A SIMPLE QUEUE MODEL FOR AN APPOINTMENT SYSTEM AND APPLICATIONS IN A HOSPITAL CT SCAN FACILITY

F. Boenzi

DMMM (Department of Mechanics, Mathematics and Management) – Engineering Faculty at Politecnico di Bari (Italy)

boenzi@poliba.it

ABSTRACT

Key features of modern healthcare services in developed countries, in a general context characterized by strict cost control and aging population, are flexibility in access to care for different types of patient flows (e.g. inpatients, ED patients and outpatients) and capability of adapting to quantitative service demand change, commonly posed by outpatients. An expensive resource, such as a CT scanner or an MRI machine, is often shared by diverse patient flows and this single server has to assure satisfactory performances (in terms of waiting time) to both the random and the planned service demand component, implementing appropriate buffering and priority rules. For outpatients, requesting access to an appointment service in general, waiting time is represented by waiting time for the appointment day and waiting time experienced at the facility. The governing rules determining these performance results can be treated as separate problems and, in the present paper, the first kind of wait, commonly related to the existence of long waiting lists, is considered. A simple general analytical model, adopting statistical considerations, is proposed and, successively, applications in a CT scan facility in a public hospital, supported by simulation, are illustrated.

Keywords: appointment scheduling, queues, health care, simulation

1. INTRODUCTION

In many European countries, healthcare services still represent a basic pillar of the welfare system; however, public expenditure for new investment is often unrealizable by the local NHS (National Health Service) Departments/Authorities.

As a consequence, it is not uncommon that, at the examination facility, an expensive resource for digital imaging diagnosis (e.g. a CT scanner, MRI equipment) is unique and shared among diverse patient flows. On the other hand, for patients in critical conditions (e.g. patients from the Emergency Department in a hospital), an immediate response is crucial, but, ideally, it would require dedicated equipment. As a trade-off solution, in a single-server system, appropriate buffering and priority rules have to be implemented in order to assure satisfactory performances (in terms of waiting time) to both the random and the planned service-demand component.

The last consists of outpatients who are given appointments well in advance with respect to the examination date; for them a booking method, in most cases supported by IT tools, is in place. The coexisting presence of a random component can be considered beneficial in order to attain a "satisfactory" utilization level of expensive resources, but, on the other hand, as well-known, "emergency medical (ambulance) and fire service systems operate at only 10-15% utilization. Short response time is essential for these systems and hence excessive waiting times are simply not allowable. Since variability is unavoidable (people don't schedule their heart attacks), the only way to keep waiting to a minimum is to have a great deal of excess capacity" (Hopp 2008). Therefore, two opposite needs must be conciliated. On the one hand "hospital managers have every incentive to keep these machines fully utilized" and "this is often done by filling most, if not all, examination slots during the day with outpatient appointments" (Green et al. 2006). On the other hand, this is in contrast with the outlined features of an emergency service, since, in order to keep short waiting time, urgent cases should find, ideally, the server idle upon their arrival (null waiting time) or working at low utilization levels (which means the server is found idle most of the times). Because of the presence of appointed patients and/or inpatients (when the service facility is inside a hospital), priority rules must necessarily be adopted and, inevitably, longer waiting times for outpatients result.

Extensive literature exists, regarding the analysis of system performance in terms of waiting time at the facility for different patient classes and the risk of overtime (prolonged working time beyond the daily shift), in relation to appointment scheduling and other problem parameters. However, the investigation on specific characteristics of healthcare appointment systems in relation to the existence of long waiting lists for outpatients appear to be overlooked.

The present paper is organized as follows: in Section 2, a brief literature overview is given; in Section 3, some general considerations about the behavior of a generic appointment system in steady-state conditions are illustrated. In Sections 4 and 5 a possible analytical

approach, adopting statistical results, for the description of the outpatient buffer in unsteady-state conditions is proposed. In Section 6, basic simulation models, adopting the illustrated concepts for the specific case-study of a CT scan facility, are introduced and results are discussed. Finally, remarks and general conclusions are drawn in Section 7.

2. LITERATURE SURVEY

The kind of system illustrated in the introduction can be treated of as a single-server queue model with a non-preemptive priority rule (Adan and Resing 2002).

In fact, even though ED patients are served as soon as possible (and the FCFS, first-come-first-served, rule is adopted for patients of the same class), the process cannot be interrupted once started for any patient, even if he belongs to a lower priority class. Using Kendall's notation, it isn't possible to describe the system in a short comprehensive manner, because of the diverse nature of these patient flows. For the appointed patients, an appropriate representation could be in the form *D/G/1* (deterministic arrivals based on the appointment schedule in place and general *pdf* for service times). However, the mathematical model becomes easily untreatable when service time moves away from a Markovian distribution or when it is not assumed to be deterministic, or when phenomenons like no-shows or unpunctuality must be taken into account (*D + noise/M/1* models, cited, for example, in (Hassin and Mendel 2008). In (Kuiper 2012), the author illustrates the solution method for finding the optimal appointment schedule (which minimizes a cost function of the sum of expected patients' waiting times and server's idle times) by means of a phase-type distribution approximation of service time (*D/PH/1* setting), neglecting the mentioned disturbances. (Kaandorop and Koole 2007) find the optimal outpatient appointment schedule, with no-shows and exponential service time, for the minimization of an objective function which includes tardiness (time exceeding the given session period), discretizing the time session and employing a local search algorithm. Exploring over a limited number of time points for appointments, derived dividing the continuous time session into discrete intervals, makes sense for practical applications and doesn't limit the general principles behind the optimal solution, found to be in "dome shapes", as observed in literature. (Hassin and Mendel 2008) use a continuous time analytical model *D/M/1*, including no-shows and an objective function to be minimized (linear combination of waiting time cost and server's availability time cost). They obtain "what is believed to be an optimal solution by applying sequential quadratic programming (SQP)".

The possibility of including in a mathematical model additional types of patients aside the appointed ones often implicates the assumption of deterministic service times and rigidly fixed slot durations. For example, (Green et al. 2006) formulate the real-time capacity allocation problem for a medical diagnostic facility (MRI) accessed by outpatients, inpatients and emergency clients (with associated arrival probabilities and a daily expected total reward function, with overtime implicitly included in the form of penalty costs for unseen patients), as a finite horizon dynamic programming problem. They highlight the relationship between the "strategic" decision level of the system represented by the Appointment Schedule (AS) rule, and the "tactical" decision level regarding the set of dynamic priority rules for admitting patients into service. For the latter, they investigate the properties of the optimal capacity allocation policy, and propose a simpler linear approximation (LA) heuristic too. Results are illustrated for different heuristic appointment policies ("fill all slots", "balanced" and "newsvendor") and problem parameters. The mentioned decision levels could also be framed in the more general context of healthcare planning-and-control as "offline" and "online" operational levels (Hans et al. 2011), in which the first "concerns the in-advance planning of operations" and the second "involves control mechanisms that deal with monitoring the process and reacting to unforeseen or unanticipated events". (Kolisch and Sickinger 2008) examine the same maximization mathematical model, in the form of a Markov Decision Process, for a CT scan facility, serving the above mentioned patient classes. They illustrate the results of the optimal tactical decision policy (found by the "backward induction algorithm") and compare them with simpler decision rules, which can be applied manually (Linear approximation, FCFS and random choice), under three commonly adopted AS schemes. (Sickinger and Kolisch 2009) illustrate the results of the objective function, obtained optimally solving the associated stochastic dynamic decision problem, for a Generalized Bailey-Welch (GBW) schedule, compared to optimal and Neighborhood Search (NS) heuristic, under various combinations of problem parameters. (Patrick 2012) develops a Markov Decision Process model of an outpatient clinic and "demonstrates that a short booking window does significantly better then open access". The model takes into account the no-show rate, but assumes a deterministic service-time.

In order to remove the numerous assumptions that a mathematical model poses, extensive literature, adopting discrete event simulation models, exists and the authors follow an experimental approach. (Klassen and Rohleder 1996) carry out a full factorial ANOVA analysis on a simulation model (with lognormal service time and no-shows), taking as decision variables 1) three possible positions of a couple of open slots to accommodate urgency; 2) a total of 10 known pre-defined AS rules. They illustrate results for server idle time, total client waiting time and a linear combination of the two, for various environmental factors. In (Rohleder and Klassen 2002), attention is shifted to the observation of the system from the wider perspective of the scheduler of appointments, considering a rolling-horizon environment with fluctuating demand loads. The model takes into account the additional client

performance measure of the number of days between call and appointment, and the server oriented measure of the ending time of the session (possible overtime). Different trade-off solutions for the combination of overloading rules and rule delay choices are illustrated, under six examination demand patterns, highlighting that the managerial choice depends on which performance measure is considered to be the most important. In (Hutzschenreuter 2004) a simulation model, which accounts for no-shows and unpunctuality of patients, is utilized to test the performances of a general outpatient clinic (in terms of mean patient waiting time and mean physician's idle time), adopting three well-known AS rules and under different combinations of mean service time and variability. (Zhu et al 2009) find empirically, on a simulation model of a specialist outpatient clinic, the optimal number of scheduled patients per session for three system configurations, considering four performance measures (queue length, patient waiting time, overtime and doctors' utilization) and a previously established AS rule for each configuration. They point out how the result can vary according to the chosen performance measure. (Wijewickrama 2006) examines the simulation model of an outpatient department with different patient flows (appointed and not, with priority rule for the first ones), not including no-shows and unpunctuality. He illustrates the efficient frontier of results in terms of weighted average of patient waiting time and servers' idle time, adopting four basic AS rules and some combinations.

