a submodular function that describes the industrial environment. We explore integer linear programming models with cardinal, knapsack and multiple knapsack capacity constraints, analyze the performances of a greedy algorithm and a local search by comparing them to the integer linear programs solved by a standard solver on both industrial and randomly generated instances.

The idea of risk comes with the fear of unexpected outcomes, which often are unexpected costs or losses. In manufacturing, the risk is related to quality issues, like a capability shift on a production machine leading to the production of non-conformed items. Such events can be detected by some control operations. (Feigenbaum 1951) was one of the first author to put together costs related to quality actions or failures in manufacturing. This is called the cost of quality and represents the balance between the cost of controls and the cost of non-conformance of products. Therefore, it is essential for manufacturers to identify the elements that bring variability in the process. Risk is also a key notion when optimizing

portfolios in financial analysis. In this context, (Holton 2004) separates the idea of risk into the combination of two components: Uncertainty and exposure. Hence, managing risk consists in minimizing uncertainty, exposure or a combined indicator to handle the balance between uncertainty and exposure. The Value At Risk (Jorion 2007) is an indicator combining uncertainty and exposure and is widely used by banks, often with stochastic techniques, such as in (Rockafellar, Uryasev, et al. 2000), to plan their investments. In manufacturing, the financial risk is linked to the quality of the products. To ensure their conformity, control or measurement operations are performed on some products, and these operations are particularly expensive in semiconductor manufacturing. The risk then corresponds to the number of wafers produced on a machine between two lots that are controlled, and is called W@R (Wafers at Risk) in semiconductor manufacturing. Selecting the lots to control is called "sampling", and should be optimized to minimize the balance between the cost of controls and the cost of reworked or scrapped wafers.

Our computational experiments on industrial data show that a greedy heuristic and a integer programming solver are efficient to solve the sampling problem, with the greedy heuristic often finding the optimal solution. Risk modeling in semiconductor manufacturing using W@R (Wafers at Risk) levels is presented in Section 2. Section 3 describes the problem and shows its submodular nature. The properties of submodular functions and the algorithms used to solve problems related to these set functions are introduced in section 4. Section 5 presents and analyzes computational experiments on random and industrial instances. Conclusions and perspectives are discussed in Section 6.

2 RISK MODELING IN SEMICONDUCTOR MANUFACTURING

In many manufacturing systems, items are produced by lots. Finished lots are inspected to check the health of the production machines and to decide to start new lots on the machines or to perform maintenance operations. Acceptance sampling is very popular in these cases, the main idea is to only control some products in the lot, and then decide whether to accept, discard or fully inspect the lot based on the conformity of the sampled products. It is also interesting to consider fully integrated plans as in (Bouslah, Gharbi, and Pellerin 2016), where the manufacturing costs are minimized at once, hence the sampling plan is part of a larger system.

In semiconductor manufacturing, lots require hundreds of operations, each of them needs to be processed in workshops made up of parallel machines. There are also re-entrant flows, often many different products. These characteristics make the design of a fully integrated plan very difficult. In most cases, the scheduling, sampling and maintenance plans are considered independently. For our concern, we do not make any hypotheses on the scheduling or maintenance plans. In this context, performing a sampling consists of selecting products (or lots) to be controlled. These measurements are carried out on dedicated measurement tools, often also called metrology tools. We consider that sampling decisions are taken in real time, based on the lots that are available to be controlled and the available measurement capacity. This is the main idea in the *dynamic methods* presented below. Cycle times in semiconductor manufacturing are large, and defects can occur at any operation in the processing route. To limit the risk of losing very large quantities of products, manufacturers have to perform inline control operations in addition to 100% inspection at the end of the line. Control operations do not add value to the products. The literature review (Nduhura-Munga, Rodriguez-Verjan, Dauzère-Pérès, Yugma, Vialletelle, and Pinaton 2013) splits the inline sampling plans in three categories: Static, adaptive and dynamic. Static sampling rates define the sampling rates with rules such as *measure one lot every 10 lots processed on this production machine*. The main advantage is that static rules are easy to implement. Adaptive sampling rules mostly enhance the static rules by computing them based on some knowledge of the schedule plan or machine failure modes. The inconvenient of adaptive sampling rules is that the schedule in a factory is complex, making it hard to find a simple hypothesis to model the design of adaptive sampling rules. In this paper, we are interested in dynamic sampling rules,

where the idea consists in building indicators that model the industrial environment in real time and decide in real time which lots should be measured.

Static and adaptive rules are not efficient enough when there are several production operations between a production operation and its measurement operation as shown in Figure 1.

