Simulation and integer programming for efficient shift design

Lorraine Trilling, Frédéric Albert, Eric Marcon

SIMULATION AND INTEGER PROGRAMMING APPROACH
FOR EFFICIENT SHIFT DESIGN

Lorraine Trilling*, Frédéric Albert and Eric Marcon****

**Laboratoire PRISMa, INSA de lyon, France*
Email: lorraine.trilling@insa-lyon.fr

***Laboratoire d'Analyse des Signaux et des Processus Industriels, Roanne, France*
E-mail: frederic.albert@univ-st-etienne.fr

Abstract: Operation research approaches and heuristics are often used in the staff organization literature. The main hypothesis considered by shift scheduling approaches is that the workload variability is negligible. Faced to a real organization problem of cleaning staff in a multidisciplinary surgical suite, the lack of exploratory approaches adapted to a context of variable activity has been observed. The purpose of this paper is to implement a hybrid approach combining integer linear programming and a simulation model for assessing performance. Finally, the hybrid approach can save 30% of the cost compared to the solution given by operation research approach alone.

Keywords: Integer programming, discrete-event simulation, shift scheduling algorithm, heuristic algorithm.

1. INTRODUCTION

The classical process of scheduling staff in any organization includes several stages (Tien and Kamiyama, 1982). The first one is to determine the number of required employees for each time slot of the planning period. In other words, the purpose of this stage is to build the workforce requirement curve that would have to be covered. Next, the shifts have to be designed. These shifts as well as the days-off are then assigned to the employees in order to produce individual schedules. The design and the assignment of the shifts can be coordinated and solved in the same problem, or they can be analyzed separately. The following of this paper focuses on the shift scheduling which designate the problem of designing and determining the number of employees required for each shift, without considering the shift assignment.

Over the last decades the shift scheduling problem has been addressed by many scientists. The first integer programming formulation to the shift scheduling problem is a set covering type model proposed by Danzig (1954), which aims to find the requirement curve coverage that minimizes cost of the shifts. In this formulation type the number of variables corresponds to the number of feasible shifts that increases when a large number of shift start times, shift lengths, break placements and break lengths are considered. Feasible shifts have to be first explicitly enumerated before determining the number of employees for each selected shift, in order to cover the requirement curve and optimize the cost function.

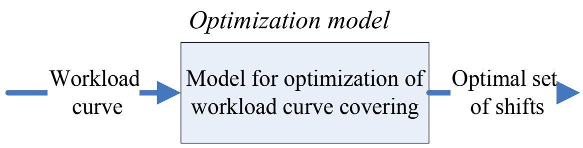

Figure 1: Model for workload curve covering

Some authors make use of the advantages of the set covering model (ease of modeling) and overcome the difficulties inherent to this model recognized in case a great number of feasible shifts, by developing their own implicit optimization technique based on column generation (Mehrotra et al., 2000).

Other authors modeled shifts implicitly rather that explicitly, but still using integer programming (Bechtold and Jacobs, 1991). In implicit approaches, shifts are no more listed, but constructed while the solution is generated. The variables are the shift start and end times, the shift lengths and the number of employees required for each shift. More recently, Musliu et al. (2004) propose an implicit model by using local search methods. The objective is to minimize the number of distinct shifts, over- and under-staffing, and the differences in the average of duties per week. After the generation of a good initial solution obtained by the choice of a suitable, they use basic principles of tabu search for the exploration of the neighborhood and choose the most accurate moves.

1.1. Weaknesses of the current operation research approaches

In the approaches presented above the main objective is to achieve the best coverage of the workload. This optimization criterion is focused on the minimization of the difference between the workload curve and the envelope of the set of shifts. This kind of approach has two main disadvantages: (1) the quality of the shift coverage is related to the quality of the workload curve; (2) the optimization criterion used is not the real criterion that must be optimized but a surrogate. Indeed the real optimization goal is to find a shift coverage that minimizes different endpoints of the process, to minimize either the completion time for all the tasks, or the delays caused by a shortage of resources. Thus, we can easily think that an optimal solution provided by these approaches will lead to poor performances when the shifts would be implemented on real world processes.

