Optimal outpatient appointment scheduling

Guido C. Kaandorp[†] & Ger Koole[‡]

[†] Department of Epidemiology & Biostatistics, Erasmus Medical Center Rotterdam, The Netherlands

[‡] Department of Mathematics, VU University Amsterdam, The Netherlands

Health Care Management Science **10**:217–229, 2007

Corresponding author: Ger Koole
Department of Mathematics, Vrije Universiteit
De Boelelaan 1081a, 1081 HV Amsterdam, Holland
Tel. +31 20 5987755, fax +31 20 5987653
koole@few.vu.nl

Optimal outpatient appointment scheduling

Abstract

In this paper optimal outpatient appointment scheduling is studied. A local search procedure is derived that converges to the optimal schedule with a weighted average of expected waiting times of patients, idle time of the doctor and tardiness (lateness) as objective. No-shows are allowed to happen. For certain combinations of parameters the well-known Bailey-Welch rule is found to be the optimal appointment schedule.

Keywords: patient scheduling, health care, local search, multimodularity

1 Introduction

Outpatient appointment scheduling has been the subject of scientific investigation since the beginning of the fifties of the previous century when Bailey and Welch wrote [1]. The objective of appointment scheduling is trading off the interests of physicians and patients: the patients prefer to have a short waiting time, the physician likes to have as little idle time as possible, and to finish on time. Bailey & Welch [1] introduced the first advanced scheduling rule and tested it through simulation. Since then many papers have appeared that analyzed appointment scheduling in various settings (see Cayirli and Veral [2] for an overview). Most of them use simulation to analyze the performance of different appointment scheduling rules. A new method is introduced to determine optimal scheduling rules for arbitrary numbers of patients. Service time durations are exponentially distributed and patients arrive on time. No-shows are allowed to happen. The setting is discrete time, i.e., there is a finite number of (equally spaced) potential arrival moments.

A local search method is described that, starting from an arbitrary appointment schedule, tries to find neighboring appointment schedules that are better. From Koole & Van der Sluis [7] it follows that when the objective has a certain property related to convexity (called *multimodularity*) then a locally optimal schedule is globally optimal. The main technical result of this paper is the proof that our objective is indeed multimodular. This objective is a weighted sum of the average expected patient waiting time, the idleness of the doctor during the session length, and the tardiness. The tardiness is the probability that the session exceeds the planned finishing time multiplied by the average excess.

The local search method is also implemented and available for public use on the world wide web at obp.math.vu.nl/healthcare/software/ges. For big instances (many intervals) the computation times can be quite long. A faster local search method with a smaller neighborhood is also implemented. It is not guaranteed that it terminates with a global optimum solution, but it gives very good results, also for big instances, within a reasonable amount of time.

We give a short literature overview. The seminal paper on outpatient scheduling is Bailey & Welch [1]. For an overview of the results obtained since then, see Cayirli and Veral [2]. Roughly speaking we can classify the papers as follows: there are those that evaluate schedules (often using simulation) and those that design algorithms to find good schedules. A recent example of the former, not included in [2], is Hutzschenreuter [6]. In addition to no-shows she considers patients not arriving on time, and non-exponential service times. Those papers that present algorithms to design schedules can also be divided in two: those that focus on continuous time, which deal with finding the optimal interarrival intervals, and those that focus on discrete time, where the question is how many arrivals should be scheduled at each potential arrival moment. Pegden & Rosenshine [10] consider a continuous-time model. Their algorithm finds the optimal arrival moments, assuming convexity of the objective in the interarrival times. Also Lau & Lau [8] give a procedure for finding optimal arrival instants, again assuming convexity. Hassin & Mendel [5] extend this work to no-shows. Wang [16, 17] proves optimality, for phase-type service-time distributions, but for a limited number of patients. Denton and Gupta [3] formulate the problem as a two-stage stochastic linear program. Their algorithm is a good approach for quickly approximating large-scale systems. Also Robinson and Chen [12] consider a stochastic linear program. They derive a fast heuristic for finding good and robust interarrival times, using the fact that interarrival times are *dome-shaped*, meaning that they are shorter at the beginning and near the end of the session, and longer in the middle.

Let us now consider papers that are most relevant to the current work as they are dealing with discrete time, i.e., a finite number of potential arrival moments. In Liao et al. [9] a branch-and-bound method is used to find the optimal schedule. This works only for small instances. Vanden Bosche, Dietz & Simeoni [15] use a method that resembles the method of this paper in a number of ways. They derive upper and lower bounds for the optimal appointment schedule. To show these bounds they use what they call submodularity (Lemma 1 of [15]), which is in fact multimodularity on a subset of the equations that we use (see the appendix). Using the results of [15] upper and lower bound schedules (which often coincide) can be made starting from specific schedules. Our results give convergence to the optimal schedule starting from any schedule. The results of [15] are extended to different types of patients in Vanden Bosche & Dietz [13], and also to no-shows in Vanden Bosche & Dietz [14]. The inclusion of different types of patients is relatively straightforward, the sequence is optimized using local search, and for each sequence the optimal schedule is determined using the method of [15]. Also our proofs nowhere use the fact that service times are equally distributed. Summarizing, compared to the work of Vanden Bosche and co-authors, our stronger sub/multimodularity results allow us to formulate an algorithm that converges from any initial schedule to the optimal one.

The paper is structured as follows. In Section 2 a model is defined, in which we can compute for an arbitrary appointment schedule the objective. In Section 3 the local search algorithm is described. Section 4 is devoted to numerical results. The proof that our objective is multimodular is given is Appendix A.

2 Model

For the scheduling problem we have to introduce some variables. A treatment/operation room is operational during T intervals with length d (for example a day from 8.00AM till 4.00PM split in intervals with length 10 minutes gives $T = 48$ and $d = 10$). Within these T intervals a total of N patients should be scheduled. Patient service times are assumed to be exponentially distributed with rate μ (and expectation μ^{-1}).

Let $x_t \in \{0,\ldots,N\}$ be the number of patients scheduled at the start of interval t. A schedule is a vector $(x_1,\ldots,x_T)$ with $\sum_{t=1}^{T} x_t = N$. So we have:

- $\beta = \frac{1}{\mu}$: average service time
- T: number of intervals
- d: length of interval
- N: total number of patients
- x_t: number of patients scheduled at the start of interval t, $t = 1,\ldots,T$

In the model we make the following assumptions:

- The service times of patients are independent and exponentially distributed.
- Patients always come on time (no-shows are modeled later on).

In the following sections we will give the formulas for calculating for a given schedule the mean waiting time, idle time and tardiness (lateness), which we call $W(x)$, $I(x)$ and $L(x)$, respectively. To compare schedules we give weights α_W, α_I, and α_L to the three main factors to obtain the overall objective function $C(x) = \alpha_W W(x) + \alpha_I I(x) + \alpha_L L(x)$. Our problem can now be stated as follows:

$$\min \left\{ \alpha_W W(x) + \alpha_I I(x) + \alpha_L L(x) \middle| \begin{array}{c} \sum_t x_t = N \\ x_t \in \mathbb{N}_0 \end{array} \right\} \tag{1}$$

For a given schedule ($x = (x_1,\ldots,x_T)$) the probabilities of having i patients in the queue just before new arrival(s) and just after arrival(s) can be calculated. This can be used to calculate the mean waiting time, idle time and tardiness. We introduce the following notation:
$p_{t^-}(i) = \mathbb{P}(i$ patients in queue just before the arrival(s) at interval $t)$ and
$p_{t^+}(i) = \mathbb{P}(i$ patients in queue just after the arrival(s) at interval $t)$.