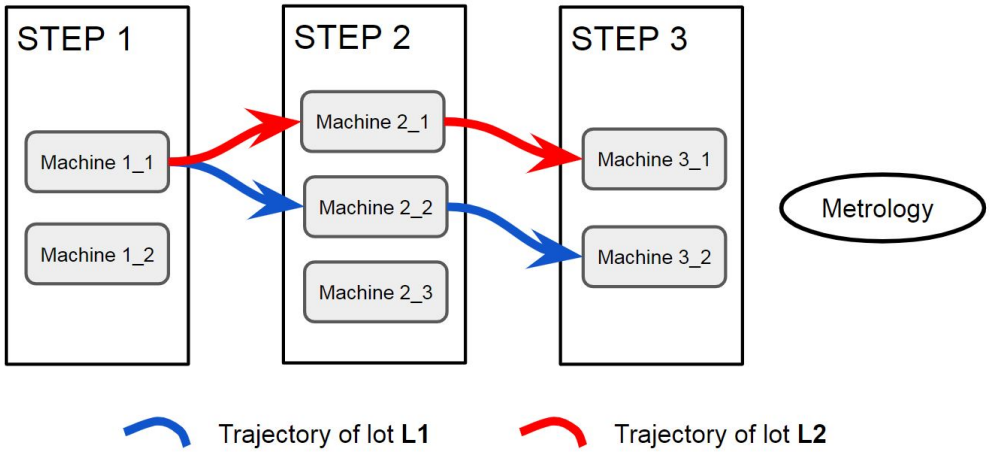

Figure 1: Small example of product routes (series of processing steps) in semiconductor manufacturing

With static or adaptive rules, the idea is to decide at step 1 which lots to measure when they arrive at the measurement (metrology) station. This may not be efficient because of what can happen in steps 2 and 3. For example, the sample lots could be held for an arbitrary long time. The main idea of dynamic sampling methods is to only decide the lots to measure between steps 3 and the metrology station, when there is no uncertainty on the available lots to measure. Also, manufacturers often want to control steps 1, 2 and 3, and the problem has to be modelled and solved in depth (several steps) and height (several parallel machines); this is a set covering problem. To use dynamic sampling methods, we need to have an indicator to discriminate the important lots from the others. This indicator relies on two hypotheses:

1. The failure mode of a production machine is irreversible, a machine that starts to shift does not return to its normal behavior without maintenance.
2. Measurements are always correct, there are no false positives or false negatives.

Then, we can define the number of items produced since the last measurement for a given risk as the worst case exposure, since the status of these items is unknown and they could be scrapped or reworked. In semiconductor manufacturing, this number is known as W@R for wafers at risk. The W@R has been used to build two risk exposure indicators: the GSI (Global Sampling Indicator), presented in (Dauzère-Pérès, Rouveyrol, Yugma, and Vialletelle 2010) and recalled in section 3.2, and the MAR (Material At Risk) (Bean 2008). The latter is the sum of the W@R for all risks weighted by the probability of wafers to be defective for each risk.

These indicators can be used to optimize and update the sampling decisions in real time. Several environmental parameters can change very rapidly, for example the number of metrology tools available for a control operation, or the number of products at a given process operation. Therefore, decisions taken earlier should regularly be re-evaluated. For example, if some events occur and the measurement capacity decreases from 5 lots to 1, the set of lots sampled earlier should be modified, so that only the most promising lots will be measured. The main goal is to handle the balance between the overall risk level and the use of the measurement capacity. A first approach is to skip (i.e. not measure) the sampled lots that do not bring

enough information, as in (Rodriguez-Verjan, Dauzère-Pérès, Housseman, and Pinaton 2013). Studies on the qualifications (the set of products that can be measured by a measurement tool) of measurement tools and their impact have been conducted in (Sendón, Dauzère-Pérès, and Pinaton 2015). In this paper, we consider the qualifications of measurement tools as fixed.

3 PROBLEM DESCRIPTION AND MODELING

3.1 General formulations

Our sampling problem can be formulated as follows: *Select a set of lots to be measured to minimize the risk in the factory subject to some capacity constraints.* In the simplest case considered in this paper, there is a risk for each production machine, as machines are elements that create variability in the production process.

The following input parameters are used:

- $\mathscr{M}$: Set of measurement tools,
- $\mathscr{R}$: Set of risks (production machines in our case),
- $\mathscr{L}$: Set of lots eligible to be controlled and thus that can be sampled,
- $\mathscr{L}_M \subset \mathscr{L}$: Set of *mandatory* lots, i.e. that must be controlled,
- $t_{l,m}$: Time required to measure lot l on metrology tool m, when a single metrology tool is considered, this time will simply be noted t_l,
- $Q_{l,m} \in \{0,1\}$: Is equal to 1 if lot l can be processed on metrology tool m (i.e. m is "qualified" to measure l), and 0 otherwise.

