CENTRE DE RECHERCHE
SUR LES STRATÉGIES ÉCONOMIQUES

Modeling medical material shortage using Markov processes

Alexis Roussel, Romain Biard, Marc Deschamps, Mostapha Diss

September 2023

Working paper No. 2023 – 04

CRESE
30, avenue de l'Observatoire
25009 Besançon
France
http://crese.univ-fcomte.fr/

UFR SJEPG
Sciences juridiques économiques politiques et de gestion
UNIVERSITÉ DE FRANCHE-COMTÉ

Modeling medical material shortage using Markov processes

Alexis Roussel*†[1], Romain Biard[2], Marc Deschamps[1,4], Mostapha Diss[1,3]

[1] *Université de Franche-Comté, CRESE, F-25000 Besançon, France.*
[2] *Université de Franche-Comté, LMB UMR 6623, F-25000 Besançon, France.*
[3] *Africa Institute for Research in Economics and Social Sciences (AIRESS), University Mohamed VI Polytechnic, Rabat, Morocco.*
[4] *OFCE-Sciences Po*

Abstract

The management and allocation of health resources, in particular equipment such as ventilators, has been the object of significant interest by a health community that is concerned to avoid new shortages. In this article, we develop a Markov chain based model considering random arrivals and discharges of patients in an intensive care unit requiring ventilator support. We provide a methodology in order to compute the exact probability distribution of the time-shortage, which is the moment where no more ventilators are available. We propose two applications of this model: a preventive traffic signal and a tool to evaluate purchasing decisions. A calibration of parameters based on real empirical data from a French hospital is carried out in order to test the operational use of the model.

Keywords: Intensive care unit, ventilator shortage, Markov process, traffic signal, purchasing management.
JEL classification: I10, I19, C0, C44, C22.

1 Introduction

The Covid-19 pandemic has caused significant medical, economic and supply chain disruptions worldwide. As countries continue to grapple with the ongoing crisis, shortages of essential medical equipment have emerged as a critical challenge (Bhaskar et al., 2020). Among these shortages, the lack of ventilators has been particularly acute (Dar et al., 2021; Santini et al., 2022). According to the World Health Organization,[1] an estimated 5% of Covid-19 patients required mechanical ventilation to help them breathe in 2020. The Centers for Disease Control and Prevention estimates that 2.4 million to 21 million

*Corresponding author. Email: alexis.roussel@univ-fcomte.fr

†We would like to express our heartfelt gratitude to the "Région Bourgogne-Franche-Comté" for funding the PhD thesis of the corresponding author Alexis Roussel.

[1] World Health Organization (2020) Management of severe/critical cases of Covid-19 with non-invasive or mechanical ventilation: based on information as at 1st June 2020. Regional Office for Africa.

Americans required hospitalization during the pandemic, and the experience in Italy has shown that about 10 to 25% of hospitalized patients required ventilation for several weeks in some cases (Truog et al., 2020). On the basis of these estimates, the number of patients needing ventilation could range between 1.4 and 31 patients per ventilator. There is a broad range of estimates of the number of ventilators needed to care for U.S. patients with Covid-19, from several hundred thousand to as many as a million. Naturally, the estimates vary depending on the number, speed, and severity of infections. Current estimates of the number of ventilators in the United States range from 60,000 to 160,000, depending on whether the ventilators that have only partial functionality are included or not, but this remains insufficient (Ranney et al., 2020). The US Department of Health and Human Services[2] estimates that 865,000 US residents would be hospitalized during a moderate pandemic (as in 1957 and 1968) and 9.9 million during a severe pandemic (as in 1918). A moderate crisis could require 64,875 ventilators and a severe one up to 742,500, even if the limiting factor during a pandemic-level crisis would be the number of respiratory therapists. This sudden surge in demand for ventilators has put unprecedented pressure on global supply chains, leading to shortages that have left health-care workers struggling to save lives.