3. GENERAL MODEL FRAMEWORK

As initially stated, healthcare services open to different patient flows are usually subject to demand variations. These variations are quite ordinary for the random component in the course of a day (e.g. ED patients) and in simulation models are usually averaged as *HPP*.

In the longer term, also the appointed component of the demand can exhibit variations, whereas the number of appointment slots could not have been adjusted adequately in the course of time. As a result, in case of average demand greater then the number of available appointments in a time period (for example, a week), the backlog represented by the list of waiting outpatients will increase and, correspondingly, mean waiting time for an appointment. Attention can also be focused on the opposite case, when a long waiting list exists and, with the aim of reducing it, agreed decisions between managers and medical staff have to be taken about possible increments of the number of appointments (exceeding the current demand). A question in this case could regard how long time it is necessary to maintain the system's workload greater then normal to attain the objective and to reach a balanced situation.

It's to keep in mind that short waiting time for an appointment may be an external constraint, set at the legislative level for specific patient classes. The last classification is often based on urgency criteria (outpatient priority classes) and must be adopted by family or specialty doctors, when they prescribe a diagnostic examination, as in the case of CT scans (for example, for patients to be followed-up).

Setting a greater number of appointments, when needed, should be framed in the general context of a demand-offer problem, disjointed from the assignment, which could be addressed on a second step, regarding how to displace additional appointments appropriately (avoiding excessive waiting time at the facility and/or overtime).

Looking at the entire appointment process, at the origin, we can find the service request generation, for example phone-calls at call-centers, which, as a whole, constitute the demand. Service requests are differentiated for each outpatient class and enqueued into the corresponding buffer, usually processed on a FAFS (first-appointment-first-serve) basis. A patient is given an appointment date corresponding to the first available slot (for that priority class/buffer) and slots in the future are filled so on. Therefore, any appointment system, at a macroscopic scale, can be thought of as a "black-box", which processes the random demand with a set "productive capability", equal to the number of appointments given in a time period, according to the current schedule in place.

Typically, in healthcare systems, a repetitive scheduling scheme is specified over a week, with possible variations from day to day, setting time and number of appointments for each class of outpatients, when separated booking agendas are in place (commonly corresponding to the mentioned priority classes). The approach presented herein treats each buffer (making up a particular booking agenda) separately. Therefore, if we consider the time measurement unit equal to the repetitive schedule time horizon (for example, one week) and we renounce the exact quantification of patient waiting time within one time unit (the week during which he will be given the appointment), we can look at each buffer in the system as a *G/D/1* single server queue. Our attention is focused on waiting time for an appointment and on how it is related to the buffer length.

The arrival process of clients (with an associated *pdf* for inter-arrival times in general form) is represented by the service request generation. Service time is not related to real machine times, but it's determined by the set "productive capability", or number of appointments. It can be considered a fictitious deterministic time, under the hypothesis that, during each period (time unit), the stated number of appointments will be processed for sure and no patient will be re-scheduled. Therefore, disregarding the way in which appointments are really allocated in the course of the scheduling period of time length $T=1$ *[time unit]*, we are just interested in their total number per period, said N_P. Assuming an equally spaced distribution of the appointments, the fictitious deterministic service time t_0 is simply equal to T/N_P *[time units]*, with constant service rate $\mu = N_P/T$. Of course, the real utilization of

the server isn't related to t_0, because, for the feasibility of the scheduled appointments, the mean processing time must necessarily be smaller (e.g. mean time for an examination / a visit). Furthermore, the real overall utilization of the service will be the result of sharing the resource among all the different patient flows. Nevertheless, such a system *G/D/1* can describe appropriately what happens in the outpatient-buffer (not beyond it) and the above defined service rate is the fundamental parameter to be compared to the average arrival rate of clients during a time period. If we define $\overline{N_A}$ as the average number of service requests (arrivals) in T and, correspondingly, λ as the arrival rate $\overline{N_A}/T$, it is well-known that the system can remain in steady-state conditions as far as the condition $\lambda<\mu$ is met, that's to say $\overline{N_A}<N_P$. An approximate expression of the expected waiting time in the buffer is given writing Kingman's equation in the following form:

$$\varphi_b \cong \frac{c_a^{\,2}+c_0^{\,2}}{2}\frac{U}{1-U}t_0 = \frac{c_a^{\,2}}{2}\frac{U}{1-U}\frac{T}{N_P} = \frac{c_a^{\,2}}{2}\frac{\overline{N_A}}{N_P}\frac{T}{N_P-\overline{N_A}}. \qquad (1)$$

$U = \lambda/\mu = \overline{N_A}/N_P$;
c_a: coefficient of variation of the arrival process;
c_0: coefficient of variation of the fictitious service process; being the process deterministic = *0*;
T = 1 [time unit].

The U ratio, although isn't the real utilization of the server, represents the utilization of the capacity at disposal. Under the assumption of a purely random arrival process with average $\overline{N_A}$, a Poisson distribution can be adopted, the system can be denoted as *M/D/1* and in this case c_a=1. Then, it's possible to observe that, since N_P must be an integer number (scheduled number of appointments), adopting the following very simple rule that assures the demand/offer balance of the system (steady-state conditions):

$$N_P = \text{upper int}\{\overline{N_A}\}+1. \qquad (2)$$

the calculated value of φ_b would always remain lower then *0.5 T [time units]*, as shown in Fig.1.

Furthermore, from Little's law, the average number of clients in the buffer $\varphi_b \cdot \lambda$, which represent the waiting list length, would remain lower than half $\overline{N_A}$. Numerically, it is also possible to verify that for $\overline{N_A} > 7.2$ and for $\overline{N_A} > 17.1$ the value of U always results above *0.8* and *0.9*, respectively, which means that, adopting (2), the system would be over-capacitated in percentages from *+25%* to *+11.1%*, respectively, with reference to the demand.

Average waiting time for an appointment within one period T (for example, one week) implies that for the scheduler it's possible to implement an approach very similar to "open access", since service demand is met quite immediately. Of course, it will always be possible to set appointments far in the future (e.g. for examinations which require previous particular treatments), but the deriving waiting time becomes a scheduler's choice and is not determined by the governing laws of the system.

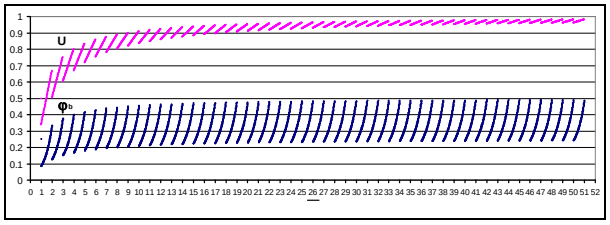

Figure 1: Expected Waiting Time in the buffer φ_b and U versus $\overline{N_A}$ for the *M/D/1* system adopting (2)

In general, it's possible to affirm that for such appointment systems (in steady-state conditions), which can be summarized by the *G/D/1* notation, filtering the demand has the advantage of isolating just one source of variability, determined by the arrival process. Moreover, as observable from Kingman's equation, assumed a Markovian arrival process, the term accounting for stochasticity is 1/2 and, differently from what happens for other generic queue systems, average waiting time can't reach elevated values, unless $\overline{N_A}$ is very close to N_P. In fact, conversely, from equation (1) and defined Δ as the difference ($N_P - \overline{N_A}$), it's possible to derive limiting values of Δ for which φ_b is k times smaller than the period T. In mathematical form:

$$\varphi_b \le kT : \Delta \ge \frac{1}{1/N_P + 2k}. \qquad (3)$$

For high values of N_P, neglecting $1/N_P$, $\Delta \ge 1/(2k)$.
In Fig. 2, example values for Δ and U are reported for $k=1,2,3$ (sub-index of labels).

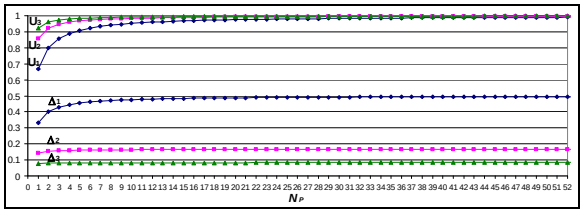

Figure 2: Δ and U derived from (3) versus N_P

As an example of a practical application, set $N_P=10$, the average waiting time in the buffer will never exceed one period T ($k=1$) as far as $\Delta \ge 0.4762$, that's to say $\overline{N_A} \le 9.5238$. It's interesting to note that, for $k=1$, Δ is always less than *0.5* and, practically, becomes independent from N_P, for increasing values.