The objective is to maximize the total information obtained by optimally selecting a subset $\mathscr{S}$ of lots in $\mathscr{L}$, i.e. such that $\mathscr{S} \subset \mathscr{L}$. When there are multiple measurement tools, $\mathscr{S}_m$ is the subset of lots assigned to measurement tool m, i.e. $\mathscr{S} = \bigcup_{m \in \mathscr{M}} \mathscr{S}_m$. Let us introduce the notations below to model the different types of capacity constraints, where the first two types correspond to the case in which a single measurement tool is available:

- *Cardinal capacity constraint:* N is the maximum number of lots that can be measured,
- *Knapsack capacity constraint:* T is the maximum measurement,
- *Multiple knapsack capacity constraint:* T_m is the maximum time that can be allocated on measurement tool m.

With a cardinal capacity constraint, this problem is well-known and presented in one of the ground articles on submodular function maximization (Nemhauser, Wolsey, and Fisher 1978). It has also been formulated as an integer program and proved to be NP-hard in (Cornuejols, Fisher, and Nemhauser 1977). Thus, the three problems studied in this article are NP-hard.

3.2 Global Sampling Indicator

To model the information brought by the measurement of each lot, the Global Sampling Indicator (GSI) models in real time the risk in the industrial environment and is detailed in (Dauzère-Pérès, Rouveyrol, Yugma, and Vialletelle 2010).

Let us introduce some additional notations:

- $W@R_r$: Number of wafers at risk on risk r, i.e. the worst case exposure to risk r,
- $NW@R_r(l)$: Value of $W@R_r$ if lot l is measured,

- $NW@R_r(\mathscr{S}) = \min_{l \in \mathscr{S}} NW@R_r(l)$: Value of $W@R_r$ if set $\mathscr{S}$ of lots is measured. Note that only the largest risk reduction for r is counted, i.e. associated to the lot l in $\mathscr{S}$ with the minimum $NW@R_r(l)$. This is because l "covers" the risk reduction on r of the other lots in $\mathscr{S}$.
- IL_r: Maximum acceptable risk level before the machine is stopped,
- $\alpha > 1$: Real number that is used to increase the risk not linearly as it increases.

The Global Sampling Indicator (GSI) is recalled below:

$$GSI(\mathscr{S}) = \sum_{r \in \mathscr{R}} \left(\frac{NW@R_r(\mathscr{S})}{IL_r} \right)^{\alpha} \tag{1}$$

It is interesting to measure the gain brought by the measurement of the lots in a set $\mathscr{S}$. Let us introduce $f(\mathscr{S})$ as follows:

$$f(\mathscr{S}) = GSI(\emptyset) - GSI(\mathscr{S}) = \sum_{r \in \mathscr{R}} \left(\frac{W@R_r}{IL_r} \right)^{\alpha} - \sum_{r \in \mathscr{R}} \left(\min_{l \in \mathscr{S}} \frac{NW@R_r(l)}{IL_r} \right)^{\alpha} = \sum_{r \in \mathscr{R}} \max_{l \in \mathscr{S}} D_{r,l} \tag{2}$$

where $D_{r,l} = \left(\frac{W@R_r}{IL_r} \right)^{\alpha} - \left(\frac{NW@R_r(l)}{IL_r} \right)^{\alpha}$ values the information brought by lot l on risk r if l is controlled.

Note that $f(\mathscr{S}) = GSI(\emptyset) - GSI(\mathscr{S})$ is submodular, monotone and $f(\emptyset) = 0$, i.e f is a β-function (Edmonds 1970). Hence, given a capacity constraint, the sampling problem in semiconductor manufacturing where the GSI is minimized can be modeled as the maximization of a submodular set-function.

The Global Sampling Indicator has already been used in industrial implementations with heuristic methods (Nduhura-Munga, Dauzère-Pérès, Vialletelle, and Yugma 2012), and the aim of our research is to evaluate the efficiency of simple heuristics and of an exact method to solve the sampling problem.