We start empty, thus $p_{1^-}(0) = 1$. Iteratively the other probabilities can be calculated as follows:

$$\begin{array}{rcll}
p_{1^-}(0) & \equiv & 1, & \\
p_{t^+}(j) & = & 0, & 0 \leq j < x_t, \\
p_{t^+}(j) & = & p_{t^-}(j - x_t), & j \geq x_t, \\
p_{(t+1)^-}(0) & = & \sum_{i=0}^{N} p_{t^+}(i) b_i, & \\
p_{(t+1)^-}(j) & = & \sum_{i=j}^{N} p_{t^+}(i) a_{i-j}, & j \geq 0.
\end{array}$$

where

$$a_i = \mathbb{P}(\#\text{ potential departures} = i) = \frac{(\mu d)^i}{i!} e^{-\mu d}$$

and

$$b_i = \mathbb{P}(\#\text{ potential departures} \geq i) = 1 - \sum_{j=0}^{i-1} a_i.$$

Because of the exponential service time distributions the potential number of departures in any interval has a Poisson distribution.

2.1 Mean waiting time of a patient

If a patient arrives and finds k patients in the queue (including the patient who is currently being treated), then the mean waiting time of that patient will be k/μ. In our model patients arrive just before a new interval alone or in groups. The ith one of that group has a mean waiting time of $\sum_{j=0}^{N} p_{t^-}(j) \cdot (j+i-1)\frac{1}{\mu}$. This is just the mean waiting time of one patient, so we must sum them all over the groups and intervals an divide that through all N patients. Thus we find the following formula for the mean waiting time:

$$W(x) = \frac{1}{N} \sum_{t=1}^{T} \sum_{i=1}^{x_t} \sum_{j=0}^{N} p_{t^-}(j) \cdot (j+i-1)\frac{1}{\mu} \tag{2}$$

2.2 Mean idle time of a doctor

For calculating the mean idle time of a doctor, we calculate first the mean makespan $M(x)$, which is the time the last patient finishes. Then it is easy to find the mean idle time $I(x)$, because $I(x) = M(x) - N/\mu$.

Set $\tilde{t} \equiv \{max\, t | x_t > 0\}$. Now we know for sure that the makespan is greater than $(\tilde{t}-1)d$. The distribution of the number of patients in the queue at time $\tilde{t}$ is known. So the average makespan is

$$M(x) = (\tilde{t}-1)d + \sum_{j=1}^{N} p_{\tilde{t}^+}(j) \cdot \frac{j}{\mu}.$$

So now we obtain the following formula for the mean idle time:

$$I(x) = \left((\tilde{t}-1)d + \sum_{j=1}^{N} p_{\tilde{t}^+}(j) \cdot \frac{j}{\mu}\right) - \frac{N}{\mu} \tag{3}$$

2.3 Mean tardiness

For the mean tardiness of the day we look at the end of the last interval. Now if there are j patients in queue, then the extension is on average $j\frac{1}{\mu}$. We know the patient distribution just after the last interval T, so the tardiness function is as follows:

$$L(x) = \sum_{j=1}^{N} p_{(T+1)^-}(j)\frac{j}{\mu} \tag{4}$$

2.4 Including No-shows

We can add no-shows to our model. This is an important generalization as no-shows occur frequently in practice. Every patient now has a probability ρ of not showing up. We assume that ρ is the same for all patients and that the patients are independent. Thus the number of arrivals at time t has a Binomial(x_t, ρ) distribution.

This changes the formulas used in the model as follows. $p_{t^-}(i)$ remains the same. $p_{t^+}(i)$ is somewhat different, because it is not known how many patients are exactly coming. We must sum over the distribution of how many patients will be arrive. This gives for $p_{t^-}(i)$ and $p_{t^+}(i)$:

$$\begin{aligned}
p_{1^-}(0) &\equiv 1, \\
p_{t^+}(j) &= \textstyle\sum_{k=0}^{x_t} \binom{x_t}{k} \rho^{x_t - k}(1-\rho)^k \cdot p_{t^-}(j-k), & j \geq 0, \\
p_{(t+1)^-}(0) &= \textstyle\sum_{i=0}^{N} p_{t^+}(i) b_i, \\
p_{(t+1)^-}(j) &= \textstyle\sum_{i=j}^{N} p_{t^+}(i) a_{i-j}, & j \geq 0.
\end{aligned}$$

2.4.1 Mean waiting time of a patient

For the mean waiting time, for all intervals we must additionally sum the waiting time over the distribution of the number of arriving patients. This gives the following equation:

$$W(x) = \frac{1}{N(1-\rho)} \sum_{t=1}^{T} \sum_{k=1}^{x_t} \binom{x_t}{k} \rho^{x_t - k}(1-\rho)^k \Big(\sum_{i=1}^{k} \sum_{j=0}^{N} p_{t^-}(j) \cdot \frac{j+i-1}{\mu} \Big) \tag{5}$$

2.4.2 Mean idle time of a doctor

Again we first calculate the makespan. Now we do not know when the last patient is coming. But we can calculate the probability that the last patient is coming at interval t. This probability is

$$\begin{aligned}
& \mathbb{P}(\text{last patient is coming at interval } t) \\
= \; & \mathbb{P}(\text{all patients after interval } t \text{ are no-shows})\mathbb{P}(\#\text{ arrivals at time } t \geq 1) \\
= \; & \rho^{N - \sum_{i=1}^{t} x_i}(1-\rho^{x_t}).
\end{aligned}$$

If the last patient is coming at interval t, we know for sure the makespan is greater than $(t-1)d$. To calculate the excess after interval t, we sum over the distribution of the number of patients that come (having in mind that at least one patient comes) times the mean excess.

What we find is then:

$$\begin{aligned} M(x) &= \sum_{t:x_t>0} \mathbb{P}(\text{last patient is coming at time } t)\mathbb{E}(\text{mean makespan}|\text{last patient is coming at time } t) \\ &= \sum_{t:x_t>0} \rho^{N-\sum_{i=1}^{t} x_i}(1-\rho^{x_t})\left((t-1)d+\sum_{k=1}^{x_t}\frac{\binom{x_t}{k}\rho^{x_t-k}(1-\rho)^k}{1-\rho^{x_t}}\sum_{i=0}^{N} p_{t^-}(i)\cdot\frac{i+k}{\mu}\right) \end{aligned}$$

The mean idle time is then given by:

$$I(x) = M(x) - N(1-\rho)\frac{1}{\mu} \tag{6}$$

The question can be asked how important this mean idle time is, because now the time between the real last patient and the last planned patient is not added as idle time. So in the case of no-shows the idle time is less relevant as objective and should have a relatively low weight.

2.4.3 Mean tardiness

The formula of the mean tardiness is the same (of course with the new probabilities $p_{t^-}(i)$ and $p_{t^+}(i)$).

3 Local search

To compute the schedule with the lowest objective value we could try all possible schedules (the solution space) and look which one has the lowest objective value. But the number of all possible schedules is huge (it is $\binom{N+T-1}{N}$), so we need a search algorithm to reduce the computation time. A local search algorithm starts with a feasible solution and tries iteratively to improve the current solution by searching a better solution in its neighborhood until a local minimum is found.

In general the local minimum is not a global minimum, but for the current problem and a well-chosen neighborhood it is possible to show that the local search algorithm finishes in the global minimum.