The elements outlined so far highlight the importance of monitoring the behavior of the demand of a healthcare service in a systematic manner, in order to take corrective actions and keep backlog under control.

In the following two sections, the behavior of the buffer of a generic appointment system in unsteady-state conditions will be investigated by means of an

analytical approach. Firstly, the case with increasing buffer ($\overline{N_A} > N_P$) and secondly the case with initial backlog and decreasing buffer ($\overline{N_A} < N_P$) will be discussed.

4. CASE WITH INCREASING BUFFER

In the following, the unbalanced situation with $\lambda > \mu$ is examined, starting from an empty system and considering the instantaneous number of patients/clients in waiting list $N_b(t)$ and the instantaneous value of $\varphi_b(t)$, defined as waiting time a patient exiting at time t has experienced. Successively, average values are calculated.

4.1. Instantaneous values

Adopting the discretization of time in intervals of length T, the following notation is introduced:

- N_{bi}: number of waiting patients in the buffer at the end of the *i-th* period;
- N_{Ai}: number of requests for appointments arrived at the system and taken into consideration for processing during the *i-th* period (indeed collected at call-centers in the course of the previous *(i-1)-th* period, for practical reasons of scheduling assignment); neglecting waiting time within one period of time, as a "physiological" wait, it's then possible to assume the number of requests N_{Ai} as arrived during the *i-th* period and the server always able to process them, within the limit of its productive capability N_P for each period.

At the end of each period:
$N_{b1} = max\{0, N_{A1} - N_P\}$;
$N_{b2} = max\{0, N_{b1} + N_{A2} - N_P\}$;
....
$N_{bi} = max\{0, N_{b\,(i-1)} + N_{Ai} - N_P\}$; $\forall i$.
Assuming that the buffer does not empty (it always remains a positive integer), also during the initial intervals, from the recursive expression of N_{bi}, it's possible to derive:

$$N_{bi} = \sum_{k=1}^{i} N_{Ak} - iN_P. \tag{4}$$

From (4), assuming that arrivals belong to a Poisson distribution, applying the *Central Limit Theorem* (see Appendix A.1 for details), it's possible to derive:

$$N_{bi} \cong i\left(\overline{N_A} - N_P\right) \pm \sqrt{\overline{N_A}} \cdot t_{value\left\{\gamma\right.^{i-1}} \sqrt{i}\ ; \ (\text{for } i \geq 2). \tag{5}$$

(of course, because of the statistical uncertainty and approximation, N_{bi} isn't anymore an integer number and negative values are not consistent). In Fig. 3, the factor in (5) due to uncertainty and depending solely on i is reported, together with power interpolation functions (dashed lines) in the form $a \cdot i^b$, in the interval *0-500*, for the two confidence levels *90%* and *95%*. In particular, setting lower limits for i so that the percentage difference between the function and its interpolation is lower than *20%* also for initial values of i (then, it rapidly decreases to around *1%*), it's possible to find ($a=1.8255;\ b=0.4814$) for $\gamma=90\%$, with $i \geq 5$ and ($a=2.2206;\ b=0.4778$) for $\gamma=95\%$, with $i \geq 6$.

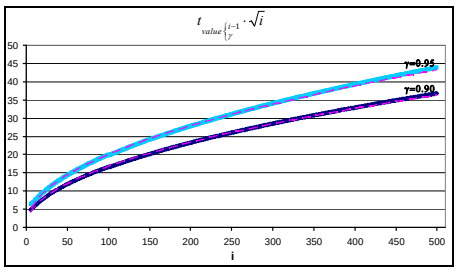

Figure 3: Values of $t_{value\ (i-1)} \cdot \sqrt{i}$ versus i for $\gamma=90\%$ and $\gamma=95\%$ and respective interpolating functions (dashed).

Considering now N_b as a time-continuous function, being $t=i \cdot T$ and substituting time as variable in the above reported interpolating function, it's possible to write:

$$N_b(t) \cong \left(\overline{N_A} - N_P\right)\frac{t}{T} \pm \sqrt{\overline{N_A}} \cdot a\left(\frac{t}{T}\right)^b.\ ^{1} \tag{6}$$

with $t \geq 5T$ and $t \geq 6T$ for $\gamma=90\%$ and $\gamma=95\%$ respectively. The second quantity to be analyzed is waiting time in the buffer for each patient and the following quantities are considered:

- φ_i: total flow-time time $(t_{out\ i} - t_{in\ i})$ for the *i-th* patient, assuming that the time instant of the appointment request generation is shifted one period later (requests must be collected and scheduled at least in a previous period and waiting time within one period of time can be considered unavoidable and neglected). In particular, it is assumed $t_{in1} = 0$.
- φ_{bi}: waiting time in the buffer for the *i-th* patient; assuming that the server is always busy (which is the same previous assumption regarding the buffer never empty) and patients are served on a FCFS basis, exit time $t_{out\ i}$ is deterministic and equal to $i \cdot t_0 = i \cdot T/N_P$; moreover, $\varphi_{bi} = \varphi_i - t_0$.

Observing that $t_{in\ i}$ (time instant of the appointment request) can be viewed as the sum of the preceding *(i-1)*, for $i \geq 2$, inter-arrival times t_{ak} between two generic contiguous requests, it's then possible to write:

$$\varphi_{b1} = \varphi_1 - t_0 = t_{out1} - t_{in1} - t_0 = 0\ ;$$

[1] Indeed, also for the function $N_b(t)$, it should be considered $max\{0;\ calculated\ value\}\ \forall t$, but a resulting empty buffer at any time instant would invalidate the expression because it would be equivalent to shifting time back to zero.

$$\varphi_{b2} = \varphi_2 - t_0 = 2t_0 - t_{a1} - t_0 = t_0 - t_{a1};$$
$$\varphi_{b3} = \varphi_3 - t_0 = 3t_0 - t_{a1} - t_{a2} - t_0 = 2t_0 - t_{a1} - t_{a2};$$
....
$$\varphi_{bi} = (i-1)t_0 - \sum_{k=1}^{i-1} t_{ak}\ ;\ (\text{for } i \geq 2). \tag{7}$$

Adopting analogous considerations illustrated for the buffer and applying the *Central Limit Theorem* (see Appendix A.2 for details), waiting time can be rewritten as:

$$\varphi_{b\,i} \cong (i-1)\left(t_0 - \overline{t_a}\right) \pm \overline{t_a} \cdot t_{value\,\{\gamma}^{i-2}\sqrt{i-1} =$$
$$= (i-1)\frac{T}{N_P}\left(1 - \frac{N_P}{\overline{N_A}}\right) \pm \frac{T}{\overline{N_A}} \cdot t_{value\,\{\gamma}^{i-2}\sqrt{i-1};\ \text{for } i \geq 3. \tag{8}$$

This discrete values function could be approximated by a time-continuous function recalling that exit time for the *i-th* patient $t_{out\ i}$ is equal to $i \cdot T/N_P$. Confounding waiting time of the *i-th* patient with waiting time of the preceding *(i–1)-th* patient, then, in (8), i could be considered in place of *(i–1)* and for the term $t_{value\ (i-2)} \cdot \sqrt{i-1}$, which is the same reported in Fig. 3, the same interpolation function of i can be adopted. Finally, the time-continuous function $\varphi_b(t)$, representing waiting time for the patient exiting from the system at time $t = i \cdot T/N_P$, can be written as:

$$\varphi_b(t) \cong \left(1 - \frac{N_P}{\overline{N_A}}\right)t \pm \frac{T}{\overline{N_A}} \cdot a\left(\frac{t}{T}N_P\right)^b. \tag{9}$$

with the reported couple of coefficients *(a,b)* for the considered values of γ (*90%* and *95%*) and lower limits $t \geq 5\ T/N_P$ and $t \geq 6\ T/N_P$ respectively.
Considering now the ratio $N_b(t)/\varphi_b(t)$ and taking the same sign for the term due to uncertainty for numerator and denominator one at a time, it's possible to write:

$$\frac{N_b(t)}{\varphi_b(t)} \cong \frac{\overline{N_A}}{T} \frac{\left(\overline{N_A} - N_P\right)\frac{t}{T} \pm \sqrt{\overline{N_A}} \cdot a\left(\frac{t}{T}\right)^b}{\left(\overline{N_A} - N_P\right)\frac{t}{T} \pm \left(N_P\right)^b \cdot a\left(\frac{t}{T}\right)^b}. \tag{10}$$

The multiplication factor of the term $\overline{N_A}/T$, with the above reported values of the coefficients *(a,b)* for *γ=90%*, is reported in Fig. 4 as a function of t (assumed *T=1 [time unit]*) and for various combinations of N_P and $\Delta = \overline{N_A} - N_P$. It's possible to observe that in the deterministic case (deterministic arrivals) $N_b(t)/\varphi_b(t) = \overline{N_A}/T = \lambda$, whereas uncertainty, represented by the reported factor, tends to unity for larger values of t. It assumes relatively low scattered values around 1, except for initial periods, for increasing values of N_P and/or Δ. From the knowledge of the number of patients in the queue at a generic time, taking into account the previous assumptions, then it could be possible to esteem grossly current waiting time of an exiting patient as $N_b(t)/\overline{N_A} \cdot T$.

