Analytical Modeling of an Accident & Emergency Department

Kevin Duijndam

Preface

This paper is part of the Master program Business Mathematics & Informatics. The objective of the paper is to investigate a business problem and present the outcomes of this investigation both on paper, as well as during an oral presentation.

I would like to thank René Bekker for his help with this research and for his comments on this paper. Also I would like to thank Annemieke van Dongen and Joost van Galen for providing the data on time spent at the accident and emergency department of the VUmc.

Abstract

In this paper an accident & emergency department is modeled as a network of queues. Then, using standard queueing theory in combination with optimization techniques to minimize differences between actual and predicted outcomes, it becomes possible to calculate various characteristics of the model like average waiting time per station and sojourn time. The big advantage of this approach is that it doesn't require knowledge of the service time distribution at every station, only the total sojourn time per customer and the stations this customer has visited are required to compute an estimate of the average service time per station.

It is shown that this approach works very well when only taking the average sojourn time for the station into account. When the entire distribution of sojourn times is taken into account, it is shown that this approach works well too. The model is only not able to predict outliers accurately, but this is not something that should be expected from such a model.

Contents

Modeling accident & emergency departments

The accident & emergency department of a hospital is in general a place most people do not like to be. And might they ever need to go there, the shorter the better. And besides this point, safety and security are also important factors why there has been a lot of attention in studying the characteristics of an A&E department. A key characteristic in this case is then the time people spend at an A&E department. However, analyzing such a department is fairly complicated because a lot of uncertainty is involved and the processes are quite complex. Think for instance of the arrival process for which it is difficult to predict how many people exactly are going to arrive in a given hour. Then there is also the time after someone has entered the emergency department, where there are a lot of different types of care that a patient could need which all take stochastic time. The type of processes a patient needs is most often only known after some tests are done. Another factor is the use of resources outside of the A&E department for which it is hard to predict whether they are available or not.

To be able to analyze and A&E department regardless, often a simulation based approach is used, see for instance Flether et al. (2007) or Kolker (2008). This is convenient, because it's possible to model virtually anything and then a simulation can be run until the required precision is achieved. A big advantage of this is the large flexibility it offers; almost any probability distribution for waiting times can be implemented and also complicated dependence between various stations can be taken into account. A major issue in a simulation-based approach is that a large amount of data is needed before good estimates of distributions per station or process can be made. This data is generally not available and also difficult to gather. Another big drawback of this approach is that it can take quite a lot of time to develop a model that's suitable for a certain hospital. Moreover, computation time can also be quite large, which is not convenient if one is interested in a lot of different

scenarios. From a mathematical point of view, analytical expressions are preferred over the outcomes of a simulation.

Therefore, in this paper another approach is used. Instead of using a simulation based approach, models from standard queueing theory are used to model an A&E department. Then, under various assumptions, it becomes possible to instantly compute the characteristics of the waiting time for various patients and, in this way, analyze the system. The big advantage of closed-form expressions that arise when using queueing models is that it becomes possible to estimate the parameters for the assumed distributions, so this is not required to be known beforehand. Some studies in this direction have been done, see for instance Mayhew, Smith (2008). However, Mayhew and Smith use models that are a lot simpler than reality, so it is still not clear how much time patients spent at various stages. To have some input for the model, data from the emergency department of the VU medical center was used.

The processes in the department

The emergency department consists of a lot of different processes, which are not all used by every patient. Except for the most urgent patients, which are not taken into account in this study, all patients start at the triage. In this stage, the situation of the patient is estimated and an urgency category is assigned to the patient. The next stage is a doctor that comes to assess what exactly is wrong with the patient and to try to develop a course of treatment. Then, to get better insight into what's wrong with the patient, various tests could be needed. One of these is taking a sample from the patient and sending this to the lab for analysis; this could for instance be a blood-test. Besides this lab-test, an x-ray or CT-scan might be needed. It's also possible the doctor of the A&E department determines that a doctor with a certain specialism from elsewhere in the hospital is required to examine or treat the patient. After this the required treatment can be determined and the patient can be treated. Sometimes a patient then needs to be admitted into the hospital, so an available bed has to be found and prepared, but other times, the patient can then immediately leave the emergency department.