We introduce our neighborhood. Define the vectors

$$\left\{\begin{array}{c} u_1, \\ u_2, \\ u_3, \\ \vdots \\ u_{T-1}, \\ u_T \end{array}\right\} = \left\{\begin{array}{c} (-1,0,\ldots,0,1), \\ (1,-1,0,\ldots,0), \\ (0,1,-1,0,\ldots,0), \\ \vdots \\ (0,\ldots,1,-1,0), \\ (0,\ldots,0,1,-1) \end{array}\right\},$$

and take $\mathcal{V}^* = \{u_1,\ldots,u_T\}$. As the neighborhood of schedule x we take all vectors of the form $x+v_1+\cdots+v_k$ with $v_1,\ldots,v_k \in \mathcal{V}^*$ such that $x+v_1+\cdots+v_k \geq 0$. Then the algorithm is as follows.

Algorithm for computing an optimal schedule

```
1. Start with some schedule $x$
2. For all $\mathcal{U} \subsetneq \mathcal{V}^*$:
    for $y = x + \sum_{v \in \mathcal{U}} v$ such that $y \geq 0$ compute $C(y)$;
    if $C(y) < C(x)$ then $x := y$ and start again with step 2
3. $x$ is the optimal schedule
```

A vector u_t can be interpreted as moving a patient from time slot t to time slot $t-1$. Thus the neighborhood of x consists of all combination of single-interval shifts starting from x. In Appendix A we prove that with this neighborhood the local search algorithm converges to the global optimal solution.

In the online tool we also implemented a smaller neighborhood that gives much faster results. Under this option we simply take $y = x + u$ for all $u \in \mathcal{U}$ in step 2 of the algorithm, thus we only consider $\mathcal{U}$ with $|\mathcal{U}| = 1$.

4 Numerical examples

In this section we give some numerical examples. All these computations were done with our webtool which is available for experimentation at obp.math.vu.nl/healthcare/software/ges.

Let the following be the base-case scenario. A medical practice is operational between 8.00AM and 12.00AM. We split this interval up in 48 intervals of 5 minutes. Thus $T = 48$ and $d = 5$. A treatment duration is on average 20 minutes ($1/\mu = 20$) and the percentage of no-shows is 10% ($\rho = 0.10$). We want to plan 10 patients ($N = 10$).

To analyze this model with the small neighborhood (which is not guaranteed to give the optimal solution) took a few seconds, analyzing the full neighborhood (what we did for all cases considered in this section) took around 12 hours for each instance.

First we compute for base-case scenario the the optimal schedule, for different weights in our objective function. The weight for the tardiness is taken 1 ($\alpha_L = 1$), for the idle time it is taken 0.2 ($\alpha_I = 0.2$). The idle time has a relatively low weight because of the no-shows. We took four different weights for α_W (0.5, 1, 2, and 10), and determined the optimal schedules for each of these cases. The schedules are given in Figure 1.

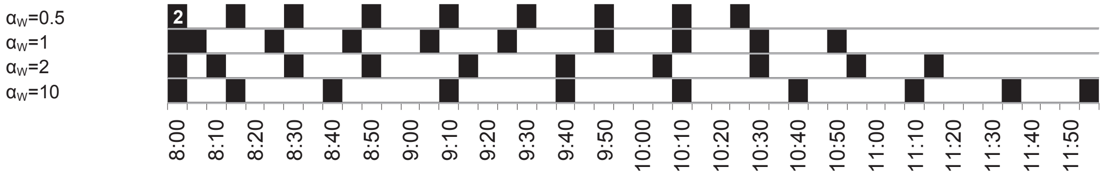

Figure 1: Base-case scenario with different weights

It is seen that if the waiting time has given a bigger weight then the patients are more spread out to the end of the schedule, as one would expect. In the optimal schedule with $\alpha_W = 0.5$ there are two patients scheduled at the beginning of the day. Note that the optimal schedule for $\alpha_W = 0.5$ is close to the Bailey-Welch rule. In all cases the times between consecutive arrivals first increases and then decreases again. This is the dome-shaped form that we discussed in the literature overview.

To have a better look on the results we compare the optimal schedules with two existing schedules: the individual block schedule and the Bailey-Welch rule. With the individual block schedule the working day is divided in the same number of intervals as there are patients. In each block exactly one patient is scheduled. The Bailey-Welch rule is similar as the individual block schedule, but with the last patient moved to the beginning of the day. So in our base-case scenario the individual block schedule and the Bailey-Welch rule plan every 24 minutes a patient, with the exception that the Bailey-Welch rule schedules two patients at 8.00AM and none at 11.36AM.

	$\alpha_W = 0.5$	$\alpha_W = 1$	$\alpha_W = 2$	$\alpha_W = 10$	Individual	Bailey-Welch
Mean Waiting Time	26.46	19.90	15.35	9.85	12.37	16.75
Mean Idle Time	21.86	36.69	54.02	88.58	72.14	50.07
Mean Tardiness	7.99	9.60	12.61	29.79	19.62	11.42
Object value ($\alpha_W = 0.5$)	25.59				40.23	29.81
Object value ($\alpha_W = 1$)		36.83			46.41	38.18
Object value ($\alpha_W = 2$)			54.12		58.78	54.94
Object value ($\alpha_W = 10$)				146.00	157.72	188.95

Table 1: Outcome values for different schedules

The results of the schedules are given in Table 1. The optimal schedules are of course better than the two existing schedules, but it can been seen for $\alpha_W = 2$ that the Bailey-Welch schedule is almost as good as the optimal one.

Now we will look what happens with the optimal schedules if we change some parameters. The changes are chosen such that the total workload does not change. The workload for the base-case scenario is $N\beta(1-\rho) = 10 * 20 * 0.9 = 180$ minutes. We change the parameters two at a time, ρ and β, N and β, and N and ρ, respectively. Let $\alpha_W = 2$ and the other parameters fixed as in the base-case scenario. The optimal schedules are given in Figure 2. The corresponding outcome values are shown in Table 2.

From Table 2a we see that if ρ becomes larger (thus β decreases) the mean waiting time, idle time and tardiness all becomes larger, because of the higher uncertainty. From the results of Table 2b it is seen that if β becomes smaller (thus N increases) then the mean waiting time, idle time and tardiness all becomes smaller because of reduced uncertainty. The results of Table 2c show us that if ρ becomes larger (thus N decreases) the mean waiting time, idle time and tardiness all becomes larger because of the higher uncertainty.

A final change in parameters would be changing T and d. This would evidently lead to more simultaneous arrivals.

5 Conclusions

In this paper a method is presented to obtain optimal outpatient schedules in case of a finite number of possible arrival epochs. The proof of the optimality relies on showing that the objective is multimodular, which is a generalization of convexity to lattices.

Numerical results are presented. The interarrival times have a dome shape, as observed earlier in the literature: the first interarrival times are short, then they get longer, and become again short.