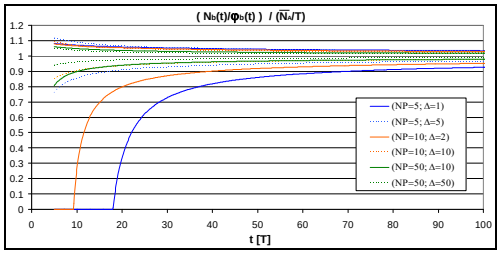

Figure 4: Values of the multiplication factor of $\overline{N_A}/T$ in (10) as a function of t for *γ=90%*.

4.2. Average values

The time-average $\overline{N_{b\,i}}$ of the first i terms N_{bk} *(k=1,2,...,i)* can be calculated assuming, for the sake of simplicity, that each value N_{bk} is maintained over the time interval of the *k-th* period. From the definition of average (see Appendix A.3 for details), it's possible to write:

$$\overline{N_{bi}} \cong \frac{1}{2}(i+1)\left(\overline{N_A} - N_P\right) \pm \sqrt{\overline{N_A}} \cdot \frac{1}{i}\sum_{k=2}^{i} t_{value\,\{\gamma}^{k-1}\sqrt{k};\ \ (\text{for } i \geq 2). \tag{11}$$

In Fig. 5, the right hand term in (11) depending solely on i and *t-value* and due to uncertainty is reported, together with power interpolation functions (dashed lines) in the form $c \cdot i^d$, in the interval *0-500*, for the two confidence levels *90%* and *95%*. In particular, setting lower limits for i so that the percentage difference between the function and its interpolation is lower than *15%* also for initial values of i (then, it rapidly decreases to around *2%* in the considered interval), it's possible to find *(c=1.3993; d=0.4585)* for *γ=90%*, with *i≥8* and *(c=1.802; d=0.4456)* for *γ=95%*, with *i≥14*.

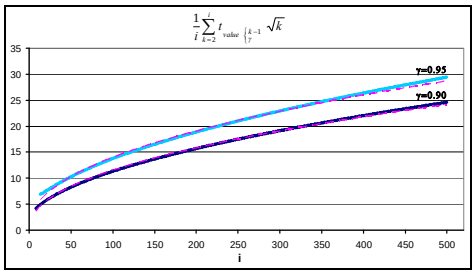

Figure 5: Values of the term depending on i and *t-value* in (11) versus i (for *γ=90%* and *γ=95%*) and respective interpolating functions (dashed).

Finally, considering for $\overline{N_b(t)}$ a time-continuous function, confounding the term *(i+1)* with i, which is equal to t/T, and taking into account the interpolation functions, it's possible to write:

$$\overline{N_b(t)} \cong \frac{1}{2}\left(\overline{N_A} - N_P\right)\frac{t}{T} \pm \sqrt{\overline{N_A}} \cdot c\left(\frac{t}{T}\right)^d; \tag{12}$$

with the shown values of *(c, d)* for γ*=90%* and γ*=95%* and with $t \geq 8T$ and $t \geq 14T$ respectively.
As regards the average value $\overline{\varphi_{bi}}$ of the first i terms φ_{bk} *(k=1,2,...,i)*, it's possible to write (see Appendix A.4 for details):

$$\overline{\varphi_{bi}} \cong \frac{1}{2}(i-1)\frac{T}{N_P}\left(1-\frac{N_P}{\overline{N_A}}\right) \pm \frac{T}{\overline{N_A}} \cdot \frac{1}{i}\sum_{k=3}^{i} t_{value\left\{\gamma\right.^{k-2}}\sqrt{k-1}; \quad (\text{for } \; i \geq 3).$$

Adopting the same previous approximation consisting of considering i in place of *(i-1)*, for the factor depending on *t-values* and i, the same previously found interpolation function can be used[2]. Moreover, considering $\overline{\varphi_b}$ as a time-continuous function:

$$\overline{\varphi_b(t)} \cong \frac{1}{2}\left(1-\frac{N_P}{\overline{N_A}}\right)\frac{T}{N_P}t \pm \frac{T}{\overline{N_A}} \cdot c\left(\frac{t}{T}N_P\right)^d; \qquad (13)$$

with the lower limits $t \geq 8\ T/N_P$ and $t \geq 14\ T/N_P$ for γ*=90%* and γ*=95%*, respectively.
Considering the ratio of the average values:

$$\frac{\overline{N_b(t)}}{\overline{\varphi_b(t)}} \cong \frac{\overline{N_A}}{T}\,\frac{\frac{1}{2}\left(\overline{N_A}-N_P\right)\frac{t}{T} \pm \sqrt{\overline{N_A}} \cdot c\left(\frac{t}{T}\right)^d}{\frac{1}{2}\left(\overline{N_A}-N_P\right)\frac{t}{T} \pm \left(N_P\right)^d \cdot c\left(\frac{t}{T}\right)^d} \qquad (14)$$

The multiplication factor of the term $\overline{N_A}/T$, with the above reported values of coefficients *(c,d)* for γ*=90%*, is reported in Fig. 6 for various combinations of N_P and $\Delta = \overline{N_A} - N_P$ as a function of t. As observable, also the ratio of average values tends to $\overline{N_A}/T = \lambda$, for increasing t and the same considerations for the ratio of instantaneous values apply. Of course, this result derives from the linearity with t of the part of the functions $N_b(t)$ and $\varphi_b(t)$ not related to uncertainty, which makes also their average values, apart from the second term, proportional to *1/2 t*.

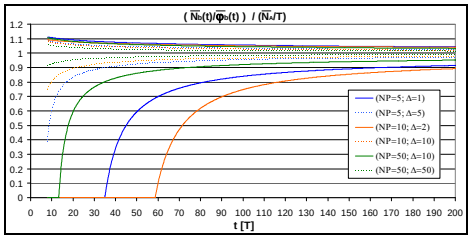

Figure 6: Values of the multiplication factor of $\overline{N_A}/T$ in (14) as a function of t for γ*=90%*.

5. CASE WITH DECREASING BUFFER

In this section, the opposite unsteady-state situation with $\mu > \lambda$ (that's to say $N_P > \overline{N_A}$) is discussed, starting with a buffer containing a backlog of N_{b0} patients. Of course, once the backlog has been cancelled, the system will be in the condition illustrated in Section 3, with fluctuating instantaneous values, but constant limit values (for $t \to \infty$) for averages. Therefore the following analysis is limited to the period during which the buffer level remains above zero.

5.1. Instantaneous values

With the same meaning of the above defined quantities:

$N_{b1} = max\ \{0,\ N_{b0} - (N_P - N_{A1})\}$;
$N_{b2} = max\ \{0,\ N_{b1} - (N_P - N_{A2})\}$;
....
$N_{bi} = max\ \{0,\ N_{b\,(i-1)} - (N_P - N_{Ai})\}$;

under the assumption that the server is always busy. Moreover, as far as the second term in the brackets isn't negative, the recursive expression N_{bi} can be written as:

$$N_{bi} = N_{b0} - iN_P + \sum_{k=1}^{i} N_{A\,k} \quad \left(\text{for } N_{bi} \geq 0 \Rightarrow i \leq \left(N_{b0} + \sum_{k=1}^{i} N_{A\,k}\right) \Big/ N_P\right).$$

Adopting the same above considerations:

$$N_{bi} \cong N_{b0} - i\left(N_P - \overline{N_A}\right) \pm \sigma_{POP} \cdot t_{value\left\{\gamma\right.^{i-1}}\sqrt{i}; \quad \left(\text{for } \; 2 \leq i \leq \left(N_{b0} + \sum_{k=1}^{i} N_{A\,k}\right) \Big/ N_P\right).$$

The function $t_{value\ (i-1)} \cdot \sqrt{i}$ could be interpolated by the same power function $a \cdot i^b$, already shown, for given values of γ. Finally, stated that $t = i \cdot T$ and for Poisson arrivals $\sigma_{POP} = \sqrt{\overline{N_A}}$, considering N_b as a time-continuous function:

$$N_b(t) \cong N_{b0} - \left(N_P - \overline{N_A}\right)\frac{t}{T} \pm \sqrt{\overline{N_A}} \cdot a\left(\frac{t}{T}\right)^b. \qquad (15)$$

