www.stefancreemers.be · ✉ info@stefancreemers.be

Evaluation of appointment scheduling rules

Stefan Creemers
Pieter Colen
Marc Lambrecht

1 Introduction

Professionals in healthcare and other services face the problem of allocating time windows to customers. This allocation can be done by means of so-called "appointment scheduling rules" (ASR). ASR determine when a customer receives service during a service session. Although the literature on ASR is mainly focussed on healthcare, the research topic is generic and applicable in many industries: attorneys, faculty receiving students, barbers, automobile service centers and many others.

Conducting scheduled appointments on time is becoming ever more important across service industries. Timeliness of appointments is a key concern both for patients seeking treatment [1] and for customers who wait for field service [2]. The customer waiting time consequently is a relevant performance measure. A second important objective of appointment scheduling has to do with the efficiency of the service. For private companies, the impetus to efficiency comes naturally. But also healthcare systems are under pressure to use their capacity effectively and efficiently. Doctors' (or more general servers') idle time and overtime are hereby important performance measures. The objective of this article is to identify ASR that minimize customer waiting time, server idle time and overtime. This has to be done in an environment where both demand and supply characteristics are highly uncertain and subject to many sources of variability.

ASR determine the planned (scheduled) arrival rate of customers during a service session. The actual arrival time may of course be different from the planned arrival time. We therefore assign each customer a probability of being too late or too early. In addition, we fix for each customer a probability of not showing up. The performance of ASR is not only influenced by the arrival rate and service rate characteristics. Other types of outages during the service session are also important. We therefore allow for delays at the start of a service session due to late arrival of the doctor or due to

setup activities at the start of a session. We also allow preemptive and non-preemptive interrupts during the service session. All these extensions allow us to model real-life appointment systems and to identify ASR that have a robust performance across different settings.

We develop an analytical model that uses an efficient (in terms of computational and memory requirements) algorithm to assess the performance of ASR. The validity and accuracy of the model are supported by a simulation study. We use the model to assess the performance of a set of 314 ASR in an elaborate computational experiment. To compare the performance of these ASR (in terms of waiting time, idle time and overtime), we apply a data envelopment analysis (DEA).

2 Model Description

We focus on static ASR (i.e., decisions are made prior to the start of the service session). The Markov model used to evaluate the performance of an ASR has the following properties:

- Customers are served by a single server.
- Customers have i.i.d. service time distributions.
- All customers that arrive during the service session receive service.
- Customers that arrive early (i.e. prior to their scheduled arrival time) receive service if the server is idle. Note that this implies the possibility of overtaking other customers.
- The arrival process can differ for each individual customer

Contrary to other existing models, we allow for an individual characterization of the arrival process for each customer. In addition, computational performance and model accuracy (and hence practical applicability) of our model significantly exceed the capabilities of comparable models in the literature on appointment systems. The performance measures of interest are: (1) the expected customer waiting time; (2) the expected amount of server overtime; (3) the expected amount of server idle time. Our model may be used to obtain these performance measures for any given schedule of customers. However, we limit ourselves to the comparison of the performance of schedules that are generated by a set of 314 ASR.

In general, ASR are classified in terms of:

- A_i, the scheduled arrival time of customer i,
- μ^{-1}, the mean service time requirement of a customer,
- σ_i, the standard deviation of the service time requirement of customer i,
- N, the number of customers that require scheduling.

We implement a set of 314 ASR and use an analytical model to perform an extensive computational experiment in which the performance of these rules is assessed with respect to the three performance measures in a wide variety of settings. The adopted set of ASR is an extension of the set of 50 ASR selected in [3, 4]. These ASR are common in daily practice or have been shown to yield good and robust results [3, 4].

The ASR may be summarized as variations of: (1) the individual appointment rule; (2) the block appointment rule; (3) early-lateness rules (hereafter referred to as the EL rules).

The individual appointment rule schedules the arrival times of customers as follows:

$$\begin{array}{lcll} A_i & = & ia\mu^{-1} & \forall i : i < l, \\ A_i & = & A_{i-1} + \mu^{-1} + h\sigma_i & \forall i : l \leq i < N. \end{array} \tag{1}$$

Where: (1) a is a multiplier to delay the start of the first arriving customers; (2) l denotes the number of customers scheduled for arrival at the start of a service session; (3) h is a multiplier used to adjust the impact of σ_i. We implemented 91 variants of the individual appointment rule by allowing parameters a, l and h to vary over set $\{0, 0.3, 0.5\}$, set $\mathbf{L} = \{1, 2, 3, 4, 5\}$ and set $\mathbf{H} = \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$ respectively.

The block appointment rule may be summarized as follows:

$$\begin{array}{lcll} A_i & = & 0 & \forall i : i < b, \\ A_{nb} & = & A_{(n-1)b} + b\mu^{-1} + h\sqrt{b\sigma_{nb}} & \\ & & \forall n : 1 \leq n < \frac{N}{b}, & \\ A_{nb+i} & = & A_{nb} & \\ & & \forall i : 1 \leq i < b. & \end{array} \tag{2}$$

Where b denotes the block size (i.e. the number of customers assigned to arrive at a single time instance). Varying parameters b and h over set $(\mathbf{L}\backslash\{1\})$ and set $\mathbf{H}$ respectively, we obtain 28 ASR.