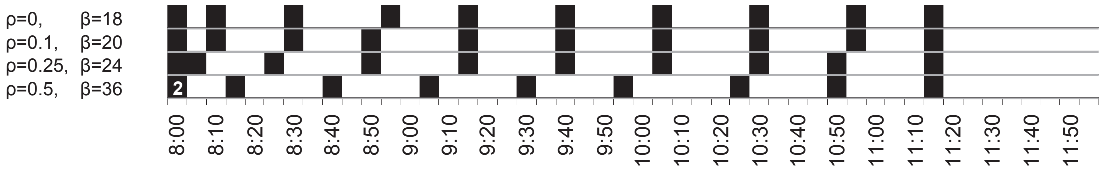

Figure 2a: Optimal schedules (ρ against β)

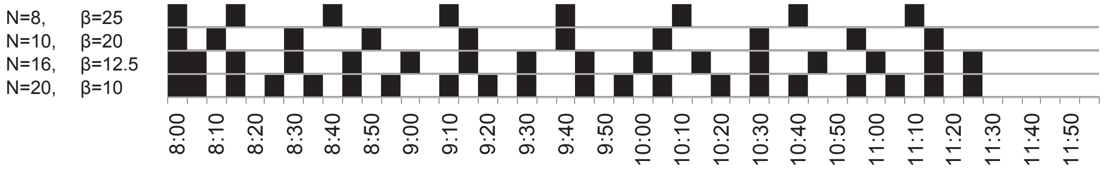

Figure 2b: Optimal schedules (*N* against β)

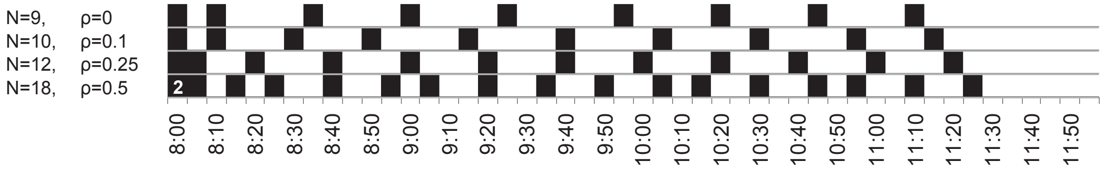

Figure 2c: Optimal schedules (*N* against ρ)

Note that in certain cases the optimal rule is close to the Bailey-Welch rule. For certain parameter values the Bailey-Welch rule is indeed optimal.

Acknowledgment The authors would like to thank the three anonymous referees for their valuable suggestions.

References

[1] N. T. J. Bailey and J. D. Welch Appointment systems in hospital outpatient departments. *The Lancet* **259**, 1105–1108, 1952

[2] T. Cayirli, E. Veral Outpatient scheduling in health care: a review of literature. *Production and Operations Management* **12**, 519–549, 2003

[3] B. Denton and D. Gupta A sequential bounding approach for optimal appointment scheduling. *IIE Transactions* **35**, 1003–1016, 2003

[4] B. Hajek Extremal splitting of point processes. *Mathematics of Operations Research* **22**, 543–556, 1985

[5] R. Hassin and S. Mendel Scheduling arrivals to queues: a model with no-shows. Working paper, 2006

	$\rho = 0, \beta = 18$	$\rho = 0.1, \beta = 20$	$\rho = 0.25, \beta = 24$	$\rho = 0.5, \beta = 36$
Mean Waiting Time	13.43	15.35	18.93	27.29
Mean Idle Time	51.67	54.02	56.96	60.66
Mean Tardiness	10.04	12.61	17.28	28.59
Object value	47.24	54.12	66.53	95.29

Table 2a: Outcome values (ρ against β)

	$N = 8,\beta = 25$	$N = 10,\beta = 20$	$N = 16,\beta = 12.5$	$N = 20,\beta = 10$
Mean Waiting Time	16.74	15.35	11.83	11.09
Mean Idle Time	54.82	54.02	53.53	49.30
Mean Tardiness	15.56	12.61	8.10	5.60
Object value	60.00	54.12	42.47	37.63

Table 2b: Outcome values (N against β)

	$N = 9, \rho = 0$	$N = 10, \rho = 0.1$	$N = 12, \rho = 0.25$	$N = 18, \rho = 0.5$
Mean Waiting Time	14.44	15.35	17.48	21.73
Mean Idle Time	50.12	54.02	56.43	58.07
Mean Tardiness	10.83	12.61	14.63	17.35
Object value	49.73	54.12	60.89	72.43

Table 2c: Outcome values (N against ρ)

[6] A. HUTZSCHENREUTER Queueing models for outpatient appointment scheduling. M.Sc thesis, University of Ulm, 2005

[7] G. KOOLE AND E. VAN DER SLUIS Optimal shift scheduling with a global service level constraint. *IIE Transactions* **35**, 1049–1055, 2003

[8] H. LAU AND A.H. LAU A Fast Procedure for Computing the Total System Cost of an Appointment Schedule for Medical and Kindred Facilities. *IIE Transactions* **32**, 833–839, 2000

[9] C. LIAO, C.D. PEGDEN AND M. ROSENSHINE Planning Timely Arrivals to a Stochastic Production or Service System *IIE Transactions* **25**, 36–73, 1993

[10] C.D. PEGDEN AND M. ROSENSHINE Scheduling Arrivals to Queues *Computers & Operations Research* **17**, 343–348, 1990

[11] R. RIGHTER Scheduling. Stochastic Orders and their Applications. Eds. M. Shaked and J. G. Shanthikumar *Academic Press*, 1994

[12] L.W. ROBINSON AND R.R. CHEN Scheduling doctors' appointments: optimal and empirically-based heuristic policies *IIE Transactions* **35**, 295–307, 2003

[13] P.M. VANDEN BOSCHE AND D.C. DIETZ Minimizing expected waiting in a medical appointment system *IIE Transactions* **32**, 841–848, 2000

[14] P.M. VANDEN BOSCHE AND D.C. DIETZ Scheduling and sequencing arrivals to an appointment system *Journal of Service Research* **4**, 15–25, 2001

[15] P.M. VANDEN BOSCHE, D.C. DIETZ AND J.R. SIMEONI Scheduling Customer Arrivals to a Stochastic Service System *Naval Research Logistics* **46**, 549–559, 1999

[16] P.P. WANG Static and Dynamic Scheduling of Customer Arrivals to a Single-Server System *Naval Research Logistics* **40**, 345–360, 1993

[17] P.P. WANG Optimally Scheduling N Customer Arrival Times for a Single-Server System *Computers & Operations Research* **24**, 703–716, 1997

A Local search method

To prove that the local search algorithm converges to the global optimum, we first show that our objective function is multimodular. We start by defining multimodularity.

A.1 Multimodularity

Multimodularity (Hajek [4]) is a property of functions on $\mathbb{Z}^m$. Define the vectors $v_0, \ldots, v_m \in \mathbb{Z}^m$ as follows:

$$\begin{aligned} v_0 &= (-1, 0, \ldots, 0) \\ v_1 &= (1, -1, 0, \ldots, 0) \\ v_2 &= (0, 1, -1, 0, \ldots, 0) \\ &\vdots \\ v_{m-1} &= (0, \ldots 0, 1, -1) \\ v_m &= (0, \ldots, 0, 1) \end{aligned}$$

Let $\mathcal{V} = \{v_0, \ldots, v_m\}$. Now:

Definition A.1 *A function $f : \mathbb{Z}^m \to \mathbb{R}$ is called multimodular if for all $x \in \mathbb{Z}^m, v, w \in \mathcal{V}, v \neq w$,*

$$f(x+v) + f(x+w) \geq f(x) + f(x+v+w) \tag{7}$$

Central in the theory of multimodular functions is the concept of an atom.

Definition A.2 *For some $x \in \mathbb{Z}^m$ and σ a permutation of $\{0, \ldots, m\}$, we define the atom $S(x, \sigma)$ as the convex set with extreme points $x + v_{\sigma(0)}, x + v_{\sigma(0)} + v_{\sigma(1)}, \ldots, x + v_{\sigma(0)} + \cdots + v_{\sigma(m)}$.*

It is shown in Hajek [4] that each atom is a simplex, and each unit cube is partitioned into $m!$ atoms; all atoms together span $\mathbb{R}^m$.

In Koole and Van der Sluis [7] the following theorem is shown. It forms the basis of our neighborhood choice.