4 ALGORITHMS FOR SUBMODULAR SET FUNCTION MAXIMIZATION

A function f defined on all the subsets of a ground set V is submodular if it satisfies, for all $X, Y \subseteq V$:

$$f(X \cap Y) + f(X \cup Y) \leq f(X) + f(Y)$$

This property is similar to convexity for set functions. As in (Krause and Golovin 2014), we define the discrete derivative of f at S with respect to e: $\Delta_f(e|S) = f(S \cup \{e\}) - f(S)$. Another way to prove the submodularity and monotonicity of f is to show that $\Delta_f(e|A) \geq \Delta_f(e|B)\ \forall A \subset B$. This shows a major idea behind the concept of submodularity: it can be understood as decreasing gain or cost.

Several simple and more sophisticated algorithm approximations have been characterized in the worst case in (Nemhauser, Wolsey, and Fisher 1978). In this paper, we consider the greedy algorithm, the exchange procedure to enhance a given solution and a standard solver with a mixed integer linear programming formulation of the problem.

4.1 Integer Linear Programming (ILP) models

For the general case with multiple knapsack capacity constraint, the following variables are used to model the assignment of sampled lots to the parallel measurement tools:

- $y_{l,m} \in \{0,1\}$: Is equal to 1 if lot $l \in \mathscr{S}_m$, i.e. is assigned to metrology tool m, and 0 otherwise,
- $x_{r,l} \in \{0,1\}$: Is equal if lot l is used to decrease the risk level of risk r, and 0 otherwise.

The most general integer linear programming model with multiple knapsack capacity constraints is written below:

$$\max \sum_r \sum_l x_{r,l} D_{r,l} \tag{3}$$

$$\text{subject to: } x_{r,l} \leq \sum_{m \in \mathscr{M}} y_{l,m}, \qquad \forall\, l,r \tag{4}$$

$$\sum_l x_{r,l} \leq 1, \qquad \forall\, r \tag{5}$$

$$\sum_m y_{l,m} = 1, \qquad \forall\, l \in \mathscr{L}_M \tag{6}$$

$$\sum_m y_{l,m} \leq 1, \qquad \forall\, l \notin \mathscr{L}_M \tag{7}$$

$$\sum_l y_{l,m} t_{l,m} \leq T_m, \qquad \forall\, m \tag{8}$$

$$y_{l,m} \leq Q_{l,m}, \qquad \forall\, l,m \tag{9}$$

$$x_{r,l} \in \{0,1\}, \qquad \forall\, l,r \tag{10}$$

$$y_{l,m} \in \{0,1\}, \qquad \forall\, l,m \tag{11}$$

The objective function (3) is the total value of the information gathered by selecting a set of lots $\mathscr{S}$. Constraints (4) ensure that lot l is assigned to a metrology tool if it is used to decrease risk r. Constraints (5) ensure that at most one element of $\mathscr{L}$ is chosen to decrease risk r, which corresponts to the maximum in the objective function (2). Constraints (6) force mandatory lots to be selected. Constraints (8) limit the workload of the metrology tools to their maximum capacity, while Constraints (9) ensure that a lot is only assigned to a metrology tool on which the lot is qualified.

When a single knapsack capacity constraint is considered, i.e. a single measurement tool, we can simplify the notations. Let us consider now the variable $y_l \in \{0,1\}$, which Is equal to 1 if lot $l \in \mathscr{S}$, and 0 otherwise. In this case, the integer linear programming model above becomes:

$$\max \sum_r \sum_l x_{r,l} D_{r,l} \tag{12}$$

$$\text{subject to: } x_{r,l} \leq y_l, \forall\, l,r \tag{13}$$

$$\sum_l y_l . t_l \leq T \tag{14}$$

$$\sum_l x_{r,l} = 1, \forall\, r \tag{15}$$

$$x_{r,l} \in \{0,1\}, \forall\, l,r \tag{16}$$

$$y_l \in \{0,1\}, \forall\, l \tag{17}$$

With a cardinal capacity constraint, Constraint (14) in the integer linear programming model above is replaced by the following constraint:

$$\sum_l y_l \leq N \tag{18}$$

4.2 Greedy Algorithms

4.2.1 Multiple Knapsack Capacity Constraints

For the case of multiple knapsack capacity constraints, the assignment of the mandatory lots on the metrology tools is a major decision because it limits the available capacity that can be allocated for risk minimization.

As this assignment is a constraint, we decided to design a greedy heuristic that works in two phases: (1) The mandatory lots are first assigned to the qualified metrology tools, and (2) the non-mandatory lots are greedily assigned, and thus sampled, to a machine until the capacity constraints are reached.

The function ASSIGN MANDATORY in Algorithm 1 corresponds to the first phase, where all the mandatory lots of $\mathscr{L}_M$ are assigned to the metrology tools. Then, in the second phase, the function GREEDY greedily selects the remaining lots and assign them to the metrology tools until the capacity constraints are reached.