Since it's relevant to find, in an immediate way, values of i (or of t) for which $N_b = 0$ and observing that the coefficient b is around 0.5, the slightly different interpolation function $a' \cdot i^{0.5}$ can be adopted (with neglectable differences in the same interval)[3]. In particular, $a' = 0.675$ for γ*=50%*, with $i \geq 4$, $a' = 1.65$ for γ*=90%*, with $i \geq 8$, and $a' = 1.97$ for γ*=95%*, with $i \geq 7$ (percentage difference between the function and its interpolation lower than *15%*). Therefore, it's possible to find, under the illustrated assumptions, the roots of the equation representing the time interval extremes (for a given confidence level) within which the buffer is expected to become empty:

$$t \cong T \cdot \left[\frac{N_{b0}}{N_P - \overline{N_A}} + \frac{1}{2}\frac{a'^2\,\overline{N_A}}{\left(N_P - \overline{N_A}\right)^2}\left(1 \pm \sqrt{1 + \frac{4N_{b0}\left(N_P - \overline{N_A}\right)}{a'^2\,\overline{N_A}}}\right)\right]. \qquad (16)$$

In Fig. 7, two simple examples of $N_b(t)$ are reported, both starting from $N_{b0} = 200$ and with $\overline{N_A} = 38$, for two

[2] In detail, $\frac{1}{i}\sum_{k=3}^{i} t_{value\left\{\gamma\right.^{k-2}}\sqrt{k-1} = \frac{1}{i}\sum_{j=2}^{i-1} t_{value\left\{\gamma\right.^{j-1}}\sqrt{j}$
and *(i–1)*, upper limit of the summation term, is substituted by *i*.

[3] This means assuming a constant value of $t_{value\ (i-1)}$ for a given γ.

possible values of N_P (40 and 45). These last parameters are realistic values, implemented in the successive simulation model.

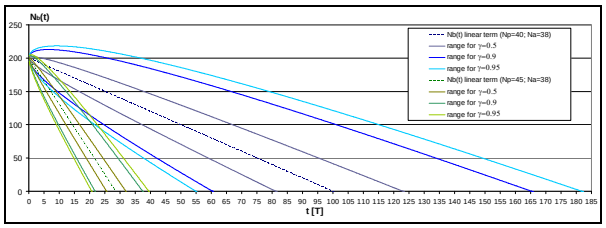

Figure 7: $N_b(t)$ and Time Limits from (16) for $N_b(t)=0$ for different γ ($N_{b0}=200$, $\overline{N_A}=38$, N_P=40 and N_P=45)

It's possible to observe how, for each alternative, the limit values of the interval aren't symmetric around the zero for the linear term of (15) (dashed lines), but are shifted to the right. Moreover, comparing the two alternatives, the range of the uncertainty interval, with the same γ, is lower for greater values of N_P. Therefore, if the objective of management is the abatement of the waiting list, doing so in a shorter period of time could be preferable also for the reduced uncertainty.
As regards waiting time in the buffer, with the same assumptions of the case with an increasing buffer (the server is always busy and patients are served on a FCFS basis), exit time $t_{out\ i}$ is deterministic and equal to $(N_{b0}+i)\cdot t_0$. Also in this case, $t_{in\ i}$ can be written as the sum of the preceding *(i-1)* inter-arrival times t_{ak}, for $i\geq 2$, and therefore:

$$\varphi_{b1} = \varphi_1 - t_0 = t_{out1} - t_{in1} - t_0 = N_{b0}\cdot t_0 ;$$
$$\varphi_{b2} = N_{b0}\cdot t_0 + t_0 - t_{a1} ;$$
....
$$\varphi_{bi} = N_{b0}t_0 + (i-1)t_0 - \sum_{k=1}^{i-1} t_{a\,k}; \quad (\text{for } i \geq 2).$$

Recalling (20) in Appendix A.2:

$$\varphi_{bi} \cong N_{b0}\frac{T}{N_P} + (i-1)\frac{T}{N_P}\left(1-\frac{N_P}{\overline{N_A}}\right) \pm \frac{T}{\overline{N_A}}\cdot t_{\,value\,\{\gamma}^{i-2}\sqrt{i-1}; \quad (\text{for } i \geq 3).$$

Confounding waiting time of the *i-th* patient with waiting time of the *(i– 1)-th* patient, then *i* could be used in place of *(i–1)* and for the term linked to uncertainty the same previous interpolating function can be adopted. Considering φ_b as a time-continuous function and deriving *i* as a function of exit time *t* as $(t/t_0 - N_{b0})$, it's possible to write:

$$\varphi_b(t) \cong N_{b0}\frac{T}{\overline{N_A}} - \left(\frac{N_P}{\overline{N_A}}-1\right)t \pm \frac{T}{\overline{N_A}}\cdot a\left(\frac{t}{T}N_P - N_{b0}\right)^b.$$

with the lower limits $t \geq (N_{b0}+5)\cdot t_0$ and $t \geq (N_{b0}+6)\cdot t_0$ for γ=90% and γ=95% respectively.
Finally, taking into consideration the ratio $N_b(t)/\varphi_b(t)$:

$$\frac{N_b(t)}{\varphi_b(t)} \cong \frac{\overline{N_A}}{T}\,\frac{N_{b0} - \left(N_P - \overline{N_A}\right)\frac{t}{T} \pm \sqrt{\overline{N_A}}\cdot a\left(\frac{t}{T}\right)^b}{N_{b0} - \left(N_P - \overline{N_A}\right)\frac{t}{T} \pm a\left(\frac{t}{T}N_P - N_{b0}\right)^b} \qquad (17)$$

The multiplication factor of the term $\overline{N_A}/T$, with the above reported values of the coefficients *(a,b)* for γ=90%, is reported in Fig. 8 as a function of *t* (*T=1 [time unit]*), for various combinations of $\overline{N_A}$, $\Delta = N_P - \overline{N_A}$ and different values of N_{b0}, set in such a way that the liner term of (15) is zero at *t=200*.

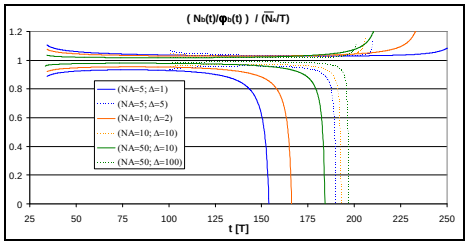

Figure 8: Values of the multiplication factor of $\overline{N_A}/T$ in (17) as a function of *t* for γ=90%.

It's possible to observe that in the deterministic case (deterministic arrivals) $N_b(t)/\varphi_b(t) = \overline{N_A}/T = \lambda$, whereas uncertainty, represented by the reported factor, is around 1 in the previous period approaching the empty buffer condition.

5.2. Average values

For the average value of the buffer, it's possible to write:

$$\overline{N_{b\,i}} \cong N_{b0} - \frac{1}{2}(i+1)\left(N_P - \overline{N_A}\right) \pm \sqrt{\overline{N_A}}\cdot\frac{1}{i}\sum_{k=2}^{i} t_{\,value\,\{\gamma}^{k-1}\sqrt{k}; \quad (\text{for } i \geq 2).$$

With the same previous assumptions and interpolation function for the term depending on *t-values* and *i* and considering a time-continuous function:

$$\overline{N_b(t)} \cong N_{b0} - \frac{1}{2}\left(N_P - \overline{N_A}\right)\frac{t}{T} \pm \sqrt{\overline{N_A}}\cdot c\left(\frac{t}{T}\right)^d.$$

As regards the average value $\overline{\varphi_{bi}}$ of the first *i* terms φ_{bk} *(k=1,2,...,i)*, proceeding analogously as before:

$$\overline{\varphi_{bi}} \cong N_{b0}\frac{T}{N_P} - \frac{1}{2}(i-1)\frac{T}{N_P}\left(\frac{N_P}{\overline{N_A}}-1\right) \pm \frac{T}{\overline{N_A}}\cdot\frac{1}{i}\sum_{k=3}^{i} t_{\,value\,\{\gamma}^{k-2}\sqrt{k-1}; \quad (\text{for } i \geq 3).$$

With the same previous assumptions, considering a time-continuous function:

$$\overline{\varphi_b(t)} \cong \frac{1}{2}N_{b0}T\left(\frac{1}{N_P}+\frac{1}{\overline{N_A}}\right) - \frac{1}{2}\left(\frac{N_P}{\overline{N_A}}-1\right)t \pm \frac{T}{\overline{N_A}}\cdot c\left(\frac{t}{T}N_P - N_{b0}\right)^d.$$

with the lower limits $t \geq (N_{b0}+8)\cdot T/N_P$ and $t \geq (N_{b0}+14)\cdot T/N_P$, for γ=90% and γ=95% respectively.