Theorem A.3 *For f multimodular, a point $x \in \mathbb{Z}^m$ is a global minimum if and only if $f(x) \leq f(y)$ for all $y \neq x$ such that $y \in \mathbb{Z}^m$ is an extreme point of $S(x, \sigma)$ for some σ.*

Our problem (1) is a $T-1$ dimensional problem: given $x_1,\dots,x_{T-1}$ we derive x_T by $x_T = N-\sum_{t=1}^{T-1} x_t$. We will show that it has a multimodular objective function. The set of allowable solutions is given by $\{x\in\mathbb{Z}^{T-1}|x\geq 0,\sum_{t=1}^{T-1} x_t \leq N\}$. This domain is not equal to $\mathbb{Z}^{T-1}$, so the question arises if the local search algorithm still converges to the global minimum. According Lemma 2 in Koole and Van der Sluis [7] Theorem A.3 remains valid for this subset of $\mathbb{Z}^{T-1}$. Proving that our objective function is multimodular for the $T-1$-dimensional problem (1) is equivalent to showing that the objective function in T dimensions satisfies Equation (7) for $v,w\in\mathcal{V}^*$, where

$$\mathcal{V}^* = \left\{\begin{array}{c} u_1, \\ u_2, \\ u_3, \\ \vdots \\ u_{T-1}, \\ u_T \end{array}\right\} = \left\{\begin{array}{c} (-1,0,\dots,0,1), \\ (1,-1,0,\dots,0), \\ (0,1,-1,0,\dots,0), \\ \vdots \\ (0,\dots,1,-1,0), \\ (0,\dots,0,1,-1) \end{array}\right\}.$$

Note that u_t is nothing else then moving a patient from time slot t to time slot $t-1$. Now we show that our objective function is multimodular and that it can be minimized by a local search algorithm that is guaranteed to terminate in the global minimum. Our neighborhood is the set of all possible combinations of the vectors u_t added to the current schedule.

Theorem A.4 *The waiting time function $W(x)$, the idle time function $I(x)$ and the tardiness function $L(x)$, as defined in Equations (2)-(4), are multimodular for all $u_i,u_j\in\mathcal{V}^*$ for which $i\neq j$.*

Proof of theorem A.4 It is easy to see that if the makespan is multimodular then also the idle time is multimodular. Thus it is sufficient to show that the makespan, the waiting time and the tardiness are multimodular. Thus it has to be shown that

$$\begin{aligned} W(x+u_i)+W(x+u_j) &\geq W(x)+W(x+u_i+u_j)\ , \\ M(x+u_i)+M(x+u_j) &\geq M(x)+M(x+u_i+u_j) \text{ and} \\ T(x+u_i)+T(x+u_j) &\geq L(x)+T(x+u_i+u_j) \end{aligned}$$

for every possible i and j with $1\leq i<j\leq T$. We use coupling (see Righter [11]) for this proof, to compare the different schedules x, $x+u_i$, $x+u_j$ and $x+u_i+u_j$. For every possible combination of i and j, all different possibilities of patient flows are distinguished to detect the difference between the number of patients in queue for each schedule for each time interval. First the proof is given for $2\leq i<j\leq T$.

In Figure 3, different paths are shown for the different schedules.

(A) Let us start with Case A. Schedule (A1) and schedule (A3) are following the same path until time $j-1$. Also Schedule (A2) and schedule (A4) are following the same path until that time. In Case A the queue empties between time i and time $j-1$, so from that time on all the paths are the same. Thus just before time $j-1$, there are say k patients in queue.

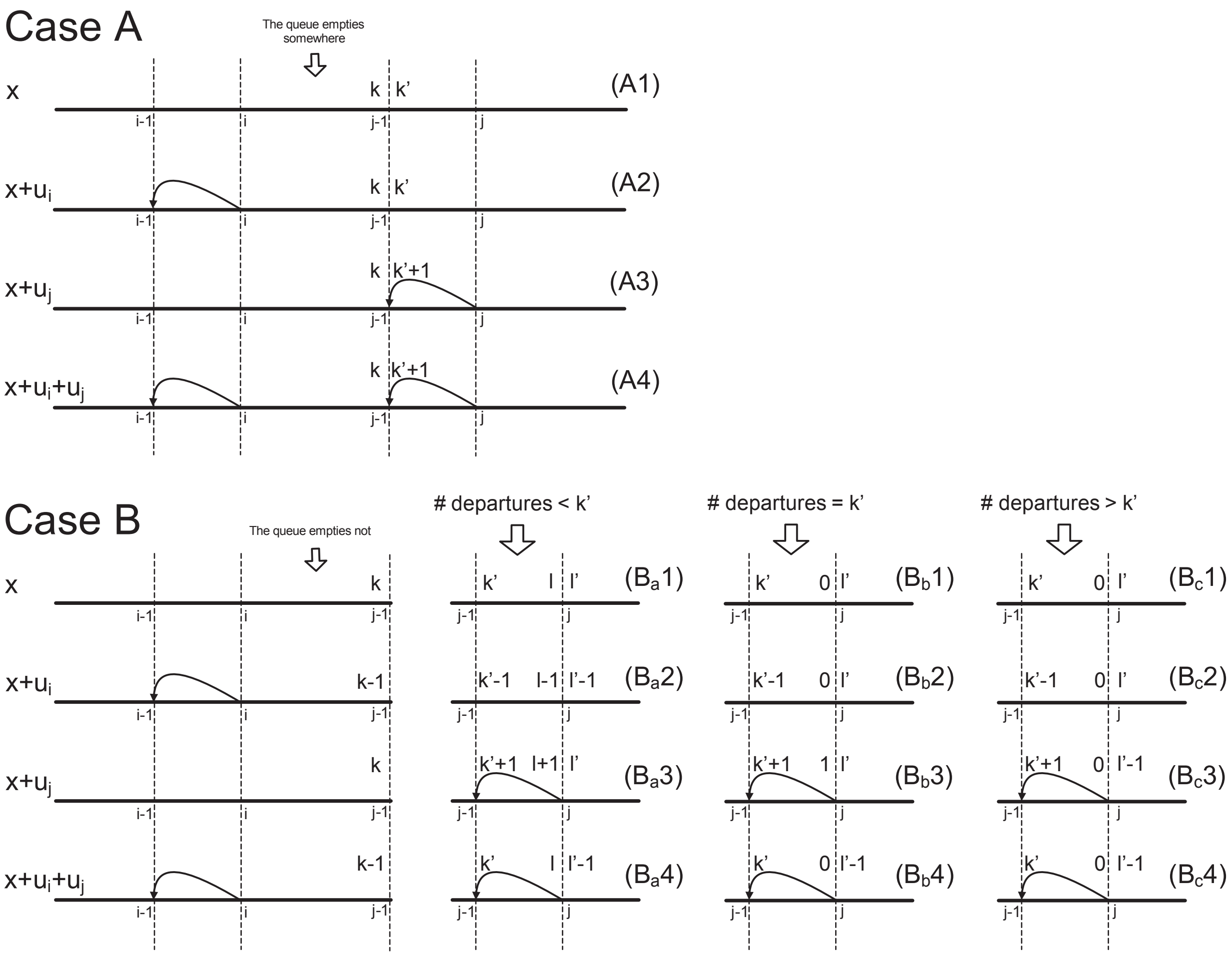

Figure 3: Case A & B ($2 \leq i < j \leq T$)

Let $k' = k + x_{j-1}$. Then just after time $j-1$ there are k' patients in queue for schedules (A1) and (A2) and $k'+1$ for schedules (A3) and (A4). Thus after time $j-1$ schedule (A1) and schedule (A2) are following the same path and also schedule (A3) and schedule (A4) are following the same path.