In the following of this section, we develop these two disadvantages in order to justify the relevancy of our approach. Firstly, the workload curve is commonly the average workload curve or another one upper than 50th percentile curve (i.e., 70th, 80th or more). Whatever the curve chosen, it would be hard to assess, a priori, the relevancy of this choice. In the different papers we studied, the authors did not suggest any methods for choosing the best curve. We think that the choice of the good curve will contribute to the search for an optimal solution particularly when the workload could change from one period to another (i.e., one workday, one week…). For example: the hourly workload curve for the Monday is obtained by merging the hourly workload of the last 6 months Mondays. In this situation, if there would have been a large variation between the different Mondays for the same range of hours, using the average or another percentile would not be accurate.

Secondly, optimization criterion based on the best coverage is not relevant because the over-coverage of some hours as well as the under-coverage during some other hours do not have the same impact on the process functioning and thus on the shifts' end time. For example, the performance of a system could be assessed by the maximum of delays for starting the tasks. In that case, the over-coverage will not generate some delays whereas shortage of resources will produce some.

In their paper the authors did not give any arguments on how to build an objective function that takes into account the impact of the over- and under-coverage (except in salary cost objective).

1.2. Application field of our contributions

The workload curve design problem gathers cases of shift scheduling problems in which the workload variability cannot be neglected. This workload could vary for different reason, mainly caused by the demand and duration variation. Many industrial processes are faced to this reality, especially: (1) services subject to an unpredictable demand like call centers, emergency services, maintenance services…; (2) but also processes in which the scheduled activity takes into account only the most critical resources, whereas non-critical resources have to perform a variable workload.

This problematic is studied in different fields as: mail processing centre (Júdice et al., 2005), maintenance (Slomp and Sureshb, 2005), army logistics, call centres, airlines companies, industrial plants, and hospitals (Ernst et al., 2004).

1.3. Approaches based on simulation

Shift scheduling approaches using the simulation techniques seem to overcome some of the drawbacks of the studies presented above. Ernst et al. (2004) define the simulation as "a technique for imitating the behavior of a real system by means of an analogous computer model. The cause-and-effect relationships of a system are captured in the simulation model, which is then used to predict the behavior of the system. One of the main uses of simulation is to carry out what-if analyses on different system scenarios." This technique is particularly efficient for describing, and assessing performances of a system, as well as assessing the performance of a shift schedule in our case.

Among these application fields, our interest is focused on transversal staff of the Surgical Suite (SS), and more precisely on the cleaning staff who sets up the Operating Room (OR) between two surgeries.

The SS activity is scheduled considering the most critical resources, such as, ORs, surgeons and technical facilities. Moreover, the case duration can vary significantly depending on unforeseeable medical factors. The consequences of these two statements are that the cleaning workload significantly varies for the transversal cleaning staff.

In hospital field, we found few publications using simulation for the staffing. The studies mainly tackle problems met in the emergency room (Kumar and Kapur, 1989; Draeger, 1992; Evans et al., 1996;

Rossetti et al., 1999), and are based on the comparison of several schedule alternatives (i.e., less than 10). These approaches do not ensure whether the a priori alternatives tested are good ones or not. According to this, finding the best staffing should be to run through the entire space of alternatives. Unfortunately, this approach would last for unreasonable time for a computing system.

To avoid such drawbacks, we suggest an approach that uses a better approximation of the workload curve and uses as objective function the endpoints of the real system. This approach is based on a hybrid model built on two sub-models. The first sub-model is based on a simulation of flow model. The second sub-model is a heuristic model for optimization. The two sub-models work together in an iterative process where the simulation flow model calculates the workload curve, as well as the value of the endpoints of the objective function. The heuristic process calculates, at each step of the iterative process, the new neighbor shifts that will minimize the objective function.

This paper is organized as follows. Section 2 is dedicated to explain in what field of application the classical approaches of the shift scheduling problems are not accurate. In section 3, we present the structure of our approach based on simulation and optimization tools, and the description of the industrial case. In section 4, two hybrid methods involving specific algorithms are detailed and applied to solve the same problem. Finally, section 5 compares the results obtained using classical and hybrid approaches and draws conclusions.

2. ASSESSMENT OF SHIFTS COVERING APPROACH

This section is focused on the performance assessment of set coverage algorithm obtained by using integer linear programming.