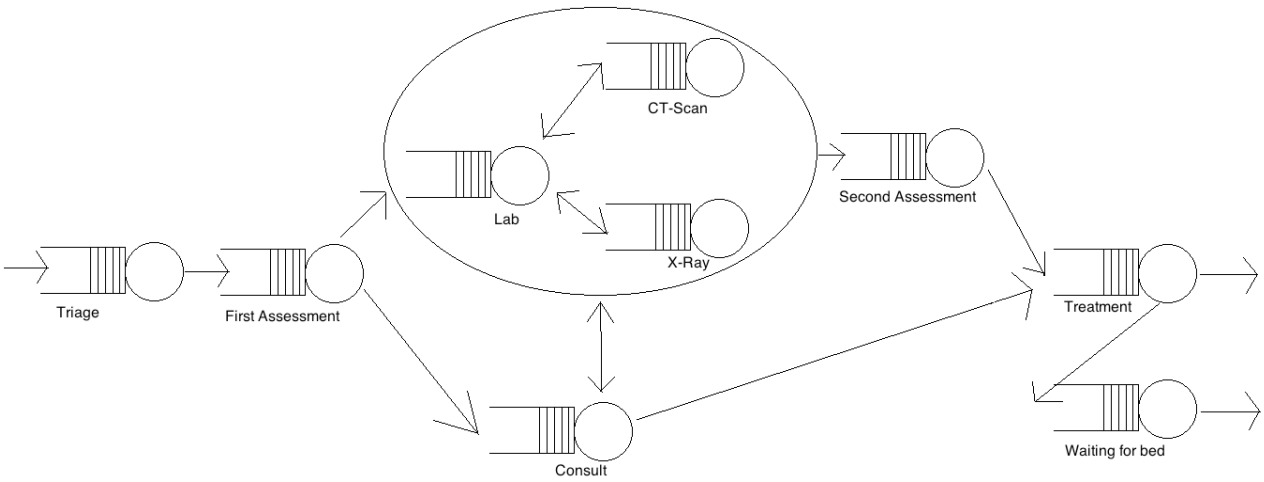

Figure 1. The processes in the A&E department.

There are certain difficulties with this process description. One of these is that there is no information available on the characteristics of the various steps. Therefore it's not possible to determine beforehand what on average the waiting- and processing time at a certain station is. In addition, there are various dependencies within the system. One example is that the doctor that tries to determine beforehand what the course of treatment for a patient should be, is the same one that determines the treatment after, for example, an x-ray has taken place. If this doctor is busy with another patient after the x-ray has finished, a patient will have to wait longer. Also, if a certain specialist from the hospital is required, it's hard to tell when this specialist will be available. This depends on a lot of factors that are hard to capture in a model. This may differ between the various specialisms. All these issues makes modeling the department in a mathematically sound way fairly complicated. So in the model, the department will be considered on an abstract level, rather than trying to fit in all the details.

General approach

The general approach can best be explained by first using independent M/M/1 queues for all processes in the emergency department. Only the consult and waiting for bed station are assumed to be infinite server queues, therefore these stations only cause a delay. So the arrival process is assumed to be Poisson, which would seem like a probable assumption; the service at every station is assumed to be exponential. It is well known that the output process of any station is then again a Poisson process, so the next queue can be analyzed in the same way. In this case all the queues are easy to analyze and it's a good start to build a model from. Of course this is not very realistic, there is not only one server at every 'service station' and there are dependencies between the queues. Also the assumption of exponential service times might not be a good one.

Based on the data from the VU hospital emergency department the probabilities of going from one server to another were computed. These probabilities are shown in Table 1.

	Triage	First Assessment	X-Ray	CT-scan	Lab	Second Assessment	Treatment
Triage	0	1	0	0	0	0	0
First assessment	0	0	0,4227	0,09	0,0836	0	0,4036
X-Ray	0	0	0	0	0,3977	0,6022	0
CT-Scan	0	0	0	0	0,7477	0,2522	0
Lab	0	0	0	0	0	1	0
Second Assessment	0	0	0	0	0	0	1

Table 1. Probability of moving from one station to the next.