When there is a shortage of medical resources, making choices regarding the allocation of scarce resources becomes crucial. The ongoing Covid-19 pandemic has brought these allocation issues to the forefront and has highlighted the importance of effective resource management and optimal anticipation and prediction of such complicated situations. In this article, we design a time-based model focusing on the time of shortage of ventilators, that is the moment no more ventilators are available for future incoming patients in an intensive care unit (ICU). Patients, with multiple profiles, transit in and out of the service with random distributions for arrivals and departures and with independence between patients. We manage to numerically compute the exact probability distribution of this random time. Consequently, we can assess the risk of an impeding shortage event, and, using "traffic signals", help health-care professionals anticipate difficult moments that will arise if no significant measures are taken. Scenario-based applications of the model are provided and our claims are finally strengthened by a calibration based on empirical data.

An abundance of literature has emerged since the outbreak of Covid-19 concerning the topic of medical scarcity. The primary problematic, when it comes to medical shortages, is the ethical dimension of decision-making. This is a complex issue in ICU since it involves choices regarding life support and end-of-life care. This may include difficult decisions such as choosing who gets access to life-saving equipment like ventilators and ICU beds, and these ought not be made in the heat of the action by clinicians in the field (e.g., Rosenbaum, 2020). In such situations, health-care providers are forced to make ethical decisions based on available guidelines and protocols. For instance, Emanuel et al. (2020) proposed a list of 6 recommendations to guide triage protocols for health-care

[2]US Department of Health and Human Services (2005) HHS pandemic influenza plan. Available at https://www.cdc.gov/flu/pandemic-resources/pdf/hhspandemicinfluenzaplan.pdf.

resources. For Truog et al. (2020), a strategy for avoiding debilitating distress over these decisions is to use a triage committee. Many health-care institutions have developed allocation frameworks to guide these decisions, based on factors such as the patient's age, overall health, and likelihood of survival. These frameworks are generally designed to ensure fairness and transparency in the allocation process, but they can be challenging to implement in practice because it implies refusing patients treatment on the basis of harsh and inflexible selection criteria. However, for Rosenbaum (2020), such transparency and inclusiveness is paramount to make those decisions acceptable. As far as we are concerned, we do not pretend to establish any sort of guideline or protocol triage. As a matter of fact, we do not consider sorting patients in our model since all of them, regardless of their conditions, are admitted into ICU as long as there is at least one ventilator available. The whole ethical debate about who to admit or not is henceforth not considered in our model, and hence our paper deviates in this respect from the existing literature related to these medical allocation problems.

Another important aspect concerning medical scarcity that has caught the attention of researchers, is the reusable or non-reusable nature of the resource that is lacking. With hospitals operating beyond capacity during the pandemic, health-care providers have struggled to keep up with the demand for non-reusable medical equipment. For example, hospitals have been forced to reuse single-use items like gowns and masks. Hence, the study of mask allocation (e.g., Chen et al., 2021), considering the hidden information that some people already have masks, is relevant. Distributing vaccines in the most inclusive and popular way is also essential for Nganmeni et al. (2022), whereas for Akbarpour et al. (2021) the optimal allocation of vaccines must take into account externalities and equity concerns. Employing the tools of game theory, they found that the number of vaccine doses necessary to generate such an allocation is greater than the one necessary to obtain an allocation that is only popular. Saha and Ray (2019) investigate medicine inventory management using a stochastic model, finding on the basis of empirical evidence that when medication demand is based on the patient's condition, the total inventory-related cost is significantly lower compared to demand based on historical daily usage.

The shortage of reusable medical equipment has also highlighted the need for more sustainable and efficient health-care systems for the use of items such as wheelchairs, hospital beds (Lee and Lee, 2018) and mechanical ventilators, as investigated by Bonneuil (2021), Olmos and Borzone (2021), and Pathak et al. (2020), among others. Notice that the rate at which a patient consumes a certain resource depends on the type of the resource itself. Hence, a patient will need one hospital bed and one ventilator, but maybe three to four nursing staff members or respiratory doctors and many more medicine pills, vaccine doses, masks, etc. Having pointed this out, we assume, in our model, that a patient needs one and only one ventilator to be properly treated (as long as it is functional) so that the "consumption rate" is one. The other feature regarding the missing resource is its reusable nature, and it is clear that when a patient enters the service, he/she uses a ventilator and when he/she exits the service, for an unknown reason among which we may

include recovery or death, his/her ventilator becomes available again for a future patient, which is exactly the definition of a renewable/reusable object.[3]

