Understanding COPD patients in the hospital system via administrative data

Henry Wilde, Vincent Knight, Jonathan Gillard

Abstract

This work presents an analysis of how patients with chronic obstructive pulmonary disorder (COPD) interact with the hospital system in South Wales.

1 Introduction

This introduction will briefly summarise the literature review for studying a patient corpus via clustering. Following this, a condensed data analysis is presented highlighting the main conclusions of the clustering and the overall benefits compared with traditional condition-treatment segmentation.

2 Constructing the queuing model

The simplest model using the data available is an $M|M|c$ queue with multiple classes. In this model, the following assumptions are made:

1. Inter-arrival and service times of patients are each exponential with some mean.
2. There are c servers available to arriving patients at a single node representing the overall resource availability at the hospital.
3. There is no queue or system capacity.
4. A first-in first-out service policy is implemented.

Each group of patients has its own arrival distribution. The parameter of this distribution is taken to be the reciprocal of the mean inter-arrival times for that group.

Like arrivals, each group of patients has its own service time distribution. This will be calculated approximately via the length of a patient's stay. The length of stay is the total time spent in the system. Without full details of the process order or idle periods during a spell, some assumption must be made about the true 'service' time in relation to the time spent in hospital. This work considers the mean service time, $\frac{1}{\mu}$, to be proportional to the mean total system time, $\frac{1}{\phi}$, such that:

$$\mu = p\phi \tag{1}$$

where $p \in (0, 1]$ is some parameter to be determined for each group.

As the full details of how the patients move through the hospital system, and the details of the system itself, are unknown, an appropriate number of servers c must be found as well as p.

In order to evaluate appropriate values of each p and the value of c, the system is simulated across of parameter space. Then, for each set of parameters, the total time distribution is compared with that in the available data via the (first) Wasserstein distance. This distance measures the approximate 'minimal work' required to move between two probability distributions where 'work' can be loosely defined as the product of how much of the distribution's mass must be to be moved with the distance it must be moved by. More formally, the Wasserstein distance between two probability distributions U and V is defined as:

$$W(U, V) = \int_0^1 \left| F^{-1}(t) - G^{-1}(t) \right| dt \tag{2}$$

where F and G are the cumulative density functions of U and V respectively. A proof of (2) is presented in [1].

Then the parameter set with the smallest mean distance over a number of runs is taken to be the most appropriate.

Reiterate the objective of the paper — to model a COPD ward within a hospital — and draw attention to lack of fine-grain data. Lead into how this can be overcome with the Wasserstein distance (a lot of this has been written up in `nbs/wasserstein.ipynb`). A brief summary of how the parameter set is chosen and a nice image of the queue we are building. Close out the section with best and worst case parameter set plots.

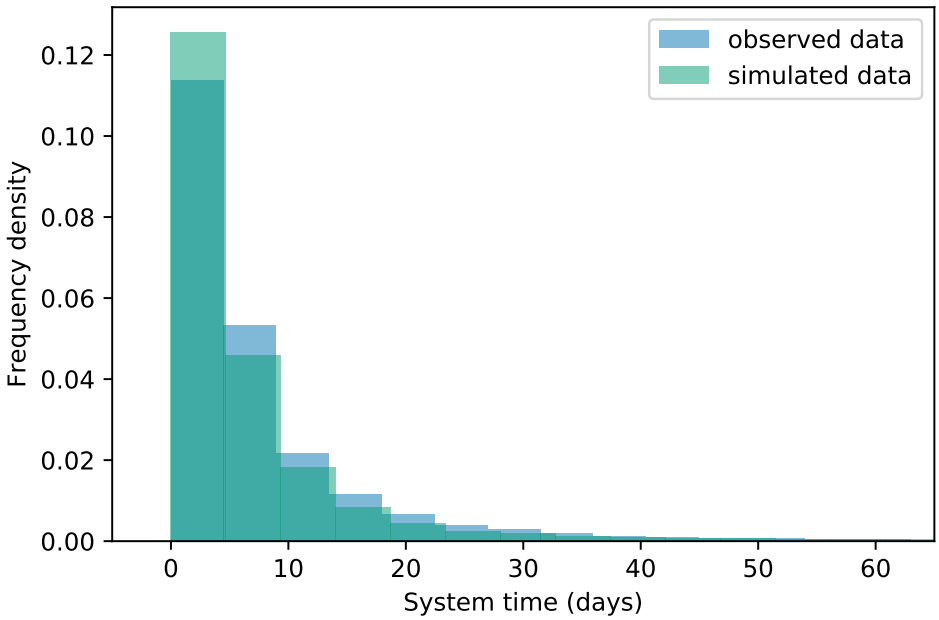

Figure 1: A histogram of the simulated and observed length of stay data for the best parameter set.

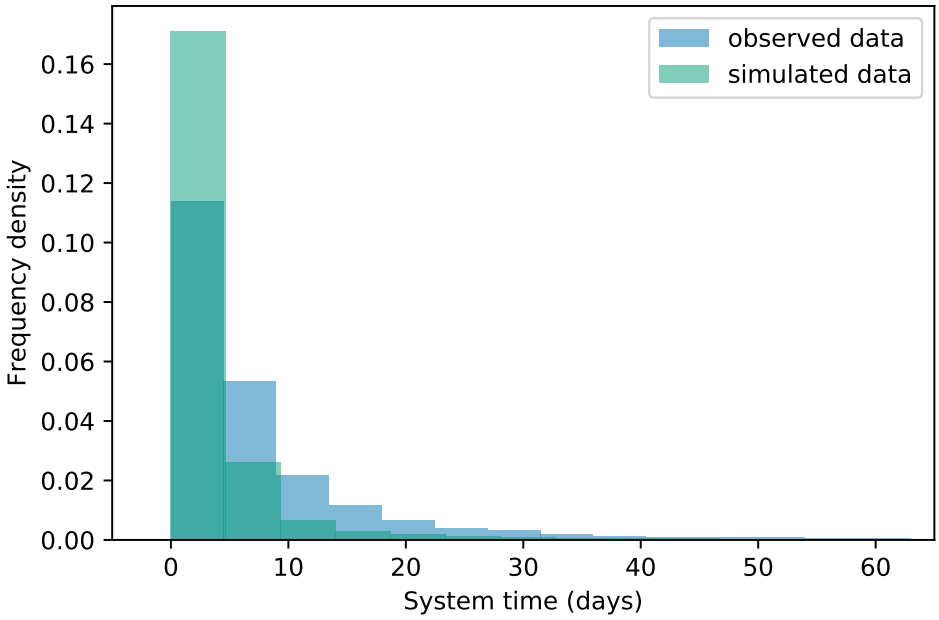

Figure 2: A histogram of the simulated and observed length of stay data for the worst parameter set.

3 Adjusting the queuing model

Body of the writing and plots come here. What can we see in the what-if scenarios? The main scenarios are:

- How would server utilisation (i.e. resource consumption) be affected by an increase in overall patient arrivals?
- How is the system affected by certain types of patients (e.g. short-stay, low-impact) arriving less frequently?
- What are the sensitivities of mean system times and server utilisation based on a change in c?

4 Conclusion

Summarise the findings and novelty of the paper: sensitivity analysis and queuing models are within reach despite a lack of data. The chosen modelling discipline for service times is very simplistic but can return good results (refer back to best-case parameter plot).

References

[1] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.