Now say that until time $j-1$ schedule (A1) has a total waiting time α_1, then schedule (A3) also has that total waiting time α_1. Say that until time $j-1$ schedule (A2) has a total waiting time α_2, then schedule (A4) has the same total waiting time. Just after time $j-1$ schedules (A1) and (A2) follow the same path, so they have the same total waiting time, say β_1. Schedule (A3) and (A4) also follow the same path, thus they also have the same total waiting time, say β_2. Now it is easy to see that the waiting time satisfies $\alpha_2 + \beta_1$(A2)+$\alpha_1 + \beta_2$(A3)=$\alpha_1 + \beta_1$(A1)+$\alpha_2 + \beta_2$(A4).

For the makespan and tardiness only the end of a day is important, so we want to know what happens at the end of the path of each schedule. Schedules (A1) and (A2) follow after time $j-1$ the same path, and therefore they have the same makespan and tardiness. Schedules (A3) and (A4) follow after time $j-1$ the same path, therefore they also have the same makespan and tardiness. So (A2)+(A3)=(A1)+(A4) for the makespan and the tardiness.

(B) Now look at "Case B". The queue does not empty between time i and $j-1$, so now just before time $j-1$ it can be that for schedules (B2) and (B4) there is one patient less in queue, because one patient more could be treated (because the movement of one patient from time i to time $i-1$). Otherwise all different schedules will have the same number in queue and then "Case A" applies. So for schedule (B2) and (B4) there are then $k-1$ patients in queue and for schedules (B1) and (B3) there are k patients in queue. Concerning the waiting time, let us say again that schedules (B1) and (B3) have a total waiting time of α_1 and that schedules (B2) and (B4) have a total waiting time of α_2.

Define again $k' = k + x_{j-1}$. Then just after time $j-1$ there are k' patients in queue for schedule (B1), $k'-1$ for schedule (B2), $k'+1$ for path (B3), one more because of the movement of one patient from time j to time $j-1$ and k' for schedule (B4).

Now we distinguish between the following three possibilities for the number of departures between time $j-1$ an j. Let $l' = l + x_j$.

a) The number of departures is less than k'.

- For schedule (B_a1) there will be say $l(\geq 1)$ patients left just before time j and just after time j it will be then l'. Let the total waiting time between time $j-1$ and j be β and after time j γ_1.
- For schedule (B_a2) the number of patient is $l-1$ just before time j. So just after time j there are $l'-1$ patients in queue and the total waiting time between time $j-1$ and j is then $\beta - d$ and after time j γ_2.
- For schedule (B_a3) the number of patient is $l+1$, just before time j. Just after time j there are l' patients in queue (one patient less arrives) and the total waiting time between time $j-1$ and j is then $\beta + d$ and after time j again γ_1.
- For schedule (B_a4) the number of patient is l, just before time j. Just after time j there are $l'-1$ patients in queue (one patient comes less) and the total waiting time between time $j-1$ and j is again β and after time j again γ_2.

Now we see that the waiting time satisfies $\alpha_2+\beta-d+\gamma_2(B_a2)+\alpha_1+\beta+d+\gamma_1(B_a3)=\alpha_1+\beta+\gamma_1(B_a1)+\alpha_2+\beta+\gamma_2(B_a4)$

The end of the path (after time j) of schedules (B_a1) and (B_a3) is the same. The same holds for (B_a2) and (B_a4). So in this case (B_a2)+(B_a3)=(B_a1)+(B_a4), for the makespan and tardiness.

b) The second possibility is that there are exactly k' departures between time $j-1$ and j.

- For schedule (B_b1) there will be $k'-k'=0$ patients left just before time j and just after time j it will be l'. Let the total waiting time between time $j-1$ and j β and after time j γ_1.

- For schedule (B_b2) the number of patients is also 0, just before time j. So just after time j there are l' patients in queue and the total waiting time between time $j-1$ and j is then $\beta - d$ and after time j again γ_1.
- For schedule (B_b3) the number of patients is $k'+1-k'=1$, just before time j. So just after time j there are l' patients in queue and the total waiting time between time $j-1$ and j is then $\beta + d$ and after time j again γ_1.
- For schedule (B_b4) the number of patients is $k'-k'=0$, just before time j. So just after time j there are $l'-1$ patients in queue and the total waiting time between time $j-1$ and j is then β (same as (B_b1)) and after time j γ_2 which is of course smaller then γ_1.

Now we see that the waiting time satisfies $\alpha_2+\beta-d+\gamma_1(B_a2)+\alpha_1+\beta+d+\gamma_1(B_a3)\geq \alpha_1+\beta+\gamma_1(B_a1)+\alpha_2+\beta+\gamma_2(B_a4)$.

The end of the path (after time j) of schedules (B_b1), (B_b2) and (B_b3) are the same so the makespan and tardiness are the same for these schedules. At the end of the path of schedule (B_b4) there is one patient less (or in the worst case the same), so the makespan and tardiness is also less or equal than the other schedules. So we can conclude that (B_b2)+(B_b3)$\geq$(B_b1)+(B_b4), for the makespan and tardiness.

c) The last possibility is that there are more than k' departures between time $j-1$ and j. So for all paths ((B_c1), (B_c2), (B_c3) and (B_c4)) there will be no patients left just before time j.
Just after time j there will be for schedule (B_c1) and (B_c2) l' patients in queue and have a total waiting time of γ_1. (B_c3) and (B_c4) have then $l'-1$ patients in queue and a total waiting time of γ_2.

Now the total waiting time between time $j-1$ and time j, if there are $s>k$ departures is $\sum_{n=1}^{m}\frac{(n-1)d}{s}=\frac{m(m-1)d}{2s}$ (the first patient has a waiting time of 0, the second $\frac{d}{s}$, the third $\frac{2d}{s}$, etc...), with m the number of patients just after time $j-1$. Because this is a convex function it is clear that the waiting time function satisfies $\alpha_2+\frac{(k-1)(k-2)d}{2s}+\gamma_1(B_c2)+\alpha_1+\frac{(k+1)kd}{2s}+\gamma_2(B_c3)\geq \alpha_1+\frac{k(k-1)d}{2s}+\gamma_1(B_c1)+\alpha_2+\frac{k(k-1)d}{2s}+\gamma_2\ (B_c4)$.

The ends of the paths of schedule (B_c1) and (B_c2) are the same and the ends of paths of schedule (B_c3) and (B_c4) are the same. Therefore is (B_c2)+(B_c3)=(B_c1)+(B_c4), for the makespan and tardiness.

All cases for $2\leq i<j\leq T$ are done. Now for $1=i<j\leq T$. For "Case C" until "Case E" (Figure 4) counts that before time $j-1$ the queue somewhere empties, so after that time all schedules will following the same path and just before time $j-1$ there are k patients in queue for all schedules. until time $j-1$ schedule (1) and (3) have a total waiting time α_1 and schedule (2) and (4) a total waiting time α_2.

Just after time $j-1$ there will be k' patients for schedule (1) and (2) and $k'+1$ patients for schedule (3) and (4). Now after time $j-1$ we can distinguish the following four possibilities.