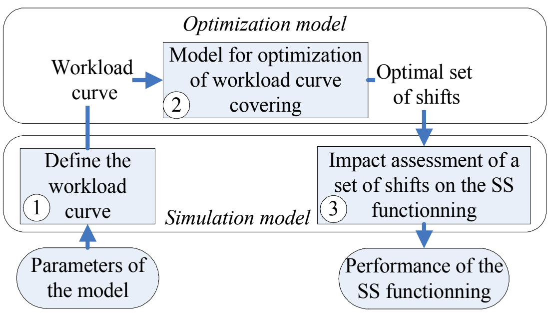

Figure 2: Process for assessing the performances of shift covering

In order to evaluate the real performances of the workload coverage, it is necessary to assess its behavior in a real world process. As shows in figure 2 a simulation model is used to calculate the workload curve to be covered (stage 1), and to assess the service performances with performance indicators (stage 3).

2.1. *Performance assessment*

First, it is necessary to define a base of common considerations in order to compare results of each approach. In that way we combine two cost Performance Indicators (PI). Those costs vary depending on the quality of the cleaning staff workload coverage. The total cost of workload coverage can be split up in two terms: the Staff Costs (SC) corresponding to the cleaning staff salaries; and the Indirect Cost (IC) generated by the malfunctioning of the SS. Costs are assessed in a specific working period defined in the application case.

- SC is the sum of working hours of the staff during the shifts.
- IC represents the efficiency of the SS, or in our case the inefficiency of the shift coverage.

After having tested various sets of shifts, we can notice that the two performance indicators move in two opposite ways. Generally, if the cleaning staff number increases, SC also increases but IC decreases and vice versa. The best performance will be a compromise between the value of IC and SC, minimizing the total cost (since total cost = IC+SC).

2.2. *Simulation model*

The simulation model describes the functioning of main process in an SS. In the case of a transversal cleaning staff, the simulation model is made up by the main surgical process, the cleaning process and the logistic process. In our model Nursing Auxiliary (NA) and Housekeeper (HK) perform the logistic and cleaning processes (figure 3).

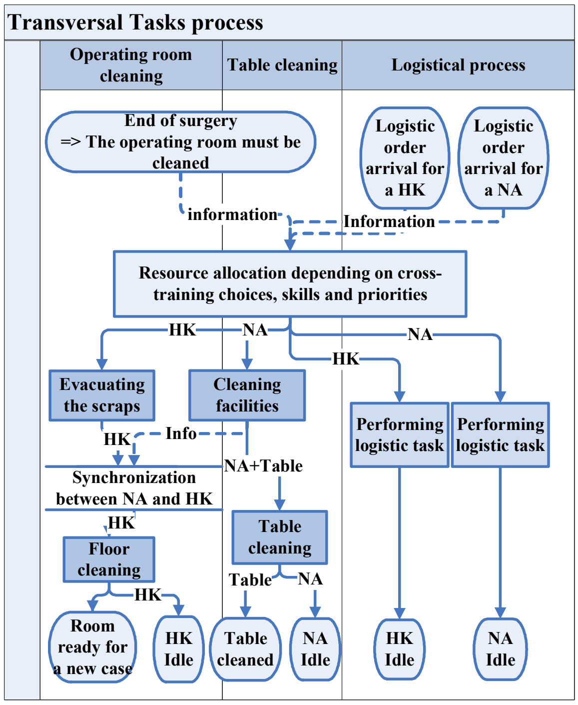

Figure 3: Cleaning and logistic processes

The OR cleaning is an important process during which operating scraps, medical supplies and linen are evacuated and OR facilities are cleaned before a new surgery. Every delay in this process implies a direct delay of the main operating process. When transversal staff has to choose between two tasks, cleaning tasks have the priority.

The logistic process includes different tasks occurring during the working day independently of the main operating process. These tasks are: (1) the delivery reception of the sterilized medical devices, or of linen (performed by the NA); (2) the evacuation of the scraps container, or the stained linen containers (performed by the HK).

For the cleaning process, orders appear at the end of a surgery. To perform the process, a Nursing Auxiliary (NA) and a Housekeeper (HK) work in parallel. When the NA cleans the equipments, the HK evacuates scraps. For cleaning floors, the HK must wait until the NA has left the OR with the operating table. Then, the floors is cleaned, the OR is ready for a new surgery

2.3. Optimization of workload curve covering

To give an example of curve covering approach, an optimization integer linear programming (ILP) model is used to find the shifts and the number of employees per shift minimizing the salary cost (PC).