The mathematical tool we consider in this article is the time-series (e.g., Box et al., 2015) of the number of patients in ICU. One type of time-series analysis is Markov chain analysis (see, for instance, Hillier and Lieberman, 2001; Norris, 1998), which is a mathematical tool used to model and analyze complex systems with changing states over time. In a Markov chain, the future state of a system depends only on the current state, and not on any previous states. This makes it a useful tool for modeling systems that have a certain degree of randomness or uncertainty.[4] The literature regarding medical shortages makes heavy use of Markov processes. We can mention Bonneuil (2021) who uses queuing theory[5] to determine an optimal age threshold minimizing a mortality rate weighted by life expectancy, so that people with lower risk of dying and better chance of recovery obtain a ventilator in priority to people with bad clinical conditions. Lee and Lee (2018) consider models where patients arriving first are first served, similar to our own work but, conversely, the arrivals of patients are dependent between them since they originate from a surge demand due to, for instance, a massive casualty; moreover, the chain is continuous with finite time horizon. The model developed by Olmos and Borzone (2021) is closely related to our framework since it aims to predict the number of ventilators available over the very short term (25 days, updating every 5 days) with empirical data and a discrete Markov process. Meisami et al. (2019) model a hospital as a complex loss queuing network, which includes a stochastic model determining the length of time patients spend in specific units and how they move between them based on their risk levels. An optimal admission control policy for the units network is estimated using a Mixed Integer Programming model.

As far as we are concerned, all patients are included without considering any kind of social or health condition information. The different states we consider are the number of patients ventilated, ranging from 0 (i.e., empty service) to the total number of ventilators. Transitions, i.e. gaining or losing a patient, occur according to very simple Bernoulli probabilities. Our vision corresponds to a very naive, but yet simple, way of imagining incoming and outgoing flows in an ICU, which enables us to compute the exact probability distribution function of what we will call the time-shortage. In other words, the time-shortage is described as a random variable measuring the moment at which no more ventilators are available, that is, all ventilators are occupied by patients in need of respiratory assistance. To the best of our knowledge, finding the properties of the moment the shortage arises has not yet been considered since most articles are concerned with the best way to provide mechanical ventilators to the weakest patients.

[3] Although our model focuses on the case of ventilators, it can also be applied to other equipment such as hospital beds or any reusable equipment.

[4] Markov chain analysis is used in a wide range of applications, from predicting stock prices (e.g., Hassan and Nath, 2005) to understanding the spread of infectious diseases (e.g., Gómez et al., 2010).

[5] He uses an M/M/C queue denoting the arrival process (Poisson distribution), the service distribution (exponential distribution) and the total number of servers.

For this kind of model, an important step consists in illustrating its effectiveness with graphical and numerical applications, which can be achieved in three different manners. One way is to calibrate the different parameters of the model based on real data. Many well-discussed studies appear to be based on real data analysis to calibrate parameters of the model considered. We can mention Bonneuil (2021), who uses follow-up data collected by the French Exceptional Health Situations[6] at peak dates of inflow for Covid-19 to estimate rates of mortality, rates of transfer to ventilators and rates of "returning home". Meisami et al. (2019) work on a data-set that spans two years, with more than 200,000 data points representing over 70,000 distinct patients, in order to calibrate the transition probabilities as well as the distribution of arrivals. Mayhew and Smith (2008) use Nu-Care data to calibrate a model that relates actual average completion times to the national target. They based their overall findings on work-flow data involving around 150,000 observations over an extended 4-year period. Olmos and Borzone (2021) use data for daily new cases of Covid-19 published by the Chilean ministry of health in order to compute the best-fit exponential regression and apply their model. A second way to illustrate the effectiveness of the method is to find and use commonly known values from the related literature. This method has been used by Chen et al. (2021) who employ values parameters governing the transition probabilities in their design, invoking Acemoglu et al. (2020), Atkeson (2020), Eikenberry et al. (2020), and Verity et al. (2020), among others. And finally, the most simple way to demonstrate effectiveness is to arbitrarily choose wise parameters to construct the simulations and build fictive, but as realistic as possible, scenarios or counter-factual situations. For instance, Lee and Lee (2018) choose some figures of lesser importance such as the peak times of immediate or delayed arrival, the expected volume of patients from an incident, or the ratios between the classes of immediate and delayed patients. Although our model renders in the first instance certain fictional scenarios and comparative situations, at the end of the paper we employ real ICU data from a French hospital in order to calibrate the parameters of our model.