Algorithm 1 Greedy - Multiple knapsack

1: **procedure** SAMPLING($\mathscr{M}$, T, t, $\mathscr{L}$, $\mathscr{L}_M$, Q)
2: ASSIGN MANDATORY($\mathscr{M}$, $load$, T, t, $\mathscr{L}_M$, Q)
3: $\mathscr{S} \leftarrow$GREEDY(f, $\mathscr{L}$, $load$, $\mathscr{M}$, $\mathscr{L}_M$, T, t, Q)
4: **return** $\mathscr{S}$

4.2.2 Knapsack Capacity Constraint

The idea in Algorithm 2 is to value more a gain if it costs less capacity. To do so, the discrete derivative is divided by the time required to measure the element l: $\frac{\Delta_f(l|\mathscr{S})}{t_l}$. (Sviridenko 2004) proves that the modified greedy algorithm has an approximation in the worst case of $1-\frac{1}{e}$ for this knapsack case. Note that the $1-\frac{1}{e}$ approximation ratio was only proved in (Sviridenko 2004) in the case where $\mathscr{S}$ is initialized with the best set of cardinality 3. The procedure we used is initialized with $\mathscr{S}=\emptyset$. Hence, we do not have guaranteed performances.

Algorithm 2 Modified greedy algorithm

1: **procedure** GREEDY(f, T)
2: $\mathscr{S} \leftarrow \emptyset$
3: **while** $\sum_{t\in\mathscr{S}} t_l < T$ **do** ▷ Add until max capacity
4: $i \leftarrow \text{argmax}_l(\frac{\Delta_f(l|\mathscr{S})}{t_l})$ ▷ Find best element
5: $\mathscr{S} \leftarrow \mathscr{S} \cup \{\ l\ \}$
return $\mathscr{S}$

4.2.3 Cardinal Capacity Constraint

The greedy algorithm, introduced by Edmonds (Edmonds 1971) and recalled in Algorithm 3, is efficient enough to solve the submodular set function maximization in our industrial case.

Algorithm 3 Greedy algorithm

1: **procedure** GREEDY(f, N)
2: $\mathscr{S} \leftarrow \emptyset$
3: **while** $|\mathscr{S}| < \mathscr{N}$ **do** ▷ Add until max capacity
4: $i \leftarrow \text{argmax}_l(\Delta_f(l|\mathscr{S}))$ ▷ Find best element
5: $\mathscr{S} \leftarrow \mathscr{S} \cup \{\ l\ \}$
return $\mathscr{S}$

Beyond its simplicity, the greedy algorithm often comes up with very good solutions, far superior to its $1-\frac{1}{e}$ guaranteed approximation ratio. In a simulation study on the selection of sensors, (Shamaiah, Banerjee, and Vikalo 2010) show that it performs better than competing techniques based on convex relaxation. (Soma and Yoshida 2017) observe that its performance is much better on real instances, sometimes reaching the optimal solution, and explains this phenomenon by studying the properties of concave relaxation specific to

the problem. This research is closely linked to the notion of curvature defined by (Conforti and Cornuéjols 1984).

4.3 The Exchange Procedure

To enhance the solutions computed by the greedy algorithm, a local search is carried out using an exchange procedure (Nemhauser, Wolsey, and Fisher 1978). For each lot $l \in \mathscr{S}$, we seek if there is a better lot $e \in \mathscr{L}$. If several good exchanges are found, the one with the maximum gain is selected. The process continues as long as carrying out an exchange brings a strictly positive gain.

5 COMPUTATIONAL EXPERIMENTS

Computational experiments have been conducted to validate the approach and study the performance of the algorithms described in the previous sections on the sampling problem with multiple knapsack, knapsack and cardinal capacity constraints. We want to study the performance of the integer linear programs and the heuristics on three types of instances: Industrial, random-industrial and random.

The integer linear programs were solved with the open source solver CBC-COIN-OR (Forrest, Vigerske, Santos, Ralphs, Hafer, Kristjansson, jpfasano, EdwinStraver, Lubin, rlougee, jpgoncal1, h-i gassmann, and Saltzman 2020), with a time limit of 180 seconds.

5.1 Data Generation

The matrix D consists of Wafer At Risk (W@R) values. We used the GSI indicator with $\alpha = 1$ and $IL_r = 1$, $\forall r \in \mathscr{R}$ and the following instances:

- *Industrial instances:* The number of lots, the number of metrology tools and their characteristics (measurement times, etc.), and the W@R values in D are taken from historical data,
- *Random-industrial instances:* The non-null coefficients in each matrix D of the industrial data are replaced by random numbers drawn from the distribution of risks observed in the historical data,
- *Random instances:* All coefficients in each matrix D of the industrial data are replaced by a random number generated between 0 and 1 with a uniform distribution.