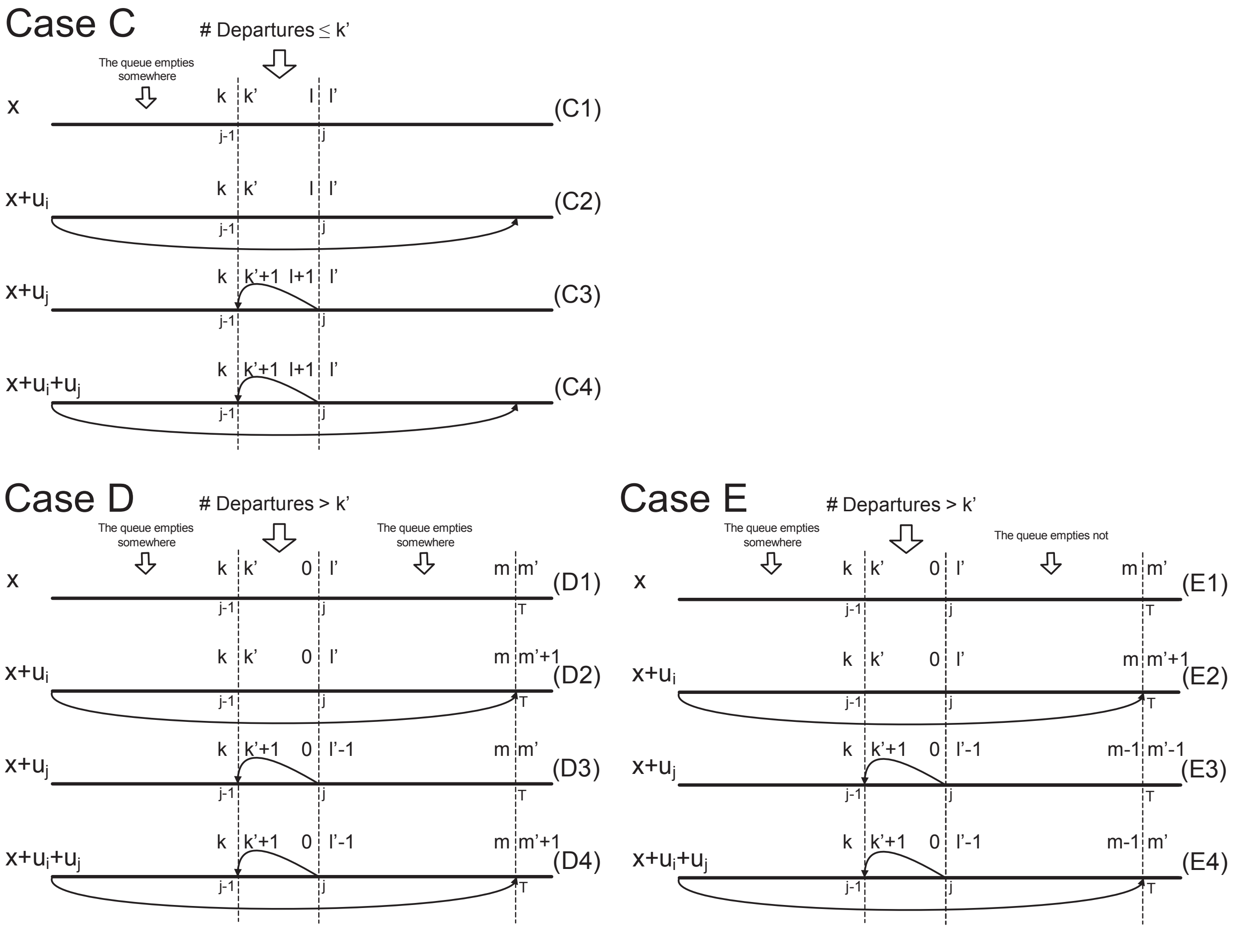

Figure 4: Case C, D & E ($1 = i < j \leq T$)

(C) Now for "Case C" there are equal or less than k' departures so just before time j there are for schedule (C1) and (C2) l patients left and for schedule (C3) and (C4) $l+1$ patients left. Between time $j-1$ and j schedule (C1) and (C2) have the same total waiting time, say β_1 and schedule (C3) and (C4) have the same total waiting time, say β_2.
Just after time j there are for all schedules l' patients. So after time j follows schedule (C1) and (C3) the same path and have a total waiting time of γ_1 and follows schedule (C2) and (C4) the same path and have a total waiting time of γ_2. Now is easy to see that the waiting time satisfies $\alpha_2 + \beta_1 + \gamma_2$(C2)+$\alpha_1 + \beta_2 + \gamma_1$(C3)=$\alpha_1 + \beta_1 + \gamma_1$(C1)+$\alpha_2 + \beta_2 + \gamma_2$(C4).

The ends of paths of schedule (C1) and (C3) are the same and the ends of paths of schedule (C2) and (C4) are the same. Therefore is (C2)+(C3)=(C1)+(C4), for the makespan and tardiness.

(D,E) Now for "Case D" and "Case E" there are more than k' departures between time $j-1$ and time j. So just before time j there are no patients left for all schedules. Between time $j-1$ and j schedule (1) and (2) have the same total waiting time, say β_1 and schedule (3) and (4) have the same total waiting time, say β_2.

Just after time j there are for schedule (1) and (2) l' patients in queue and for schedule (3) and (4) $l'-1$. Between time j and time T schedule (1) and (2) have a total waiting time of say γ_1 and schedule (3) and (4) have a total waiting time of say γ_2. Between time j and T can happen the following two cases:

(D) The queue empties. So for all schedules there are say m patients left just before time T ("Case D"). Let $m' = m + x_T$. Just after time T there will be then m' patients for schedule (D1) and (D3), with a total waiting time of δ_1 and $m'+1$ patients for schedule (D2) and (D4), with a total waiting time of δ_2. So the waiting time function satisfies $\alpha_2+\beta_1+\gamma_2+\delta_2$(D2)+$\alpha_1+\beta_2+\gamma_1+\delta_1$(D3)=$\alpha_1+\beta_1+\gamma_1+\delta_1$(D1)+$\alpha_2+\beta_2+\gamma_2+\delta_2$(D4).

The ends of paths of schedule (D1) and (D3) are the same and the ends of paths of schedule (D2) and (D4) are the same. Therefore is (D2)+(D3)=(D1)+(D4), for the makespan and tardiness.

(E) Now empties the queue not between time j and time T, so now just before time T there is one patient less $(m'-1)$ in queue for schedule (E3) and (E4). Just after time T there will be for schedule (E1) and (E4) m' patients, for schedule (E2) $m'+1$ and for schedule (E3) $m'-1$ in queue. Now the total waiting time if there is m patients left is given by $\frac{m(m-1)\frac{1}{\mu}}{2}$. Because this is a convex function it is clear that the waiting time function satisfies $\alpha_2+\beta_1+\gamma_2+\frac{(m'+1)m'\frac{1}{\mu}}{2}$(D2)+$\alpha_1+\beta_2+\gamma_1+\frac{(m'-1)(m'-2)\frac{1}{\mu}}{2}$(D3)$\geq$ $\alpha_1+\beta_1+\gamma_1+\frac{m'(m'-1)\frac{1}{\mu}}{2}$(D1)+$\alpha_2+\beta_2+\gamma_2+\frac{m'(m'-1)\frac{1}{\mu}}{2}$(D4).

Let $s = d(T-1)$. The main finishing time of the day will be at $s+m'\frac{1}{\mu}$ for schedule (E1) and (E4), $s+(m'+1)\frac{1}{\mu}$ for schedule (E2) and $s+(m'-1)\frac{1}{\mu}$ for schedule (E3). So $s+(m'+1)\frac{1}{\mu}$(E2)+$s+(m'-1)\frac{1}{\mu}$(E3)=$s+m'\frac{1}{\mu}$(E1)+$s+m'\frac{1}{\mu}$(E4) for the makespan and tardiness.

For "Case F" until "Case I" (Figure 5) counts that before time $j-1$ the queue does not empty, so just before time $j-1$ there are k patients in queue for schedule (1) and (3) and for schedule (2) and (4) one less, so $k-1$. Until time $j-1$ schedule (1) and (3) have a total waiting time α_1 and schedule (2) and (4) a total waiting time α_2.

Just after time $j-1$ there will be k' patients fore schedule (1) and (4), $k'-1$ patients for schedule (2) and $k'+1$ patients for schedule (3).