From these probabilities the incoming rate at every station, the lambda, can be computed. This is the case, since every patient first arrives at the triage station and goes on to

the first assessment. After this, for instance with a 0.42 probability a patient continues to the x-ray station. Since this is the only station from which a patient can continue to the x-ray station, the arrival rate at the x-ray station is 0.42 times the original arrival rate. Or generally:

$$\lambda_i = \sum_k P(transfer\ from\ k\ to\ i) \cdot \lambda_k$$

Now the only other thing needed are the service characteristics, the beta or mu, of every server. These were initially set to a random value, since these were not known in advance. Now for every queue the rho (or traffic load) is straightforward to compute and with this the expected sojourn time at a certain station (see for instance Koole (2009)). In particular:

$$\varrho_i = \frac{\lambda_i}{\mu_i}$$

and the expected sojourn time (under the M/M/1 assumption)

$$\mathrm{E}(sojourn\ time\ at\ station\ i) = \frac{1}{\mu_i(1-\varrho_i)}$$

Because of the independence assumption, all expected sojourn times can be added up in the right way to get an overall expected sojourn time. For example if we want to compute the expected sojourn time of a patient that goes to both the ‘lab-station’ and the ‘x-ray-station’, the sum would be

$$\mathrm{E}(sojourn\ time\ lab) + \mathrm{E}(sojourn\ time\ x\text{-}ray) + \Sigma\, P(go\ to\ station\ i\ I\ go\ to\ lab \cap go\ to\ x\text{-}ray) \cdot \mathrm{E}(sojourn\ time\ i)$$

For various patient-groups the expected sojourn times according to this model were computed and compared with the actual values. Since there are 9 stations, at least 9 equations are needed to determine the average service times from the actual data. Therefore, 9 different patient-flows were selected and their actual average staying times at the A&E department were computed from the data. All these patient-flows are disjunct to en-

sure that proportionate weight is given to every station. Next, to determine the expected service time per station, the sum over the absolute differences between the actual average sojourn time and the computed average sojourn time was minimized

$$\min\left\{\sum_i |\mathrm{E}(waiting\ at\ station\ i) - Average(waiting\ at\ station\ i)|\right\}$$

by changing these expected service times using the Generalized Reduced Gradient (GRG2) algorithm.

These differences can be seen in Table 2.

Patient type	Computed	Data
Waiting Overall:	145	142
Waiting (L+R+O-C):	210	212
Waiting (L+R+O+C):	244	212
Waiting (-L-R-C-O):	54	78
Waiting (R+L-O-C):	197	197
Waiting (R+L-O+C):	232	232
Waiting (R-L-O-C)	158	119
Waiting (L-R-O-C):	165	165
Waiting (R-L-O+C)	193	194
Waiting (CT+L+O+C)	283	283

Table 2. Comparison of actual and predicted values from the M/M/1 model

All values are sojourn times at the A&E department in minutes. An explanation of the various patient types can be found in Appendix B.

Extensions to the simple model

M/M/s queues

A natural extension to the previous model is not using queues with necessarily one server, but allowing more servers per station, still keeping the service times independent and exponential. This affects the first and second assessment stages where there are three doctors and the treatment stage where there are three nurses. This still allows for easy computation of the expected waiting times at the queues, but is a bit more realistic than assuming there is only one server at every station. Again, in the same way, the expected service times per station were computed by minimizing the absolute differences using the M/M/s expected sojourn times for appropriate stations. The results of this model can be seen in Table 3.

	Computed	Data
Sojourn time Overall:	148	142
Sojourn time (L+R+O-C):	203	212
Sojourn time (L+R+O+C):	228	212
Sojourn time (-L-R-C-O):	78	78
Sojourn time (R+L-O-C):	203	197
Sojourn time (R+L-O+C):	227	232
Sojourn time (R-L-O-C)	143	119
Sojourn time (L-R-O-C):	144	165
Sojourn time (R-L-O+C)	169	194

	Computed	Data
Sojourn time (CT+L+O+C)	283	283

Table 3. Comparison of actual and predicted values from the M/M/s model

For some patient groups this is not very close to the actual values, and this is to be expected, as the model used is not exactly like reality. Note for instance that in reality the average duration of stay is equal for (R+L+O+C) and (R+L+O-C) type patients. In the model used now, it's not possible to model this. The average waiting and service times per station can be seen in Table 4.

	Average service time	Average waiting time
Triage	8	10
First Assessment	21	2
X-ray	20	38
Lab	24	34
CT-scan	63	36
Second Assessment	7	0
Waiting for bed	-	0,001
Consult	-	25
Treatment	36	0,03

Table 4. Service and waiting times per station in M/M/s model

These values are not what one would expect. Note for instance that there is hardly any waiting for the first and second assessment and the treatment.