Considering the ratio of the two above quantities, we have:

$$\frac{\overline{N_b(t)}}{\overline{\varphi_b(t)}} \cong \frac{\overline{N_A}}{T} \frac{N_{b0} - \frac{1}{2}\left(N_P - \overline{N_A}\right)\frac{t}{T} \pm \sqrt{\overline{N_A}} \cdot c\left(\frac{t}{T}\right)^d}{\frac{1}{2}N_{b0}\left(1+\frac{\overline{N_A}}{N_P}\right) - \frac{1}{2}\left(N_P - \overline{N_A}\right)\frac{t}{T} \pm c\left(\frac{t}{T}N_P - N_{b0}\right)^d}. \quad (18)$$

Numerically, it's possible to verify that the multiplication factor of $\overline{N_A}/T$ is always greater than one (as shown in Fig. 9 for γ=90%) and therefore an upper limit for the instantaneous average waiting time in the buffer $\overline{\varphi_b(t)}$ can be assumed as $\left(\overline{N_b(t)}/\overline{N_A}\right)\cdot T$.

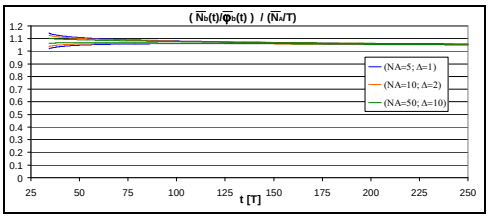

Figure 9: Values of the multiplication factor of $\overline{N_A}/T$ in (18) as a function of t for γ=90%.

6. SIMULATION MODELS

In this section two simple DES models, coded with the process algebra language χ (Hofkamp and Rooda 2007), are presented and applied to real cases of a CT scan facility inside a public hospital.

Outpatient access to the diagnostic service (one multi-slice CT scanner) is managed during two six hour work shifts from 8:00 a.m. to 8:00 p.m., from Monday to Friday. During this weekly period T, normally 8 appointments each day are scheduled. The first model reproduces exactly an M/D/1 system and a deterministic machine, with a fictitious service time t_0, is implemented in order to simulate a scheduling process with N_P served patients for each period T ($t_0=T/N_P$). Therefore, in such a model, detailed information about the really adopted scheduling plan (i.e. information regarding appointment times) is not needed, with the assumptions already illustrated in Section 3, since a deterministic service time is equivalent to assuming the scheduled patients' access uniformly distributed during the time period. Two cases are examined, respectively with N_P=40 and, adding one further slot per day, N_P'=45 weekly scheduled patients, starting, in both situations, with an assumed initial buffer N_{b0}=200 patients and an hypothesized value of $\overline{N_A}=38$ weekly average demand.

Attention is focused on the decreasing buffer case because it meets more practical interest than the opposite one. The aim is verifying that the resulting range of time necessary to empty the buffer is within the values derived from (16). Successively, once the backlog is over and the system has reached steady-state conditions (since $\overline{N_A}<N_P$), average waiting time is tracked in order to verify relation (1). It should be remarked that a huge backlog could have been caused, during preceding unbalanced periods, by greater demand and/or less planned appointments. Moreover, a generic buffer level N_{b0} implies that the first appointment slot for a patient entering the system at that time will be available in $N_{b0}/N_P \cdot T$ periods (value of φ_{b1}). If this level were a constant, with N_{b0}=200, waiting time would always be 5 (with N_P=40) or 4.4 (with N_P'=45) weeks; with N_{b0}=100 it would be reduced to 2.5 or 2.2 weeks, which can be considered generally acceptable. Therefore, also a partial drop down of N_{bo} from 200 to 100 patients could represent a managerial objective (in (16) the difference of the two values should be used in place of N_{bo} to esteem the needed time). Successively, under the hypothesis $\overline{N_A}$ is not variable, N_P should vary in order to keep the buffer level fluctuating around the desired level; this could be advantageous for freeing resource time for other purposes, fixing periods with less appointments then normal (and, correspondingly, buffer increasing).

The second model, reported in Fig. 10, is a simplified adaptation of a more complete and detailed model of a CT scan facility open to diverse patient flows, described in (Boenzi et al. 2013), in which the scheduling process of outpatients is simulated by means of a controlled release mechanism from the buffer.

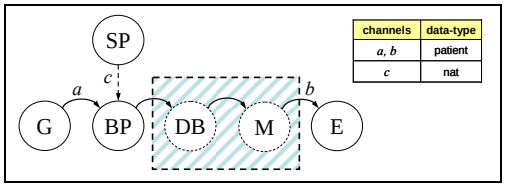

Figure 10: Complete Model Processes, truncated portion (hatch box) and resulting Simplified Model

In the simplified examined case, the illustrated general framework refers to one buffer only (for a specific class of outpatients). Moreover, for the aims of this paper, a truncated model can be considered. Patients are generated by process *G* (adopting a *HPP*) and stored in *BP*. Their access to the service is regulated by means of a scheduling process *SP*, which repeats a daily cycle, according to the particular schedule in place. *SP* controls *BP* requesting the release of established numbers of patients at particular time instants to a daily working-list buffer *DB*. From process *DB* (in which, normally, priority rules are adopted to manage diverse patient flows) a patient can access service *M*, when it is not busy (representing a common shared resource in the complete model). Finally, patients leave the system in process *E*. However, since we are interested, in terms of simulation results, only in waiting time in *BP* (time interval from the examination request to the appointment day) and not in waiting time in *DB* (at the scan facility), the processes involved beyond *BP* are out of the scope of the model. For this reason *BP* can be linked directly to the exit process *E*, in which data are collected.

6.1. Simulation results for the first model

The unity of measurement of time is the week (*T =1 week*) and, in the following figures, results for two groups of 10 simulation runs (with the mentioned parameter values and the two values of N_P, shown in different colors) are reported. In Fig. 11, the time-history of the instantaneous buffer level is reported, terminated when the buffer becomes empty.

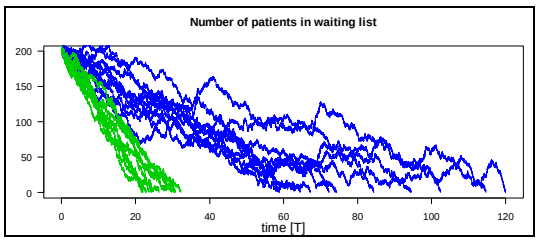

Figure 11: Buffer Level time-history for ten simulation runs for N_P=40 (blue) and N_P=45 (green)

In particular, Fig. 12 shows in detail how the resulting values of time are within the expected confidence interval limits calculated from (16) for the respective case, reported as box edges. Finally, in Fig. 13, the time-history of the average waiting time is reported for *10000* weeks. It can be observed that, after a transition period, the value stabilizes around the limiting values $\cong 0.25\ T$ and $\cong 0.07\ T$ for $N_P=40$ and $N_P=45$ respectively, in agreement with (1).

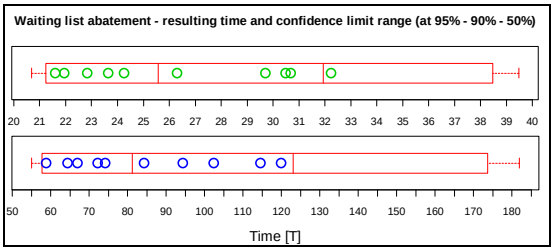

Figure 12: Time for which $N_b=0$ in ten simulation runs, (N_P=40: blue, N_P=45: green) compared to interval limits derived from (16) for γ=50%-90%-95%

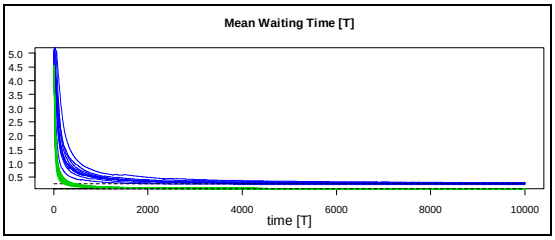

Figure 13: Time-history of the Mean Waiting Time in the Buffer for ten simulation runs for N_P=40 (blue) and N_P=45 (green)

6.2. Simulation results for the second model

In the second model, two slightly different appointment schedules have been specified in more detail and implemented. In general, at the facility, outpatients are admitted one at a time (each patient is assigned a specific slot). The time interval between two consecutive appointments is set at one hour in the morning (because of the presence of a more intense internal service request in the hospital), starting at 8:00, and half a hour in the afternoon, starting at 15:00. Maintaining the same weekly number of appointments N_P adopted for the first model, the total number of daily patients in the first schedule is equally divided as four patients in the morning shift and four patients in the afternoon shift (last appointment at 16:30). In the second schedule, an additional appointment has been set at 17:00. It's to remark that the feasibility of this new plan, in particular with respect to the risk of over-time, could be assessed by a more detailed simulation model, but is beyond the scope of the present work. The unity of measurement of time is one hour, but results are presented in weeks (*T =1 week = 60 hours* in the simulation). The following Figures from 14 to 16 are analogous to those referring to the first model and the same considerations could be made. A slight difference between the two models can be found with regard to average waiting time in the buffer. In fact, considering each group of ten simulations, time terminated after *10000* weeks, this value is $\cong 0.283\ T$ when $N_P=40$ and $\cong 0.087\ T$ when $N_P=45$. This is due to the different distribution of patient appointments, which is not perfectly uniform in the course of time in the second model.