The remainder of the present paper is structured as follows. In Section 2, we introduce the elements of the model. We use simulated trajectories in order to illustrate what we call the "time-shortage". The underlying Markov process is then studied in more depth so as to obtain the computation of the distribution of the time-shortage. In Section 3, we define the notion of "profile" and propose to extend the model from one profile to two profiles of patients in a very similar way. The type of applications of such a model are explored in Section 4. We then comment on the traffic signals plots and density/cumulative/quantile distribution curves. We also construct simulations of fictive scenarios and make comparisons between real/fictional situations, before looking at an application with real data from an ICU in France. Section 5 concludes.

[6] Better known in French as CIVIC.

2 The model

The probabilistic approach we adopt in order to study our phenomenon is Markov chain theory. As a recurrent time process, the total number of patients turns out to be a classical Markov chain, but if we add some necessary assumptions, we can guarantee that the chain is homogeneous. The required hypotheses that allow this property to be satisfied are the following:

- Each term of the considered time sequence depends only on the previous term of this sequence and of other independent events. If we force the probabilities of entering and leaving the service to not depend on time, this ensures that the process of switching from a certain time to the next one is always the same, no matter what moment it takes place. In other words, moving from time 1 to time 2 occurs in the exact same way as moving from time 100 to time 101.
- All times of arrivals of all patients must be independent. This means that the moment a patient arrives does not impact the moment any other patient arrives.
- All times of discharges of all patients must be independent. This means that the moment a patient leaves the service does not impact the moment any other patient leaves.
- For a given patient, the moment he/she arrives is independent from the moment he/she leaves. This means that the moment a patient arrives does not impact the moment he/she leaves and vice versa.

In this section, we introduce our model in a formal and technical way. The case of a "single profile" is presented first, before generalizing to two profiles and hence to n profiles (see Appendix B).

2.1 The case of one profile

2.1.1 Basic notations and the process $(X_t)_t$ of the total number of patients

The first element we need in order to set up a time process is, of course, a time step $t = 0, 1, ..., \infty$ accounting for the moment or the instant considered in time. The unit of time can be a minute, an hour, a day, etc., as long as it meets a certain condition presented later. The duration of the study is theoretically infinite, in order to witness a shortage event in any possible situation. The model has three parameters, as already mentioned:

- $p \in [0, 1]$ is the probability of arrival of a patient at each time step.
- $q \in [0, 1]$ is the probability of discharge of a patient at each time step.
- $Q \in \mathbb{N}^*$ is the total number of ventilators in the service that are functional and able to be offered to patients in need.

Hence, in our model, the triplet $(p, q, Q) \in (0,1) \times (0,1) \times \mathbb{N}^*$ characterizes a single and unique situation of the flows in and out of an ICU, and these are all the switches and levers we can adjust in order to fit the model as faithfully as possible to a concrete real situation. It is paramount to highlight that this triplet is constant with time. Concerning the random variables that represent the patients, we use the following notations:

