Methodology Behind COVID-CAT

Qian Cheng, Nilay Tanik Argon, Yufeng Liu, Serhan Ziya

Department of Statistics and Operations Research, University of North Carolina at Chapel Hill

An Infinite-server Queueing Model

COVID-CAT is based on an infinite-server queueing system known as $M_t/GI/\infty\ TVIS$ to model the occupancy process at an ED, medical ward, or ICU of a hospital, see, e.g., Section 4 in Whitt (2018). Here, M_t stands for Poisson arrivals with a time-varying rate, GI means that the service times form a sequence of independent and identically distributed (i.i.d.) random variables, and $TVIS$ stands for time-variant arrival rate and infinite servers. Because of the unlimited number of servers, every patient is taken into service upon arrival, and hence, there is no wait. The rationale behind this approximation is that a $TVIS$ model is considered as an ideal version of a time-varying queueing system with a finite number of servers, where there is enough capacity so no customer will encounter substantial wait. Since our goal is to conduct a capacity analysis, i.e., determine the number of resources to deliver a targeted service performance, this queueing model approximation would be appropriate. Furthermore, in the COVID-19 setting, it is undesirable for patients to encounter long waiting times to prevent cross-infection in a common waiting room.

The number in the system, $X(t)$, in an $M_t/GI/\infty\ TVIS$ queue, has a Poisson distribution for each $t \geq 0$ with mean

$$\begin{aligned} m(t) \equiv E[X(t)] &= \int_0^\infty \lambda(t-s)G^c(s)ds \\ &= E[S]E[\lambda(t-S_e)], \end{aligned} \tag{1}$$

assuming the system started empty in the distant past, i.e., $t = -\infty$, see, e.g., Section 1 in Eick *et al*. (1993). Here, $\lambda(t)$ is the deterministic arrival rate at time *t*, G is the cumulative distribution function of service time S, $G^c(s) \equiv 1 - G(s)$, and S_e is a random variable with cumulative distribution function

$$G_e(x) \equiv \frac{1}{E[S]} \int_0^x G^c(s)ds,\ x \geq 0. \tag{2}$$

The time-variant mean $m(t)$ in (1) is usually called the *offered load* because it represents the expected number of servers needed to serve all customers at any given time point t if we ignore the capacity constraint. Notice that in the stationary case, where λ is a constant, the offered load is $m(\infty) = \lambda E[S]$, which is consistent with the Little's Law. Therefore, the only difference here is a random time lag S_e, of which the impact could be roughly estimated using $E[S_e] = E[S]\frac{c_s^2+1}{2}$, c_s is the squared coefficient of variation of the service time distribution, see, e.g., Section 4.1 in Whitt (2018).

Modification to Incorporate Initial System State Information

The original $M_t/GI/\infty\ TVIS$ discussed above assumes that the system has started empty at a distant past. However, in reality, we can observe how many people there are in the system at time zero (current census). In order to incorporate this information, we make some modifications to the original formulation as we explain next.

Suppose at time zero, we observe that there are n_0 patients in the system. $X(t)$, for $t \geq 0$, could then be decomposed into two parts as $X(t) = Y(t) + Z(t)$, where $Y(t)$ is the number of patients that are still in the system at time t among those who were present at time 0, and $Z(t)$ stands for the number of patients that arrived after time zero and are still in the system at time t.

Using the original theoretical model after a simple modification, we find that for each $t \geq 0$, $Z(t)$ follows a Poisson distribution with mean

$$E[Z(t)] = \int_0^t \lambda(t-s)G^c(s)ds.$$

However, the same formulation cannot be used to obtain the distribution of $Y(t)$. For patients that were in the system at time zero, suppose their remaining service time distribution is approximated by the equilibrium distribution associated with G, i.e., $G_e(x)$ defined in (2). Then, $Y(t)$, the number of patients who were in the system at time zero and are still in the system at time t, follows a binomial distribution, where $Y(t) \sim B(n_0, p(t))$ and $p(t) = 1 - G_e(t)$, and thus, its mean could be calculated simply as $E[Y(t)] = n_0 p(t)$. Hence, the expected number of all patients in the system at time t is given by

$$E[X(t)] = n_0(1 - G_e(t)) + \int_0^t \lambda(t-s)G^c(s)ds. \tag{3}$$

Since $Y(t)$ and $Z(t)$ are independent, we also have

$$Var(X(t)) = Var(Y(t)) + Var(Z(t)) = n_0 G_e(t)(1 - G_e(t)) + \int_0^t \lambda(t-s)G^c(s)ds.$$

Moreover, the probability that there are $m \geq 0$ patients in the system at time t is given by

$$P(X(t) = m) = \sum_{i=0}^{\min(m,n_0)} P(Y(t) = i)P(Z(t) = m - i).$$

COVID-CAT Outputs

COVID-CAT calculates the following performance measures based on the formulation discussed in the previous sections:

- **Expected occupancy and 95% band:**

  Aside from the expected occupancy $E[X(t)]$ calculated using (3), a 95% band $[Q_{0.05}(t), Q_{0.95}(t)]$ is obtained as well, where $Q_{0.05}(t)$ and $Q_{0.95}(t)$ are the 5% and 95% quantiles of the distribution of the number of patients in the system at time t, where $Q_p(t) = \{\min x \geq 0 : P(X(t) \leq x) \geq p\}$.

- **Probability of exceeding a threshold:**

  Another performance measure of interest for practitioners would be the probability that $X(t)$ exceeds a certain threshold ψ during a given period of time $[0, t]$, which can be approximated by

$$T_t(\psi) = \max_{0 \leq u \leq t} P(X(u) \geq \psi).$$

$T_t(\psi)$ is an approximation because the probability that the maximum occupancy over period $[0, t]$ exceeds a certain threshold is not necessarily equal to the maximum of the probabilities of occupancy exceeding this threshold at each time during this period.

References

[1] Whitt, W. (2018). Time-varying queues. *Queueing Models and Service Management*, *1*(2), 79-164.

[2] Eick, S. G., Massey, W. A., & Whitt, W. (1993). The physics of the $M_t/G/\infty$ queue. *Operations Research*, *41*(4), 731-742.