The EL rules speed up and/or slow down the pace of scheduled arrivals using correction factors r_1 and r_2. The computation of scheduled arrival times is performed in two steps. First, all scheduled arrival times are initialized using the individual appointment rule where $(l = 1)$ and $(h = 0)$. Next, a correction is applied to speed up and/or slow down the pace of scheduled customer arrivals.

$$\begin{array}{l} \text{Initialization:} \\ A_0 \;=\; 0, \\ A_i \;=\; A_{i-1} + \mu^{-1} \quad \forall i : 1 \leq i < N. \\ \text{Correction:} \\ A_i \;=\; A_i - r_1(z-i)h\sigma_i \quad \forall i : 1 \leq i \leq z, \\ A_i \;=\; A_i - r_2(z-i)h\sigma_i \quad \forall i : z < i < N. \end{array} \tag{3}$$

Where: (1) r_1 and r_2 are correction factors used to speed up or slow down the succession of scheduled arrivals; (2) z is any multiple of 5 smaller than N. Parameter r_1 controls the arrival pace of the first z customers; the arrival pace of these customers increases as r_1 increases. Conversely, parameter r_2 controls the arrival pace of those customers that are scheduled to arrive after customer z; the arrival pace of these customers decreases as r_2 increases. When varying parameter h over set $(\mathbf{H} \setminus \{0\})$ and parameters r_1 and r_2 over the set $\{0, 1, 2\}$ (where $(r_1 + r_2) > 0$), we obtain 39 times times $\left\lfloor \frac{N-1}{5} \right\rfloor$ ASR.

A summary of the ASR is given in table 2. We determine the performance of the ASR in a wide range of environmental settings. In the most simple environment, all customers arrive punctually at their assigned appointment time. Complexity is introduced in the form of so-called "environmental variables". An extensive overview of these environmental variables is provided in [5]. We take the following environmental variables into account:

- The number of customers served during a service session.
- The customer service time.
- Customer unpunctuality (i.e., the early/late arrival of customers).
- Customer no-shows.
- Late start of the service session.
- The variability of service times and customer unpunctuality.

By combining different values for the environmental parameters, we obtain 243 environmental settings in which we evaluate the performance of the ASR.

Rule	$A_i = ia\mu^{-1}, \quad \forall i : i < l,$
	$A_i = A_{i-1} + \mu^{-1} + h\sigma_i, \quad \forall i : l,$
Rule no.	$1-7, 8-14, 15-21, 22-28, 29-35,$
Conditions	$l = 1, 2, 3, 4, 5 \wedge a = 0 \wedge h \in \mathbf{H},$
Rule no.	$36-42, 43-49, 50-56, 57-63,$
Conditions	$l = 2, 3, 4, 5 \wedge a = 0.3 \wedge h \in \mathbf{H},$
Rule no.	$64-70, 71-77, 78-84, 85-91,$
Conditions	$l = 2, 3, 4, 5 \wedge a = 0.5 \wedge h \in \mathbf{H},$
Rule	$A_i = 0, \quad \forall i : i < b,$
	$A_{nb} = A_{(n-1)b} + b\mu^{-1} + h\sqrt{b\sigma_{nb}}, \quad \forall n : 1 \leq n < \frac{N}{b},$
	$A_{nb+i} = A_{nb}, \quad \forall i : 1 \leq i < b,$
Rule no.	$92-98, 99-105, 106-112, 113-119,$
Conditions	$b = 2, 3, 4, 5 \wedge h \in \mathbf{H},$
Rule	$\text{initialize} A_0 = 0, \quad A_i = A_{i-1} + \mu^{-1}, \quad \forall i : 1 \leq i < N, \text{then}$
	$A_i = A_i - r_1(z-i)h\sigma_i, \quad \forall i : 1 \leq i \leq z,$
	$A_i = A_i - r_2(z-i)h\sigma_i, \quad \forall i : z < i < N,$
Rule no.	$120-125, 159-164, 198-203, 237-242, 276-281,$
Conditions	$z = 5, 10, 15, 20, 25 \wedge r_1 = 0 \wedge r_2 = 1 \wedge h \in (\mathbf{H} \setminus \{0\}),$
Rule no.	$126-128, 165-167, 204-206, 243-245, 282-284,$
Conditions	$z = 5, 10, 15, 20, 25 \wedge r_1 = 0 \wedge r_2 = 2 \wedge h \in \{0.2, 0.25, 0.3\},$
Rule no.	$129-134, 168-173, 207-212, 246-251, 285-290,$
Conditions	$z = 5, 10, 15, 20, 25 \wedge r_1 = 1 \wedge r_2 = 0 \wedge h \in (\mathbf{H} \setminus \{0\}),$
Rule no.	$135-140, 174-179, 213-218, 252-257, 291-296,$
Conditions	$z = 5, 10, 15, 20, 25 \wedge r_1 = 1 \wedge r_2 = 1 \wedge h \in (\mathbf{H} \setminus \{0\}),$
Rule no.	$141-146, 180-185, 219-224, 258-263, 297-302,$
Conditions	$z = 5, 10, 15, 20, 25 \wedge r_1 = 1 \wedge r_2 = 2 \wedge h \in (\mathbf{H} \setminus \{0\}),$
Rule no.	$147-149, 186-188, 225-227, 264-266, 303-305,$
Conditions	$z = 5, 10, 15, 20, 25 \wedge r_1 = 2 \wedge r_2 = 0 \wedge h \in \{0.2, 0.25, 0.3\},$
Rule no.	$150-155, 189-194, 228-233, 267-272, 306-311,$
Conditions	$z = 5, 10, 15, 20, 25 \wedge r_1 = 2 \wedge r_2 = 1 \wedge h \in (\mathbf{H} \setminus \{0\}),$
Rule no.	$156-158, 195-197, 234-236, 273-275, 312-314,$
Conditions	$z = 5, 10, 15, 20, 25 \wedge r_1 = 2 \wedge r_2 = 2 \wedge h \in \{0.2, 0.25, 0.3\}.$