- $X_t \in \{0, \dots, Q\} := E$ is the total number of patients at time t. We shall always denote x_0 the number of patients in the service at the initial time, so that $X_0 = x_0$. Since there cannot be more patients than the total number of ventilators, the condition $X_t \leq Q$ holds.[7] Since the number of patients is positive, $X_t \geq 0$.
- $Y_t \in \{0, 1\}$ represents that a patient requiring a ventilator either arrives (1) or does not arrive (0) at time t. Hence, all $(Y_t)_t$ follow Bernoulli distributions with parameter p and are mutually independent. The condition $X_t \leq Q$ means that once the maximum of Q patients is reached, we cannot have an additional patient admitted, which can be expressed as $\mathbb{P}(Y_t = 1 | X_t = Q) = 0$ or similarly $\mathbb{P}(Y_t = 0 | X_t = Q) = 1$.
- $Z_t \in \{0, 1\}$ represents that a currently admitted patient either leaves the service (1) or does not leave (0) at time t, thus releasing a ventilator for an unknown reason which could be either recovery or death. Hence, all $(Z_t)_t$ follow Bernoulli distributions with parameter q and are independent between them. The condition $X_t \geq 0$ means that once the minimum of 0 patients is reached, we cannot have any patient leaving, which can be expressed as $\mathbb{P}(Z_t = 1 | X_t = 0) = 0$ or similarly $\mathbb{P}(Z_t = 0 | X_t = 0) = 1$.

As stated at the beginning of this section,

- all arrivals are mutually independent: $\forall t, t' \in \mathbb{N},\ t \neq t' \implies Y_t \perp\!\!\!\perp Y_{t'}$,
- all discharges are mutually independent: $\forall t, t' \in \mathbb{N},\ t \neq t' \implies Z_t \perp\!\!\!\perp Z_{t'}$, and
- rates of flows are constant over time: $\forall t \in \mathbb{N},\ \mathbb{P}(Y_t = 1) = p$ and $\mathbb{P}(Z_t = 1) = q$.

Accounting for the total number of patients at time $t+1$ is simply done by adding the number of new patients and removing the number of patients leaving at time t:

$$\forall t \in \mathbb{N},\ X_{t+1} = X_t - Z_t + Y_t.$$

Definition 1. *The time series of the total number of patients $(X_t)_t$ can be defined as follows:*

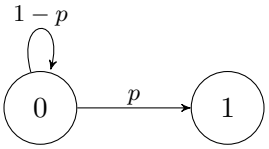
$$\forall t \in \mathbb{N},\ \begin{cases} X_0 = x_0 \\ X_{t+1} = X_t - Z_t + Y_t \\ 0 \leq X_t \ \leq Q \end{cases}$$

[7] Actually, without this condition, the chain would not be bounded and would have an infinite state space. In this case, some of theoretical results stated later would no longer hold, but as regards our main concern this would not be problematic., this would not be problematic.

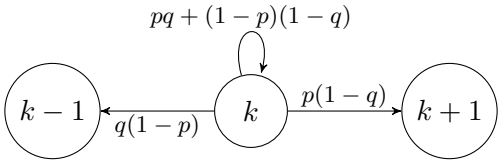
2.1.2 Properties of the Markov chain $(X_t)_t$

Let us now study the Markov chain $(X_t)_t$. The values taken by the chain are called "states" so that the space of states here is $E = \{0, \dots, Q\}$. We can proceed to the calculations of the transition probabilities $p_{i,j} = \mathbb{P}(X_{t+1} = j | X_t = i) = \mathbb{P}(X_1 = j | X_0 = i)$, i.e., the probability of switching from state i to state j with $i, j \in E$ in a time step. In other words, $p_{i,j}$ is the probability of having j patients in the service at time $t+1$ knowing there were i patients at time t. As we saw earlier, this probability does not depend on t since the chain is homogeneous and, hence, does not need to be indexed by t.