Dependence between first and second assessment and parallel stations

To provide better results, a number of extensions were implemented. In practice there are not separate stations for both the first and second assessment, the same doctor both carries out the first and second assessment. To take this into account in the model as well, the first and second assessment stations are merged into one station where most patients pass twice. This is an approximation, as in reality it is unlikely both the first and second assessment take the same amount of time. So this queue actually has more variability than is now modeled, however due to the complexity of a multi-type, multi-server queue without priorities, this approach was chosen (see for instance Van Harten and Sleptchenko (2003)).

Second, it seems reasonable to assume that if a patients needs both a lab test and an x-ray, these processes are often carried out in parallel. Therefore, the expected waiting time for a patient that requires both an x-ray and a lab test, is the expectation of the maximum of the two. A derivation of the expectation of the maximum of two exponentials can be found in appendix A. Incorporating these two modifications gives the results of Table 5. The results are a little bit closer to the actual values, but still they are not completely the same. Therefore, other extensions were made.

Patient type	Computed	Data
Sojourn time Overall:	148	142
Sojourn time (L+R+O-C):	195	212
Sojourn time (L+R+O+C):	232	212
Sojourn timeSojourn time (N-C-O):	78	78
Sojourn time (R+L-O-C):	195	197
Sojourn time (R+L-O+C):	232	232
Sojourn time (R-L-O-C)	138	119
Sojourn time (L-R-O-C):	165	165
Sojourn time (R-L-O+C)	175	194
Sojourn time (CT+L+O+C)	283	283

Table 5. Comparison of actual and predicted values for the extended M/M/s model

Fine tuning the model

A general station called ‘general waiting’ was added. This is to reflect the fact that the system in reality is not as perfect as it is in this model. Factors that may influence the sojourn times of all patients are for example acute patients arriving at the shock room, transportation times, doctors that get called away, registration times, etc. This delay was assumed to be independent of the type of patient and exponentially distributed. Moreover, a speed fac-

tor was added, to reflect the fact that certain types of patients can be treated a lot quicker than others. For instance, a patient that only needs an x-ray, is likely to be a patient that has a minor injury, like for instance a broken arm or leg. This type of patient is fairly easy to treat, so also his triage, first and second assessment and treatment will be faster than for a complicated patient. This was subtracted from the total duration of stay. Now the same model was again optimized and the results of Table 6 were obtained.

	Computed	Data
Sojourn time Overall:	140	142
Sojourn time (L+R+O-C):	212	212
Sojourn time (L+R+O+C):	212	212
Sojourn time (N-C-O):	78	78
Sojourn time (R+L-O-C):	193	197
Sojourn time (R+L-O+C):	232	232
Sojourn time (R-L-O-C)	119	119
Sojourn time (L-R-O-C):	165	165
Sojourn time (R-L-O+C)	195	194
Sojourn time (CT+L+O+C)	283	283

Table 6. Comparison of actual and predicted values for the fine-tuned M/M/s model

These values are all very close to their actual values as measured in the data. Other parameters of interest for this model are the average waiting and service times per station (in minutes). These can be found in Table 7.

	Average service time	Average waiting time
Triage	5	3
First Assessment	15	2
X-ray	23	38
Lab	29	40
CT-scan	63	36
Second Assessment	15	2
Waiting for bed	-	18
Consult	-	38
Treatment	16	11
General Waiting	-	25
Speed Factor	-	-38

Table 7. Service and waiting times in the final model

Sojourn time distribution

A natural extension to the previous analysis is to consider the entire sojourn time distribution as opposed to only the expected value. Under the same assumptions as before, the total sojourn time of a patient becomes a sum of exponential distributions (this is not entirely clear for the 'fine-tuned' model, as no distribution was assumed for the speed factor). Note that at stations that were modeled as an M/M/1 or M/M/infinity queue the sojourn time is exponential. Since waiting times and service times at an M/M/s queue are also exponential, the total result is just a sum of these exponentials. Using Koole [2004], it is possible to find the distribution function by assuming it to be a Cox distribution. Then all the stations are a node in the Cox representation and the waiting at the multi-server queues become a node as well. In a Cox distribution there is usually a probability of skipping all the nodes ahead, however in this model the probability of leaving the system earlier are zero everywhere. There are two difficulties. First, some stations have the same average service time (the first and second assessment station) and therefore also both the same service and waiting time distribution. Second, waiting does not always occur at the multi-server stations. This was handled by conditioning on waiting or not at the various stations and then later joining the distributions. Then