We believe that the industrial instances are easy to solve because they have a very specific structure:

- The production routes lead to many null coefficients in the matrix D. This is broken in the random instances, but preserved in the random-industrial instances.
- The priorities (speed) of lots in the factory lead to dominance between lots: When a lot contributes strongly to a solution, it is often an important contributor to several risks. This is broken in both the random and random-industrial instances.

5.2 Multiple Knapsack Capacity Constraints

To evaluate the performance of a set of sampled lots, we compute the approximation ratio below without considering the mandatory lots on which no decisions are made.

$$\text{ratio} = \frac{f(S_{\text{greedy}}) - f(\mathscr{L}_M)}{f(S_{\text{solver}}) - f(\mathscr{L}_M)} \tag{19}$$

The approximation ratios of the greedy algorithms shown in Table 1 are on average good enough for an industrial implementation. On the 30 industrial instances, the solver often managed to compute the optimal solution in less than one minute. It is also interesting to measure how often the various approaches find the optimal solution. The associated percentages are given in Table 1. As this general problem is

dealing with both coverage and assignment, it is expected that the greedy algorithm, combined or not with the exchange procedure, does not find optimal solutions as often as with simpler capacity constraints.

Table 1: Overall performances for multiple knapsack capacity constraints

	Average approximation ratio			**Optimal solution found**		
Instance type	**Greedy**	**Exchange**	**Solver**	**Greedy**	**Exchange**	**Solver**
random	94.50%	94.63%	99.34%	3.33%	3.33%	76.67%
random-industrial	82.70%	83.62%	98.93%	6.67%	6.67%	90.00%
industrial	79.52%	80.10%	100%	10.00%	10.00%	100%

5.3 Knapsack Capacity Constraint

The heuristics again perform well on average, as shown in Table 2. The greedy algorithm improved by the exchange procedure reaches the optimal solution in almost 60% of the cases on the industrial instances. There are only few random instances (0.6%) for which the solver CBC-COIN-OR was not able to provide the optimal solution in the given computational time of 180 seconds. For these cases, we evaluated the approximation ratio by comparing to the relaxed solution value. Note that we can identify some cases where we have instances with approximation ratio below $1-\frac{1}{e}$, because we do not initialize with the best set of cardinality 3, so we do not have guaranteed performances. Such an initialization is not always feasible because it might exceed the capacity constraint.

Table 2: Overall performances for knapsack capacity constraint

	Average approximation ratio			**Optimal solution found**		
Instance type	**Greedy**	**Exchange**	**Solver**	**Greedy**	**Exchange**	**Solver**
random	99.68%	99.76%	99.39%	31.25%	36.33%	99.38%
random-industrial	98.92%	99.05%	100%	48.51%	54.95%	100%
industrial	99.00%	99.10%	100%	52.19%	59.54%	100%

5.4 Cardinal Capacity Constraint

Table 3 compares the performances of the various approaches for the cardinal capacity constraint. The average approximation ratio for the greedy algorithm enhanced with the exchange procedure is very good. The greedy algorithm reaches the optimal solution in more than 99% of the instances for industrial and random-industrial instances. The exchange procedure enhances the solution significantly and helps to reach the optimal solution in almost all cases on industrial and random-industrial instances. This reinforces our belief that structure of the industrial instances makes the problem easier to solve.

Table 3: Average approximation ratio for cardinal capacity constraint

	Average approximation ratio			**Optimal solution found**		
Instance type	**Greedy**	**Exchange**	**Solver**	**Greedy**	**Exchange**	**Solver**
random	99.90%	99.98%	100%	92.18%	98.66%	100%
random-industrial	99.99%	99.99%	100%	99.51%	99.88%	100%
industrial	99.99%	100%	100%	99.71%	100%	100%

6 CONCLUSIONS AND PERSPECTIVES

We formalized the dynamic sampling problem for risk minimization as the maximization of a submodular function subject to various types of capacity constraints. We have shown that the considered problems are NP-hard. We also evaluated the performances of standard heuristics and of integer linear programming models on three types of instances: Randomly generated, random-industrial instances derived from the industrial ones and industrial instances. The computational experiments showed that the open source solver CBC-COIN-OR is efficient enough to solve the industrial instances in less than 180 seconds, even the most complicated instances with multiple knapsack capacity constraints. Standard greedy heuristics, combined with an exchange procedure, find solutions that are good enough for an industrial implementation, with a particularly high approximation ratio (above 99% on average) when knapsack or cardinal capacity constraints are considered. These methods are being deployed at Soitec to manage the dynamic sampling of lots in the 300mm factory.