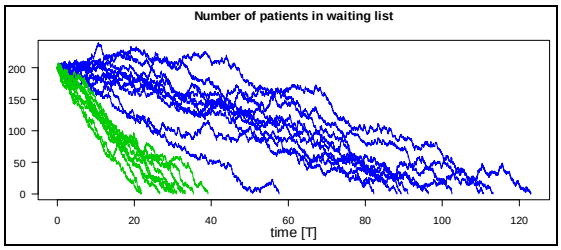

Figure 14: Buffer Level time-history for ten simulation runs for N_P=40 (blue) and N_P=45 (green)

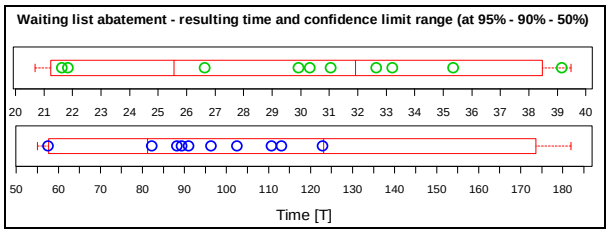

Figure 15: Time for which $N_b=0$ in ten simulation runs, (N_P=40: blue, N_P=45: green) compared to interval limits derived from (16) for γ=50%-90%-95%

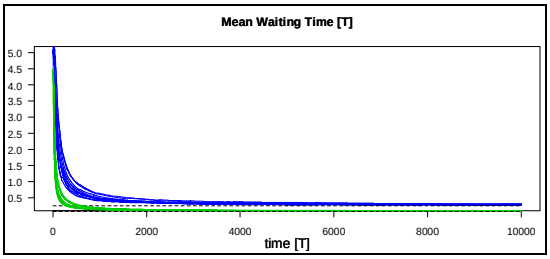

Figure 16: Time-history of the Mean Waiting Time in the Buffer for ten simulation runs for N_P=40 (blue) and N_P=45 (green)

Finally, in the model adopting the first schedule, a buffer control mechanism has been introduced, through the monitoring of the current waiting list length. The aim is maintaining the buffer at a constant level, which can be beneficial both to the facility staff (in order to manage possible cancellations with other appointments) and to the customers, who can plan their examinations well in advance (sometimes this is unavoidable because of additional examinations and/or preparation). In Fig. 17, the time-history (terminated at *210* weeks) of the instantaneous number of patients in the buffer and of current waiting time (for the patient currently exiting from the system) for three simulation runs (in different colors) are reported, starting with $N_{bo}=200$ patients.

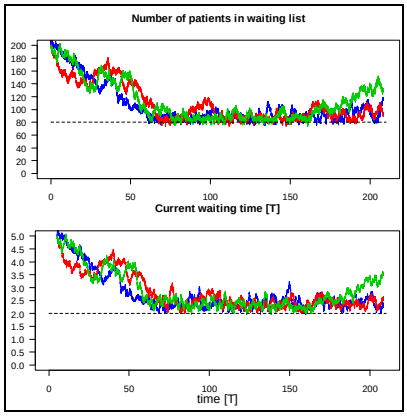

Figure 17: Time-history of Current Buffer Level (above) and Current Waiting Time (below) for three simulation runs for N_P=40 adopting Buffer Control

The lower limit for the buffer has been set at 80 patients because this implies a lower waiting time limit equal to 2 weeks (assuming $N_P=40$ patients/week), which can be considered a good compromise. In the simulation model, the buffer level control is done at the end of each day and, if the level is below the set limit, no appointments are scheduled for the subsequent day. Of course this is a simplification, being unfeasible in a real appointment system. In the last, a day without appointments could be set within the already booked period. However a problem is that as this time is set far in the future, as the lag-time between the signal and the corrective action increases. Anyway, this day could be better utilized for other patient flows (for example, additional appointments for a different booking agenda could be placed). Solutions different from control reveal to be ineffective. For example, implementing a cyclic alternate schedule, made of 19 days with 8 appointments per day and one day without appointments, would realize, for the assumed value $\overline{N_A}=38$ patients/week, the exact balance between demand and offer in a four week period. However, since this cycle is completely unrelated to variability of the demand, the system state variables are highly fluctuating. Therefore, some type of control loop in these systems should be enforced. Once adopted, demand variability could be turned into an opportunity, with the appropriate flexibility in the work organization.

7. CONCLUSION

In the present work, some common features of an appointment system in general have been highlighted, by means of an analytical approach for non-steady state conditions; successively, applications to a real case of a CT scan facility have been presented. It can be observed that, at a macroscopic level, the main factors characterizing the system are on the one side the appointment demand and, on the other side, the offer of planned appointments.

At this level, details regarding the acceptability of a scheduling plan in terms of overtime and/or excessive waiting time experimented at the facility can be put apart and analyzed successively. The only source of variability is represented by the appointment demand, since setting the number of appointments in a period of time T means specifying the system "productive capability". Therefore, it can be treated as a G/D/1 queue. In particular, in the M/D/1 case, from Kingman's equation it's possible to observe that waiting time in the buffer can't exceed one period T, as far as the difference between the set number of appointments and the expected number of requests in T remains above 0.5. Therefore, in such systems, the existence of long waiting time for appointments can only be explained by current demand greater then offer (increasing buffer and waiting time) or by former relevant backlog, created in preceding unbalanced periods. An analytical approach has been proposed in order to study the two situations represented by increasing buffer and decreasing buffer with initial backlog. In particular, in the latter case, statistical considerations permit to define confidence intervals for estimation of time necessary to clear the backlog, which could be useful in managerial decisions. It should be noticed that also an intermediate solution, aiming at maintaining a set number of patients in the waiting list, could be an objective. However, it would require a very strict control of the buffer and of the demand. Finally, two simulation models, adopting parameter values of a CT scan facility, have been illustrated. Results obtained from the two models are similar and show, on the one hand, agreement with the predicted statistical intervals; on the other hand, results show that the M/D/1 model can well describe the behavior of the buffer as with a more refined model. Future work will be focused on the detailed investigation, by means of simulation, of the appointment schedule in place at the CT facility inside a public hospital, differentiated for each day of the week, in order to verify, in particular, the impact of different alternatives with increased numbers of appointments on the probability of overtime.

APPENDIX A

A.1 The summation term of (4) can be thought of as the sample average $\overline{N_{Ai}}$ of the first i arrivals multiplied by i. Applying the *Central Limit Theorem*, this sample average can be approximated by the population average $\overline{N_A}$, with a specified confidence level γ, expressed by the amplitude of a variability range, which is the

product of the critical *t-value* of the *t-probability distribution,* for γ and *(i-1)* degrees of freedom, and $\sigma_{POP}/\sqrt{i}$, being σ_{POP} the population standard deviation:

$$\sum_{k=1}^{i} N_{Ak} = i\,\overline{N_{Ai}} \cong i\left(\overline{N_A} \pm t_{value}\left\{{}^{i-1}_{\gamma}\right\}\frac{\sigma_{POP}}{\sqrt{i}}\right) = \tag{19}$$

$$= i\,\overline{N_A} \pm \sigma_{POP} \cdot t_{value}\left\{{}^{i-1}_{\gamma}\right\}\sqrt{i};\quad (\text{for } i \geq 2).$$

In particular, when arrivals belong to a Poisson distribution, $\sigma_{POP} = \sqrt{\overline{N_A}}$.

A.2 For the summation term of (7), applying the *Central Limit Theorem*, the sample average of the first *(i–1)* terms can be approximated by the population average $\overline{t_a}$:

$$\sum_{k=1}^{i-1} t_{ak} \cong (i-1)\left(\overline{t_a} \pm t_{value}\left\{{}^{i-2}_{\gamma}\right\}\frac{\sigma_{POP}}{\sqrt{i-1}}\right) = \tag{20}$$

$$= (i-1)\overline{t_a} \pm \sigma_{POP} \cdot t_{value}\left\{{}^{i-2}_{\gamma}\right\}\sqrt{i-1};\ (\text{for } i \geq 3).$$

In particular, when the number of requests follows a Poisson distribution, inter-arrival times belong to an exponential distribution with $\overline{t_a} = 1/\lambda = T/\overline{N_A}$ and $\sigma_{POP} = \overline{t_a} = 1/\lambda = T/\overline{N_A}$.