We suggest to use the common approach, which is the optimization model derived from the Danzig (1954) set covering formulation that can be expressed as following:

$$Minimize \quad Z=\sum_{i=1}^{n} c_i X_i \qquad (1)$$

Subject to

$$\sum_{i=1}^{n} a_{ij} X_i \geq P \times b_j \qquad \forall j = 1,...m \qquad (2)$$

$$X_i \geq 0 \qquad \forall i = 1,...n \qquad (3)$$

Where n represents the number of feasible shifts, m is the number of time slots in the planning period (here it corresponds to the day), b_j is the estimated average number of workers required for the period j, c_i is the cost of assigning a worker to the shift i, and a_{ij} is equal to one if period j is a work period for shift i and zero otherwise. X_i is an integer variable defined as the number of workers required for the shift i. P is a parameter corresponding to the percentage of the workload curve covering that must be respected (it could be 100%, or less).

The 0-1 matrix a_{ij} is obtained by the generation of the set of feasible shifts. Each shift has the same effective time duration (8 hours), and can include a one-hour break or not. The breaks are placed so that employees work at least two hours before the break and at least one hour after.

The solution given by this optimization model is the set of integer values assigned to each variable X_i that satisfies the workload coverage constraint (2) and the non-negative constraint (3) while minimizing the global salary cost (1).

2.4. Implementation on the application case

In our application case, we consider a multi-disciplinary surgical suite (SS) with 20 ORs. The simulation duration is set up to 6 months that represent 130 working days.

Surgeries durations follow a lognormal distribution with a mean between 60 and 100 min depending on the operating room. We consider that the scheduling process is already done and that the scheduled occupancy rate is set to 80%. The surgical cases in an OR are scheduled by duration the longest first.

The arrival of cleaning orders is directly linked to the surgery end time. The duration of a cleaning process varies from 10 min to 20 min.

Logistic tasks occur with a uniform distribution from 30 min after the start of the work and 60 min before the end. Their duration is distributed from 12 to 20 min for HK logistic tasks and from 30 to 45 min for NA logistic tasks. The workload resulting of the logistic tasks is divided: 70% for the NA staff and 30% for the HK staff.

The staff and indirect costs are defined as following:

- SC is the sum of working hours of the staff shifts considering a cost per hour for 1 NA + 1 HK around 40 €;
- IC represents the efficiency of the SS. It comprises the over-utilized OR cost and the transversal personnel overtime cost. The over-utilized OR cost is set to 230 € for one OR per hour. The transversal personnel overtime cost is set to 40 € per hour.

When analyzing performance of the integer linear programming model alone, we can define a first solution cost, such that IC = 0 €; SC = 327,000 € and the total Cost of Workload coverage solution set to 320,000 €.

3. HYBRID APPROACHES

3.1. Hybrid approach focused on optimization model

To overcome the weaknesses of the coverage approach presented in the previous section, we suggest the use of a simulation model and the curve coverage model displayed on figure 4.

A first workload curve is defined by using the simulation model (stage 1). This curve expresses the average workload for each time slot calculated over a great number of simulations (around 500). We call this curve the *P*-curve where *P* is the parameter used in equation 2 (see previous section) that corresponds to the percentage of the average workload curve we want to cover (for the first curve P is equal to 100%, i.e., 1). Secondly, this first P-curve is the input of the optimization model based on ILP (stage 2) in order to generate the optimal set of shifts, corresponding to the optimal coverage of the workload P-curve. The performance of optimal coverage is then assessed (stage 3) on the same simulation model used to generate the first curve. The performance indicators of this solution are then saved.