Two research perspectives are being investigated. The first one aims at using the data collected from the sensors in production machines, and the findings in designing machine health indices (Chen and Blue 2009) and in Virtual Metrology (Khan, Moyne, and Tilbury 2007), to prioritize the measurements of lots that provide information on machines with poor health indices. In the second research perspective, we study approaches where lots will not only be assigned to the non-identical parallel metrology tools, but also be scheduled on them to better manage the measurement capacity. Our goal is to integrate sampling and scheduling decisions on metrology tools.

ACKNOWLEDGEMENTS

This work has been partially financed by the ANRT (Association Nationale de la Recherche et de la Technologie) through the PhD number 2018/0149 with CIFRE funds and a cooperation contract between SOITEC and ARMINES.

REFERENCES

Bean, J. 2008, 11. "Variation reduction in a wafer fabrication line through inspection optimization".

Bouslah, B., A. Gharbi, and R. Pellerin. 2016. "Integrated production, sampling quality control and maintenance of deteriorating production systems with AOQL constraint". *Omega* 61:110–126.

Chen, A., and J. Blue. 2009. "Recipe-independent indicator for tool health diagnosis and predictive maintenance". *IEEE Transactions on Semiconductor Manufacturing* 22(4):522–535.

Conforti, M., and G. Cornuéjols. 1984. "Submodular set functions, matroids and the greedy algorithm: tight worst-case bounds and some generalizations of the Rado-Edmonds theorem". *Discrete applied mathematics* 7(3):251–274.

Cornuejols, G., M. L. Fisher, and G. L. Nemhauser. 1977. "Location of bank accounts to optimize float: An analytic study of exact and approximate algorithms". *Management science* 23(8):789–810.

Dauzère-Pérès, S., J.-L. Rouveyrol, C. Yugma, and P. Vialletelle. 2010. "A smart sampling algorithm to minimize risk dynamically". In *Advanced semiconductor manufacturing conference (ASMC), 2010 IEEE/SEMI*, 307–310. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Edmonds, J. 1970. "Submodular functions, matroids, and certain polyhedra". *Combinatorial structures and their applications*:69–87.

Edmonds, J. 1971. "Matroids and the greedy algorithm". *Mathematical programming* 1(1):127–136.

Feigenbaum, A. V. 1951. *Quality control: Principles, practice and administration: An industrial management tool for improving product quality and design and for reducing operating costs and losses*. McGraw-Hill.

John J. Forrest and Stefan Vigerske and Haroldo Gambini Santos and Ted Ralphs and Lou Hafer and Bjarni Kristjansson and jpfasano and EdwinStraver and Miles Lubin and rlougee and jpgoncal1 and h-i-gassmann and Matthew Saltzman 2020, March. "coin-or/Cbc: Version 2.10.5".

Holton, G. A. 2004. "Defining risk". *Financial Analysts Journal* 60(6):19–25.

Jorion, P. 2007. *Value at Risk-The New Benchmark for Managing Financial Risk*. 3 ed. McGraw-Hill.

Khan, A. A., J. R. Moyne, and D. M. Tilbury. 2007. "An approach for factory-wide control utilizing virtual metrology". *IEEE Transactions on semiconductor Manufacturing* 20(4):364–375.

Krause, Andreas and Golovin, Daniel 2014. "Submodular function maximization.".

Nduhura-Munga, J., S. Dauzère-Pérès, P. Vialletelle, and C. Yugma. 2012. "Industrial implementation of a dynamic sampling algorithm in semiconductor manufacturing: Approach and challenges". In *Proceedings of the 2012 Winter Simulation Conference (WSC)*, edited by O. Rose, A. Uhrmacher, M. Rabe, C. Laroque, R. Pasupathy, and J. Himmelspach, 1–9. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Nduhura-Munga, J., G. Rodriguez-Verjan, S. Dauzère-Pérès, C. Yugma, P. Vialletelle, and J. Pinaton. 2013. "A literature review on sampling techniques in semiconductor manufacturing". *IEEE Transactions on Semiconductor Manufacturing* 26(2):188–195.

Nemhauser, G. L., L. A. Wolsey, and M. L. Fisher. 1978. "An analysis of approximations for maximizing submodular set functions—I". *Mathematical programming* 14(1):265–294.