A.3 From (4) and (19):

$$\overline{N_{bi}} = \frac{1}{i}\left(N_{b1} + N_{b2} + \ldots + N_{bi}\right) = \frac{1}{i}\sum_{k=1}^{i} N_{bk} = \frac{1}{i}\sum_{k=1}^{i}\left(\sum_{l=1}^{k} N_{Al} - kN_P\right) =$$

$$= \frac{1}{i}\left(\sum_{k=1}^{i}\sum_{l=1}^{k} N_{Al} - N_P\sum_{k=1}^{i} k\right) = \frac{1}{i}\left[\sum_{k=1}^{i}\sum_{l=1}^{k} N_{Al} - \frac{i}{2}(i+1)N_P\right]$$

$$\sum_{l=1}^{k} N_{Al} = k\,\overline{N_{Ak}} \cong k\overline{N_A} \pm \sigma_{POP} \cdot t_{value}\left\{{}^{k-1}_{\gamma}\right\}\sqrt{k}\quad (\text{with } k \geq 2);$$

$$\sum_{k=1}^{i}\sum_{l=1}^{k} N_{Al} = N_{A1} + \sum_{k=2}^{i}\sum_{l=1}^{k} N_{Al} \cong N_{A1} + \sum_{k=2}^{i}\left(k\overline{N_A} \pm \sigma_{POP} \cdot t_{value}\left\{{}^{k-1}_{\gamma}\right\}\sqrt{k}\right) + \overline{N_A} - \overline{N_A} =$$

$$= N_{A1} - \overline{N_A} + \sum_{k=1}^{i} k\,\overline{N_A} \pm \sum_{k=2}^{i}\sigma_{POP} \cdot t_{value}\left\{{}^{k-1}_{\gamma}\right\}\sqrt{k} = N_{A1} - \overline{N_A} + \frac{i}{2}(i+1)\overline{N_A} \pm\ \sigma_{POP} \cdot \sum_{k=2}^{i} t_{value}\left\{{}^{k-1}_{\gamma}\right\}\sqrt{k};$$

$$\overline{N_{bi}} \cong \frac{1}{2}(i+1)\left(\overline{N_A} - N_P\right) + \frac{1}{i}\left(N_{A1} - \overline{N_A}\right) \pm \frac{1}{i}\sigma_{POP} \cdot \sum_{k=2}^{i} t_{value}\left\{{}^{k-1}_{\gamma}\right\}\sqrt{k};\ (\text{for } i \geq 2).$$

The term $\left(N_{A1} - \overline{N_A}\right)/i$ can be neglected and, assuming a Poisson arrival process, $\sigma_{POP} = \sqrt{\overline{N_A}}$.

A.4 From (7) and (20):

$$\overline{\varphi_{bi}} = \frac{1}{i}\left(\varphi_{b1} + \varphi_{b2} + \ldots + \varphi_{bi}\right) = \frac{1}{i}\sum_{k=1}^{i}\varphi_{bk} = \frac{1}{i}\sum_{k=2}^{i}\left((k-1)t_0 - \sum_{l=1}^{k-1} t_{al}\right) =$$

$$= \frac{1}{i}\left(t_0\sum_{k=2}^{i}(k-1) - \sum_{k=2}^{i}\sum_{l=1}^{k-1} t_{al}\right) = \frac{1}{i}\left[i\frac{i-1}{2}t_0 - \sum_{k=2}^{i}\sum_{l=1}^{k-1} t_{al}\right]$$

$$\sum_{l=1}^{k-1} t_{al} \cong (k-1)\overline{t_a} \pm \sigma_{POP} \cdot t_{value}\left\{{}^{k-2}_{\gamma}\right\}\sqrt{k-1}\quad (\text{with } k \geq 3);$$

$$\sum_{k=2}^{i}\sum_{l=1}^{k-1} t_{al} = t_{a1} + \sum_{k=3}^{i}\sum_{l=1}^{k-1} t_{al} \cong t_{a1} + \sum_{k=3}^{i}\left((k-1)\overline{t_a} \pm \sigma_{POP} \cdot t_{value}\left\{{}^{k-2}_{\gamma}\right\}\sqrt{k-1}\right) + \overline{t_a} - \overline{t_a} =$$

$$= t_{a1} - \overline{t_a} + \sum_{k=2}^{i}(k-1)\overline{t_a} \pm \sum_{k=3}^{i}\sigma_{POP} \cdot t_{value}\left\{{}^{k-2}_{\gamma}\right\}\sqrt{k-1} = t_{a1} - \overline{t_a} + i\frac{i-1}{2}\overline{t_a} \pm\ \sigma_{POP} \cdot \sum_{k=3}^{i} t_{value}\left\{{}^{k-2}_{\gamma}\right\}\sqrt{k-1};$$

$$\overline{\varphi_{bi}} \cong \frac{1}{2}(i-1)\left(t_0 - \overline{t_a}\right) - \frac{1}{i}\left(t_{a1} - \overline{t_a}\right) \pm \frac{1}{i}\sigma_{POP} \cdot \sum_{k=3}^{i} t_{value}\left\{{}^{k-2}_{\gamma}\right\}\sqrt{k-1}\quad (\text{with } i \geq 3);$$

$$\overline{\varphi_{bi}} \cong \frac{1}{2}(i-1)\left(t_0 - \overline{t_a}\right) - \frac{1}{i}\left(t_{a1} - \overline{t_a}\right) \pm \frac{1}{i}\sigma_{POP} \cdot \sum_{k=3}^{i} t_{value}\left\{{}^{k-2}_{\gamma}\right\}\sqrt{k-1};\ (\text{for } i \geq 3).$$

As done similarly for the buffer expression, the term $\left(t_{a1} - \overline{t_a}\right)/i$ can be neglected and, assuming a Poisson arrival process, $\sigma_{POP} = T/\overline{N_A}$.

REFERENCES

Adan, I., Resing, J., 2002. *Queueing Theory.* Eindhoven, The Netherlands: Department of Mathematics and Computing Science – Eindhoven University of Technology.

Boenzi, F., Mummolo, G., Rooda, J.E., 2013. Appointment scheduling for a computed tomography facility for different patient classes using simulation. *Proceedings of the 25th European Modeling and Simulation Symposium*, pp. 370-378, September 25-27, Athens, Greece

Green, L.V., Savin, S.V., Wang, B., 2006. Managing patient service in a diagnostic medical facility. *Oper Res* 54(1): 11–25.

Hans, E.W., Van Houdenhoven, M., Hulshof, P.J.H., 2011. *A Framework for Health Care Planning and Control*. Memorandum 1938. ISSN 1874-4850. Department of applied mathematics, University of Twente, Enschede, The Netherlands. Available from: http://www.math.utwente.nl/publications [Accessed 15 July 2014].

Hassin, R., Mendel, S., 2008. Scheduling Arrivals to Queues: A Single-Server Model with No-Shows. *Management Science* 54(3): 565–572.

Hofkamp, A.T., Rooda, J.E., 2007. C*hi 1.0 Reference Manual*. Eindhoven University of Technology. Available from: http://seweb.se.wtb.tue.nl/sewiki/_media/chi/chi10_refman1689.pdf [Accessed 15 July 2014].

Hopp, W.J., 2008. Chapter 4: Single Server Queueing Models. In: Chhajed, D., Lowe, T.J., eds. *Building Intuition - Insights from Basic Operations Management Models and Principles.* Springer International Series in Operations Research & Management Science Volume 115, pp. 51-79.

Hutzschenreuter, A., 2004. *Waiting Patiently - An analysis of the performance aspects of outpatient scheduling in health care institutes*. BMI – Paper, Vrije Universiteit Amsterdam, The Netherlands. Available from: http://www.few.vu.nl/en/Images/werkstuk-hutzschenreuter_tcm39-91363.pdf [Accessed 15 July 2014].

Kaandorop, G., Koole, G., 2007. Optimal outpatient appointment scheduling. *Health Care Manag Sci* 10: 217–229.

Klassen, K.J., Rohleder, T.R., 1996. Scheduling outpatient appointments in a dynamic environment. *Journal of Operations Management* 14: 83–101.

Kolisch, R., Sickinger, S., 2008. Providing radiology health care services to stochastic demand of

different customer classes. *OR Spectrum* 30(2): 375–395.

Kuiper, A., 2012. *Appointment scheduling in healthcare*. Master Thesis. IBIS UvA – Faculty of Economics and Business, Faculty of Science – University of Amsterdam.

Patrick, J., 2012. A Markov decision model for determining optimal outpatient scheduling. *Health Care Manag Sci* 15: 91–102.

Rohleder, T., Klassen, K.J., 2002. Rolling Horizon Appointment Scheduling: A Simulation Study. *Health Care Manag Sci* 5: 201–209.

Sickinger, S., Kolisch, R., 2009. The performance of a generalized Bailey–Welch rule for outpatient appointment scheduling under inpatient and emergency demand. *Health Care Manag Sci* 12: 408–419.

Wijewickrama, A.K.A., 2006. Simulation analysis for reducing queues in mixed-patients'outpatient department. *Int. j. simul. model.* 05 2:56-68.

Zhu, Z. C., Heng, B. H., Teow, K. L., 2009. Simulation study of the optimal appointment number for outpatient clinics. *Int. j. simul. model.* 08 3:156-165.