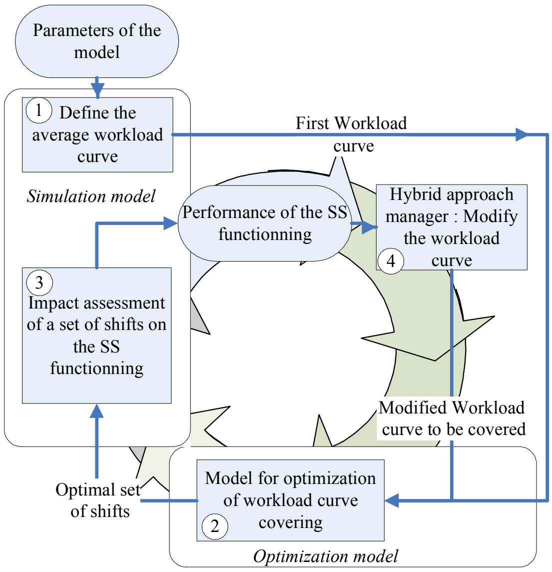

Figure 4: Hybrid approach focused on optimization model

The hybrid approach manager (stage 4) creates a new workload P-curve built with P decremented from a step of S% (i.e., 5% for this instance) and a new optimal coverage is generated. While the performance of this solution is better than the last performance saved, the manager keeps going decrementing P from S% (sweep #1). If the performance begins decreasing significantly (i.e., more than 3% of the solution where $P = 100\%$), the manager changes the value of S to $S/2$. He moves to increment P from S% (sweep #2), and so on. The manager decides to stop the search for the optimal value of P when the minimal value of the step is reached or when there are no more performance gap between two consecutive solutions.

This method has been applied to the cleaning process presented in section 2.2, and thus to the auxiliary nurses (NA) shift-scheduling problem. Results are presented on figure 5.

The best performance of the SS functioning has been found, after 20 min running, for P = 82.5 %. In this diagram we can see that all the solutions with $P \in [62.5, 82.5]$ seem to have almost the same cost, and are equivalent. However the decision maker would rather choose a solution, where the staff cost has a more important part in the total cost than the indirect cost (i.e., $P \in [75, 82.5]$). As displayed on the figure 5, this solution achieved by the hybrid approach saves a total cost reduction more than 70,000 € compared to the 100% of the average curve covering solution.

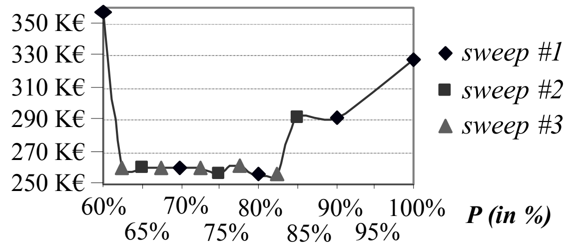

Figure 5: Results of the hybrid approach focused on optimization model

3.2. *Hybrid approach focused on simulation*

We found interesting to compare the previous hybrid approach with a different one. In this new approach, we do not consider the notion of workload curve anymore but we use a new approach focused on simulation. The algorithm explained in the following paragraph does not claim to reach an optimal solution, but just to find a satisfying solution.

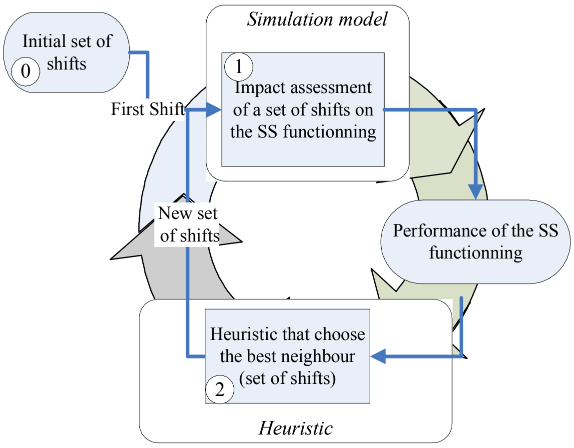

Figure 6: Hybrid approach focused on simulation approach

Figure 6 presents the main steps of this second hybrid approach. The algorithm starts (stage 0) from an extremely bad solution (i.e., 10 staff of each type of shift), and then the simulation model (stage 1) assesses the global performance and defines the waiting rate for each hour of workday.

The heuristic (stage 2) uses the curve describing the waiting rate of surgeries for defining the time slot which has the lowest waiting time rate. This algorithm assumes that when the waiting time is the lowest, available staffs are too numerous. As a consequence, the algorithm tests (i.e., assess (stage 1) the performance by the simulation) every elementary changes of shift that could reduce the number of available staffs for this specific time slot. Among these neighbor solutions the heuristic chooses the best one and then considers this solution as the current one for the next step. If the algorithm does not succeed in finding a neighbor solution at the first chosen time slot, it tries at the next time slot which has the lowest waiting time rate directly upper and so on. The algorithm stops when for each time slot of the day, the neighborhood solutions are less interesting than the current one.

When analyzing performance of the second hybrid approach, we obtain a single resulting set of shifts, such that IC = 32,000 €; SC = 182,000 € and the total cost of workload coverage solution is equal to 214,000 €.