Table 1: Summary of the different appointment scheduling rules

3 Results

After obtaining the relevant performance measures by means of the Markov model, we perform a DEA to get a performance ranking of all the ASR. This ranking enables us to select those ASR that work best under a wide range of settings and performance criteria. In order to assess the performance of an ASR, we use a composite indicator (CI):

$$\mathrm{CI}_i = \frac{1}{\sum_{k=1}^{3} w_{i,k} x_{i,k}}, \tag{4}$$

where: (1) $w_{i,k}$ is the weight assigned to performance measure k for ASR i and (2) $x_{i,k}$ is the score of ASR i for performance measure k. Weights $w_{i,k}$ are determined by means of a DEA. We use the Maverick index to assess the sensitivity of a CI-score with respect to the chosen weight set (refer to [6] for further details).

Table 2 lists the best performing ASR. We report the average CI values across the different environmental settings. In the table we also indicate the type of ASR (i.e., individual, block or EL) and the sensitivity of the chosen weight set (i.e., the Maverick index). It is striking that the six best performing ASR are all individual ASR with the common characteristics of zero delay for the first arrival, an initial number of patients of three or more and with a very small to zero adjustment for service time variance. This observation is encouraging for practitioners as these rules are simple to implement. At the downside, the maverick index indicates that the efficiency scores of these simple ASR are quite sensitive to the weight selection. The lower sensitivity to the weight selection can be a reason to opt for an EL rule, which are firmly established in the top 15 as from rank seven. Block ASR are clearly the less attractive type of ASR. The best performing block ASR have a block size of two with a small or zero adaption for service time variance. These findings corresponds with the conclusions of earlier research [7, 3]. Lastly, we note that the simple Bailey-Welch rule with two initial arrivals followed by one-by-one arrivals (ASR 8) performs rather well both in terms of the efficiency score and weight sensitivity (respectively 99,81% and 0,0959).

We can conclude that individual ASR have the best performance across environments taking into account server overtime, idle time as well as customer waiting time. EL ASR, however, seem to be more robust with respect

Rank	ASR	CI (%)	Type	Weight sensitivity
1	15	99.9980	Individual	0.195624
2	22	99.9976	Individual	0.313921
3	30	99.9964	Individual	0.389075
4	23	99.9954	Individual	0.284273
5	29	99.9941	Individual	0.412453
6	16	99.9838	Individual	0.163625
7	273	99.9724	EL	0.066703
8	162	99.9650	EL	0.171877
9	256	99.9453	EL	0.070306
10	298	99.9184	EL	0.064002

Table 2: Ranking of ASR based on average efficiency across environments

to the value judgement (i.e., the importance assigned to each of the performance measures). The performance of block ASR is dismal and they should be avoided because better alternatives clearly exist.

References

[1] Grote, K.D., Newman, J.R. and Sutaria, S.S., A better hospital experience, The McKinsey Quarterly, November, 1-11 (2007)

[2] Apte, A, Apte, U. M. and Venugopal, N., Focusing on customer time in field service: A normative approach, Production and Operations Management, 16(2), 189-202 (2007)

[3] Ho, C. J. and Lau, H. S., Minimizing total-cost in scheduling outpatient appointments, Management Science, 38(12), 1750-1764 (1992)

[4] Ho, C. J. and Lau, H. S., Evaluating the impact of operating conditions on the performance of appointment scheduling rules in service systems, European Journal of Operational Research, 112(3), 542-553 (1999)

[5] Cayirli, T. and Veral, E., Outpatient scheduling in health care: a review of literature, Production and Operations Management, 12(4), 519-549 (2003)

[6] Doyle, J. and Green, R., Efficiency and cross-efficiency in DEA - derivations, meanings and uses, Journal of the Operational Research Society, 45(5), 567-578 (1994)

[7] Welch, J.D., Appointment systems in hospital outpatient departments, Operations Research Quarterly, 15(3), 224-232 (1964)