Rockafellar, R. T., S. Uryasev et al. 2000. "Optimization of conditional value-at-risk". *Journal of risk* 2:21–42.

Rodriguez-Verjan, G. L., S. Dauzère-Pérès, S. Housseman, and J. Pinaton. 2013. "Skipping algorithms for defect inspection using a dynamic control strategy in semiconductor manufacturing". In *Proceedings of the 2013 Winter Simulations Conference*, edited by R. Hill, M. Kuhl, R. Pasupathy, S.-H. Kim, and A. Tolk, 3684–3695. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Sendón, A., S. Dauzère-Pérès, and J. Pinaton. 2015. "Simulation model to control risk levels on process equipment through metrology in semiconductor manufacturing". In *Proceedings of the 2015 Winter Simulation Conference*, edited by C. M. Macal, M. D. Rossetti, L. Yilmaz, I.-C. Moon, W. K. Chan, and T. Roeder, 2941–2952. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Shamaiah, M., S. Banerjee, and H. Vikalo. 2010. "Greedy sensor selection: Leveraging submodularity". In *Decision and Control (CDC), 2010 49th IEEE Conference on*, 2572–2577. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Soma, T., and Y. Yoshida. 2017. "A New Approximation Guarantee for Monotone Submodular Function Maximization via Discrete Convexity". *arXiv preprint arXiv:1709.02910*.

Sviridenko, M. 2004. "A note on maximizing a submodular set function subject to a knapsack constraint". *Operations Research Letters* 32(1):41–43.

AUTHOR BIOGRAPHIES

ÉTIENNE LE QUÉRÉ is a Ph.D. student at Ecole Nationale des Mines de Saint-Etienne, France. He works as an Engineer in the Yield and Statistic Process Control Department at Soitec Bernin. He received the Master's degree in Industrial Engineering and Operations Research from Ecole Nationale des Mines de Saint-Etienne, France in 2017. His Ph.D. subject is on modeling and minimizing risk in semiconductor manufacturing. His email address is etienne.lequere@soitec.com

STÉPHANE DAUZÈRE-PÉRÈS is Professor at the Center of Microelectronics in Provence (CMP) of Mines Saint-Etienne in France and Adjunct Professor at BI Norwegian Business School in Norway. He received the Ph.D. degree from the Paul Sabatier University in Toulouse, France, in 1992; and the H.D.R. from the Pierre and Marie Curie University, Paris, France, in 1998. He was a Postdoctoral Fellow at the Massachusetts Institute of Technology, U.S.A., in 1992 and 1993, and Research Scientist at Erasmus University Rotterdam, The Netherlands, in 1994. He has been Associate Professor and Professor from 1994 to 2004 at the Ecole des Mines de Nantes in France. His research interests broadly include modeling and optimization of operations at various decision levels (from real-time to strategic) in manufacturing and logistics, with a special emphasis on semiconductor manufacturing. He has published 84 papers in international journals. He has coordinated multiple academic and industrial research projects. He was runner-up in 2006 of the Franz Edelman Award Competition, and won the Best Applied Paper of the Winter Simulation Conference in 2013. His email address is dauzere-peres@emse.fr.

KARIM TAMSSAOUET is a postdoctoral researcher at the Center of Microelectronics in Provence (CMP) of Mines Saint-Etienne in France. In 2019, he completed his Ph.D thesis on scheduling in semiconductor manufacturing. He reveived an M.Sc degree in Supply Chain Management from Paris Dauphine University in 2015 and he studied Industrial Engineering at École Nationale Polytechnique of Algiers where he graduate in 2014. His email adress is karim.tamssaouet@emse.fr.

CÉDRIC MAUFRONT works as a technical leader within the yield, sampling and process control system group at Soitec. He received a Ph.D degree from Paris-Sud Orsay University in 2005. His PhD subject was on characterizations and definitions of MRAM architectures. He also has an engineering degree in material science from Polytech Paris-Sud, as well as a MS in nanotechnology from Paris-Sud Orsay University. After holding several positions at ST, he joined Soitec as a senior yield enginer in 2011. His email address is cedric.maufront@soitec.com.

STÉPHANE ASTIE is the manager of a team with data analysts and six sigma experts to improve products Yield and process control deployment. He received a Ph.D in Material Science from the Paul Sabatier University in Toulouse, France, in 1998

and has a Six Sigma Black Belt. He worked at Motorola/Freescale semiconductor for 15 years, in R&D, Defectivity and Device teams before joining Soitec in 2011. His email address is stephane.astie@soitec.com.