$$\bar{F}_k(t) = \sum_{i=1}^{k} c_{ik} t^{m(i)} \mathrm{e}^{-\mu_i t}$$

for all $t \geq 0$, *with* c_{ik} *as follows:*

$$c_{ik} = \begin{cases} 1 & \text{if } i = k = 1; \\ \dfrac{\mu_k c_{ik-1} \bar{\alpha}_k - c_{h(i,k)k}(m(i)+1)}{\mu_k - \mu_i} & \text{if } \mu_i \neq \mu_k; \\ \dfrac{\mu_k c_{n(i)k-1} \bar{\alpha}_k}{m(i)} & \text{if } \mu_i = \mu_k, m(i) > 0, k > 1; \\ 1 - \displaystyle\sum_{1 \leq j < k : m(j) = 0, j \neq i} c_{jk} & \text{otherwise, i.e., if } \mu_i = \mu_k, m(i) = 0, k > 1. \end{cases}$$

with

- $m(j) = \#\{i \mid \mu_i = \mu_j , 1 \leq i < j \}$,
- $h(j, k) = \min_i\{ \mu_i = \mu_j , j < i \leq k\}$,
- $n(j) = \max_i\{ \mu_i = \mu_j , 1 \leq i < j\}$ if $m(j) > 0$, 0 otherwise,
- $l(k) = \min_i\{ \mu_i = \mu_k , 1 \leq i \leq k\}$

Implementing the numerical algorithm in Koole (2004) with probabilities of leaving the system set to zero and unconditioning allows to analyze the length of stay at the A&E department. Comparing the obtained results to the actual data gives the following result for the patients that require lab, x-ray and a consult:

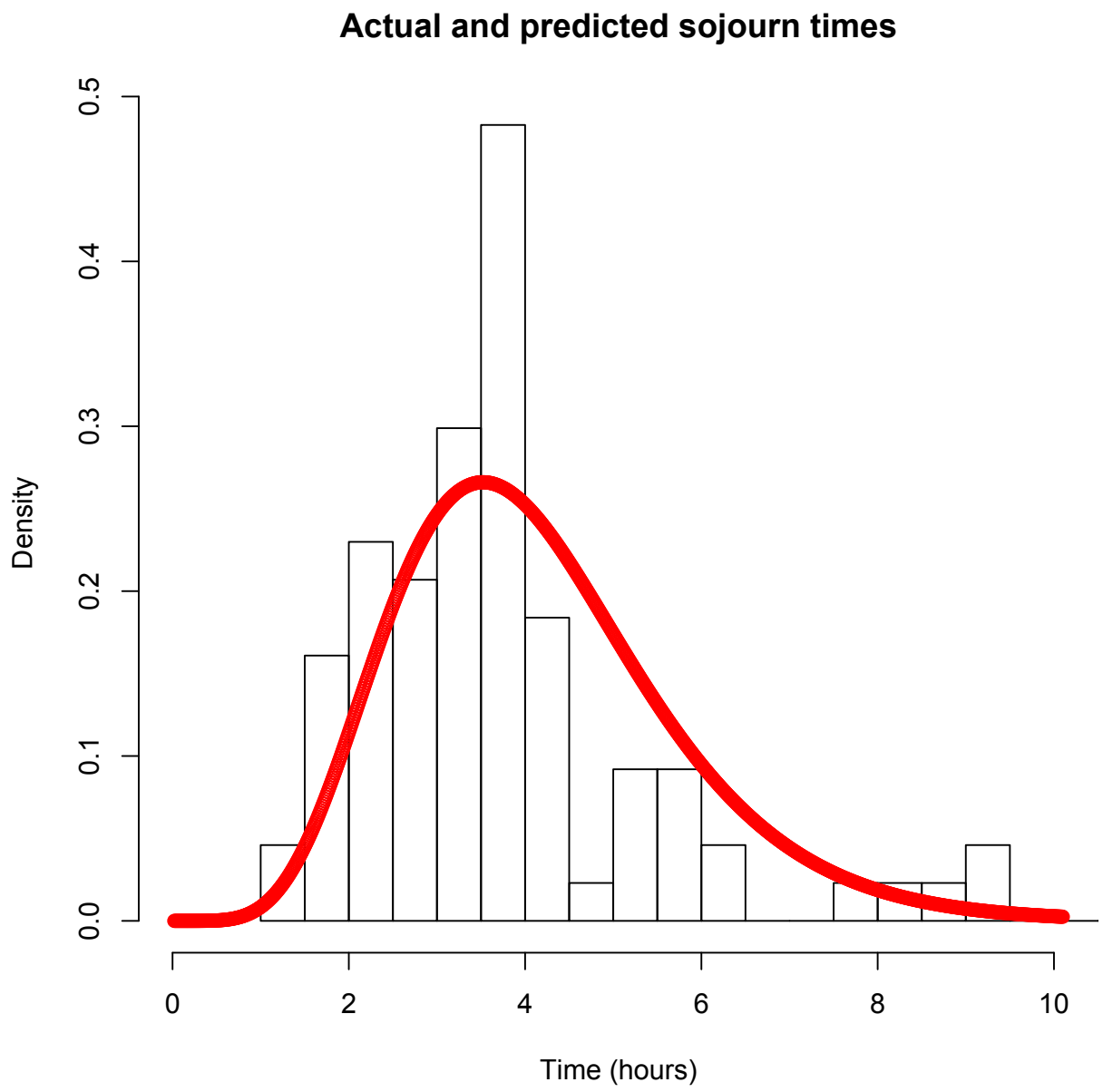

Figure 2. Comparison of actual and predicted sojourn time distribution for (L+R-O+C) patients

For the patients that require only a first assessment and treatment, the following results were obtained:

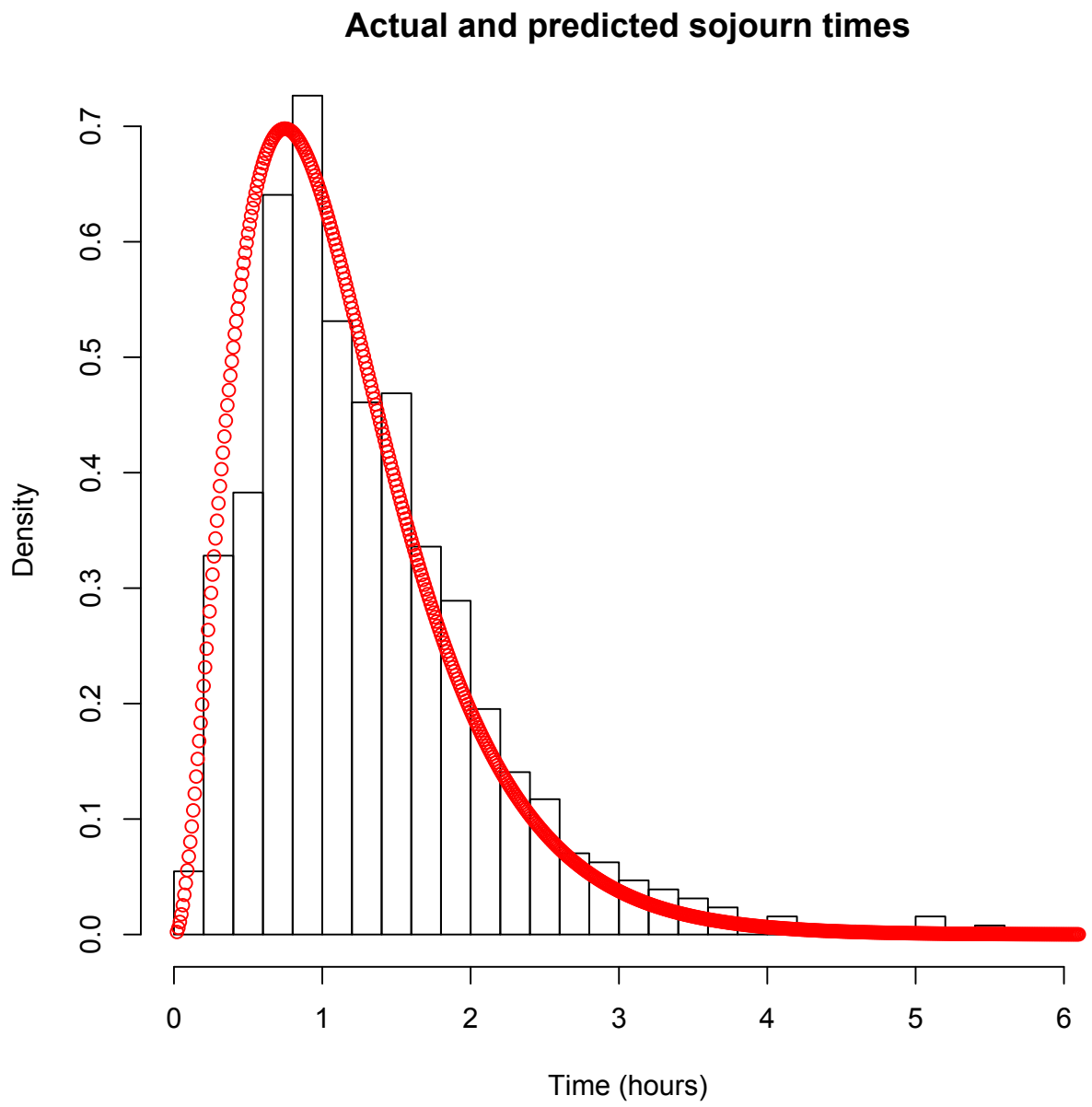

Figure 3. Comparison of actual and predicted sojourn time distribution for (-L-R-O-C) patients

These patient-flows were chosen because they are very different, one requiring a lot of care and the other hardly any, and because there was enough data about these patients to reliably compare with a prediction. The average service times were taken from the M/M/s

model with the coupled first and second assessment, but without the parallel servers, the general waiting and the speed-factor, as it is not so apparent what type of distribution these have.

It becomes clear that the prediction by the model does have the same form and is also fairly close to the data. However the empirical distribution is slightly more peaked and has slightly fatter right tail.

Conclusion

Modeling an emergency ward as a system of queues makes it possible to estimate the average waiting time at every separate station and if enough detail is added, provides fairly accurate results for the average waiting and service times. If the entire distribution is taken into account, this also yields fairly accurate results. The difference between the predicted and empirical sojourn time distribution is mainly caused by the outliers in the right tail. But this is not something that can be expected from such a model. However, it remains an open question whether minimizing the difference between the predicted and actual distribution yields even better results.

Appendix A: Derivation of the expectation of the maximum of two exponentials

Assume S_1, S_2 are both independently, exponentially distributed with parameters η and ξ.

Then

$$P(max(S_1,S_2)<x)=P(S_1<x;S_2<x)=P(S_1<x)P(S_2<x)=$$

$$(1-e^{-\eta x})(1-e^{-\xi x})=1-e^{-\eta x}-e^{-\xi x}+e^{-(\eta+\xi)x}$$

,so

$$P(max(S_1,S_2)>x)=e^{-\eta x}+e^{-\xi x}-e^{-(\eta+\xi)x}.$$

Therefore,

$$\mathrm{E}[\max(\mathrm{S}_1,\mathrm{S}_2)]=\int_0^\infty P(\max(\mathrm{S}_1,\mathrm{S}_2)\geq x)dx=\int_0^\infty e^{-\eta x}+e^{-\xi x}-e^{-(\eta+\xi)x}dx=\frac{1}{\eta}+\frac{1}{\xi}-\frac{1}{\eta+\xi}$$

Appendix B: Explanation of various patient types

In all the tables the type of patients were abbreviated to be able to show this information quickly. The letters mean the following:

R: X-ray

L: Lab test

CT: CT-Scan

O: Admitted to the hospital

C: External consult needed.

In the table, a '+' sign before such a symbol means a patient needs this type of care, whereas a '-' sign before such a symbol means that the patient does not need this type of care. For compactness, the first symbol is always assumed to be needed. So type (R-L+O+C) is a patient that requires x-ray, admittance to the hospital and an external consult, but does not require a lab test.

References

Fletcher, A., Halsall, D., Huxham, S., Worthington, D. (2007), "The DH Accident and Emergency Department model: a national generic model used locally", *Journal of the operational research society* (**58**)3:1554-1562

Harten, A. van, Sleptchenko, A. (2003), "On multi-class multi-server queueing and spare parts management.", *Queueing systems* (**43**)4:307-328

Kolker, A. (2008), "Process modeling of emergency department patient flow: Effect of patient length of stay on ED diversion", *Journal of medical system* (**32**)5:389-401

Koole, G. (2004) "A formula for tail probabilities of Cox distributions", *Journal of applied probability* (**41**) 935-938

Koole, G. (2009) "Optimization of business processes: An introduction to applied stochastic modeling"

Mayhew, L., Smith, D. (2008), "Using queueing theory to analyze the Government's 4h completion time target in Accident and Emergency departments", *Health care management science* 11:11-21