3.3. Results comparison

The application of the approaches on a real world surgical suite context highlights the inefficient results of the operation research approach alone. In this application case, the hybridizing of sub-models allows to achieve benefits from 21% to 33%.

Table 1: Synthesis of approach performances

	Staff Cost	Indirect Cost	Total Cost
OR alone approach	327,000 €	0 €	327,000 €
1st hybrid approach 82.5%	255,000 €	3,000 €	257,000 €
1st hybrid approach 65%	218,000 €	43,000 €	261,000 €
2nd hybrid approach	182,000 €	32,000 €	214,000 €

If we compare the two hybrid approaches, we notice that the hybrid approach focused on optimization model achieves a good solution in less time (i.e., 20 min) than the hybrid approach focused on simulation (i.e., 6 hours). The optimization hybrid approach covers the workload in a global way that limits the set of shifts assessed. If the hybrid manager would adjust the coverage percentage for each time slot, it would probably give best results, but it would also increase significantly the number of simulations to perform.

The hybrid approach focused on simulation works in a different way. The hybrid manager is focused on the delay time smoothing all day long. This second objective allows to reduce the number of alternatives tested by reducing the neighborhood of solutions. For the field studied, the hybrid approach focused on simulation gives the best results compared to the others approaches.

4. CONCLUSION

The main hypothesis considered by current shift scheduling approaches is that the workload variability is negligible. Faced to a real organization problem we were challenged by the lack of exploratory approaches adapted to a context with variable activity. Using simulation and optimization heuristics, the hybrid approaches that we suggest for solving the shift scheduling problem allow to assess a cost function closer to real world case than an operation research approach alone. These approaches give solutions that allow achieving a decrease in staff cost between 21% and 33%. In our study, the economical gain achieves more than 70,000 € for 6 months. After having highlighted possible benefits of these approaches, some tracks for future works can be drawn: (1) the improvement of the current algorithms by developing more efficient ones, (2) the experimentation of other approaches

ACKNOWLEDGMENTS

This study has been performed within the scope of the HRP² French regional project.

REFERENCES

Bechtold, S. E. and L. W. Jacobs (1991). Improvement of labour utilisation in shift scheduling for services with implicit optimal modeling. *International Jounral Operation Production Management*, **11**, 54-69.

Dantzig, G. B. (1954). A comment on EDIE's 'Traffic delays at toll booths'. *Operations Research*, **2**, 339-341.

Draeger, M. (1992). An emergency department simulation model used to evaluate alternative Nurse staffing and patient population scenarios,. *Proceedings of 1992 Winter Simulation Conference IEEE*, Arlington, VA.

Ernst, A. T., H. Jiang, M. Krishnamoorthy, B. Owens and D. Sier (2004). An annotated bibliography of personnel scheduling and rostering. *Annals of Operations Research*, **127**, 21-144.

Evans, G. W., T. B. Gor and E. Unger (1996). A simulation model for evaluating personnel schedules in a hospital emergency department. *Proceedings of 1996 Winter Simulation Conference IEEE*, Coronado, California, USA.

Júdice, J., P. Martins and J. Nunes (2005). Workforce planning in a lotsizing mail processing problem. *Computers & Operations Research*, **32**, 3031-3058.

Kumar, A. and R. Kapur (1989). Discrete simulation application-Scheduling staff for the emergency room. *Proceedings of 1989 Winter Simulation Conference IEEE*, Washington, DC.

Mehrotra, A., K. E. Murphy and M. A. Trick (2000). Optimal Shift Scheduling: A Branch-and-Price Approach. *Naval Research Logistics*, **47**, 185-200.

Musliu, N., A. Shaerf and W. Slany (2004). Local search for shift design. *European Journal of Operational Research*, **153**, 51-64.

Rossetti, M. D., G. F. Trzcinski and S. A. Syverud (1999). Emergency Department Simulation and Determination of Optimal Attending Physician Staffing Schedules. *Proceedings of 1999 Winter Simulation Conference*.

Slomp, J. and N. C. Sureshb (2005). The shift team formation problem in multi-shift manufacturing operations. *European Journal of Operational Research*, **165**, 708-728.

Tien, J. and A. Kamiyama (1982). On manpower scheduling algorithms. *SIAM Review*, **24**, 275-287.