(F) For "Case F", after time $j-1$ we can distinguish the following four possibilities.

- For schedule (F1) there are say l patients left just before time j and just after time j it is l'. Say that the total waiting time between time $j-1$ and j is β and after time j γ_1.
- For schedule (F2) the number of patient is $l-1$ just before time j. So just after time j there are $l'-1$ patients in queue and the total waiting time between time $j-1$ and j is then $\beta-d$ and after time j γ_2.

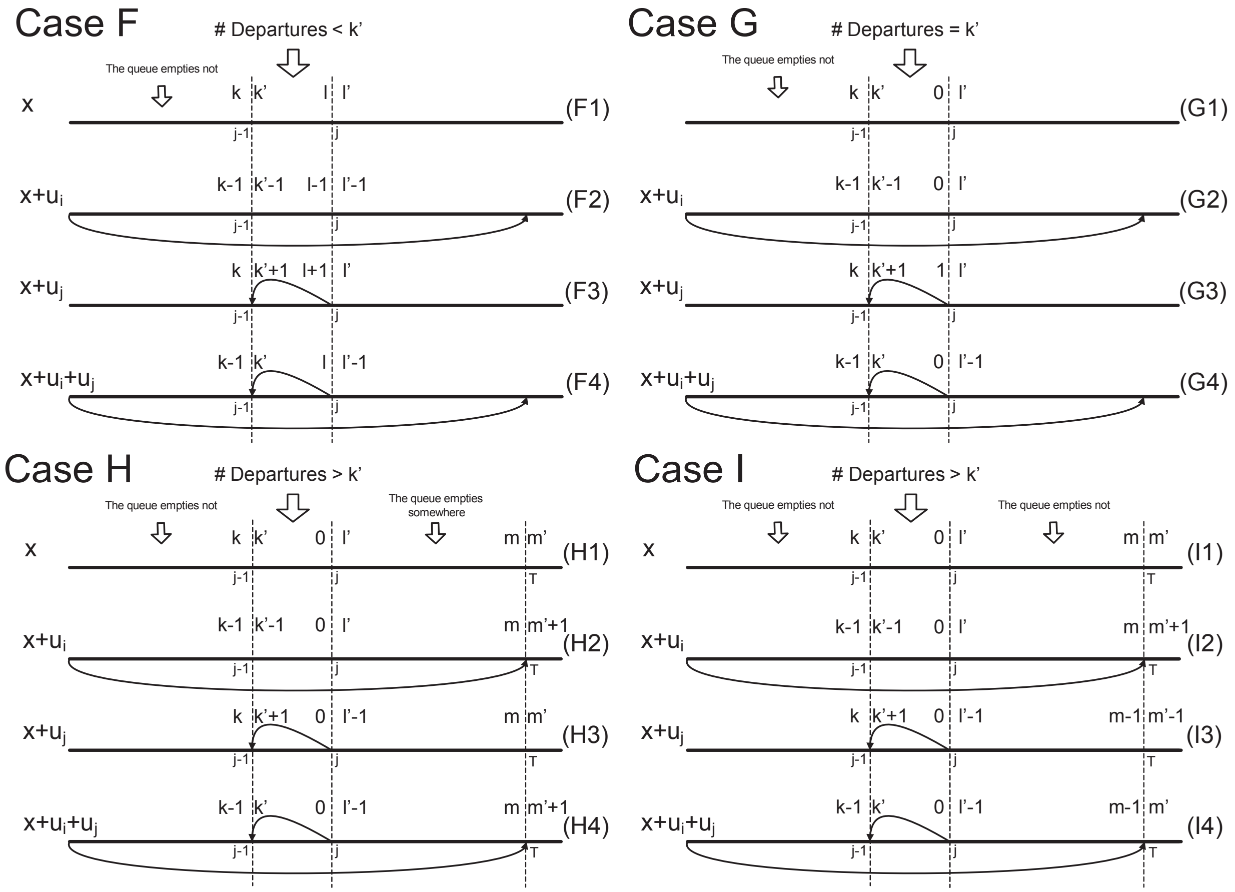

Figure 5: Case F, G, H & I ($1 = i < j \leq T$)

- For schedule (F3) the number of patient is $l+1$ just before time j. Just after time j there are l' patients in queue (one patient comes less) and the total waiting time between time $j-1$ and j is then $\beta + d$ and after time j again γ_1 (same path as schedule (F1)).
- For schedule (F4) the number of patient is l just before time j. Just after time j there are $l'-1$ patients in queue (one patient comes less) and the total waiting time between time $j-1$ and j is again β and after time j again γ_2 (same path as schedule (F2)).

Now we see that the waiting time satisfies $\alpha_2 + \beta - d + \gamma_2$(F2)+$\alpha_1 + \beta + d + \gamma_1$(F3)=$\alpha_1 + \beta + \gamma_1$(F4)+$\alpha_2 + \beta + \gamma_2$(F4)

The end of the path (after time j) of schedules (F1) and (F3) are the same, and also (F2) and (F4) are the same. So in this case (F2)+(F3)=(F1)+(F4), for the makespan and tardiness.

(G) Now for "Case G" there are exactly k' departures between time $j-1$ and j.

- For schedule (G1) there will be $k'-k'=0$ patients left just before time j and just after time j it will be l'. Say that the total waiting time between time $j-1$ and j is β and after time j γ_1.

- For schedule (G2) the number of patient shall also be 0 just before time j. So just after time j there are l' patients in queue and the total waiting time between time $j-1$ and j is then $\beta - d$ and after time j γ_2.
- For schedule (G3) the number of patient shall be $k'+1-k'=1$ just before time j. So just after time j there are l' patients in queue and the total waiting time between time $j-1$ and j is then $\beta + d$ and after time j again γ_1 (same path as schedule (G1)).
- For schedule (G4) the number of patient shall be $k'-k'=0$ just before time j. So just after time j there are $l'-1$ patients in queue and the total waiting time between time $j-1$ and j is then β (same as (G1)) and after time j γ_3 which is of course smaller than γ_2, because 1 patient is less to do.

Now we see that the waiting time satisfies $\alpha_2+\beta-d+\gamma_2$(G2)+$\alpha_1+\beta+d+\gamma_1$(G3)$\geq$ $\alpha_1+\beta+\gamma_1$(G1)+$\alpha_2+\beta+\gamma_3$(G4).

The end of the path (after time j) of schedules (G1) and (G3) are the same so the makespan and tardiness are the same for these schedules. At the end of the path of schedule (G4) there are one patient less (or in the worst case the same) than at the end of path (G2), so the makespan and tardiness shall also be less or equal than schedule (G2). So we can conclude that (G2)+(G3)$\geq$(G1)+(G4), for the makespan and tardiness.

(H,I) Now for Case “H” and “I” there are more than k' departures between time $j-1$ and time j. So just before time j there are no patients left for all schedules. Between time $j-1$ and j schedules (1) and (4) have the same total waiting time of $\frac{k'(k'-1)d}{2s}$, schedule (2) $\frac{(k'-1)(k'-2)d}{2s}$ and schedule (4) $\frac{(k'+1)k'd}{2s}$ (same as in “Case $\mathrm{B_C}$”).

Just after time j the schedules will follow the same path as in ”Case D” and “Case F”, which we already discussed, so for the makespan and tardiness it is immediately clear that it satisfies the multimodularity.

Now for the waiting time it is also clear because before time j it satisfies the multimodularity, and after time j also.

Now we distinguished all possible cases and we proved that in each case the waiting time, makespan and tardiness are multimodular. Thus the same holds for the idle time.

The proof can easily be extended to include no-shows. This is done by conditioning on the no-shows: we get the same model as without no-shows but with less patients planned.