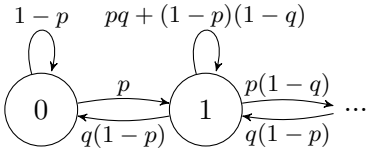
Starting from state 0 (i.e., knowing that there is no patient in the system), two situations can occur: we can only move to state 1 or stay in state 0. To move to state 1 we need to gain one patient, which happens with probability p, hence $p_{0,1} = p$. Staying at state 0 (i.e., not admitting a new patient) happens with probability $1 - p$, hence $p_{0,0} = 1 - p$. This reasoning can be summarized in the following graph:

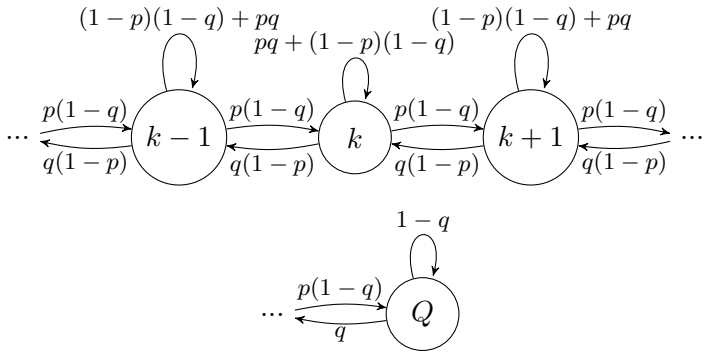


More generally, from a given state k, we can have a transition to states k, $k+1$, or $k-1$. To move to state $k+1$, we need to gain one patient with probability p but we also need to release nobody with probability $1 - q$, so the combination occurs with probability $p(1-q) := p_{k,k+1}$. To move to state $k-1$, we need to release one patient with probability q but we also need not to admit a new one with probability $1 - p$, so the combination occurs with probability $q(1-p) := p_{k,k-1}$. The remaining probability $p_{k,k}$ can be calculated in two ways. We can first use the fact that $p_{k,k} + p_{k,k+1} + p_{k,k-1} = 1$. The second way is to consider that we need to gain one patient and to release another one at the same time, which occurs with probability pq, or not to gain one and not to release one either, which occurs with probability $(1-p)(1-q)$. In both cases, we should have $pq + (1-p)(1-q) := p_{k,k}$. This discussion can be summarized again as follows:



Now we can draw the whole transition graph showing all states of the Markov chain and all transition probabilities. It has the following form:





Equivalently, the matrix filled with $p_{i,j}$ in row i and in column j, denoted by $P = (p_{i,j})$, is called the "transition matrix". Each row sums to 1 since each row is the probability distribution of transition from state i (conditional distribution).

$$P(p,q) = \begin{pmatrix} 1-p & p & 0 & & (0) \\ q(1-p) & (1-p)(1-q)+pq & p(1-q) & & \\ & \ddots & \ddots & \ddots & \\ & & q(1-p) & (1-p)(1-q)+pq & p(1-q) \\ (0) & & \dots & q & 1-q \end{pmatrix}$$

Finally, and again equivalently, the following proposition holds:

Proposition 1. *For all $i, j \in E$, we have*

$$p_{i,j} = \begin{cases} (1-p)(1-q)+pq & \text{if } j = i \\ q(1-p) & \text{if } j = i-1 \\ p(1-q) & \text{if } j = i+1 \\ 0 & \text{otherwise} \end{cases}$$

with exceptions $p_{0,0} = 1-p$, $p_{0,1} = p$, $p_{Q,Q-1} = q$ and $p_{Q,Q} = 1-q$.

Proof. See Appendix D. □

This result can be extended to more general discrete probability distributions on Y and Z (see Appendix C). It is now clear that the chain is irreducible, i.e., that there is a non-zero probability to transit from any state to any other state. Indeed, there is a non-zero probability to transit to an adjacent state (a neighbor state that is the next, the previous, or the same one), so there exists a strictly positive chance to transit from one state to another one in multiple steps. Once we have the transition matrix, we can compute the distribution of X_t for all t as follows: if we denote by

$$\mu_t = (\mathbb{P}(X_t = i))_{i \geq 0},$$

for all t, and if $\mu_0 = (0, 0, \dots, 1, 0, \dots, 0)$ (with 1 in position x_0) is the initial distribution of the chain, i.e., absolute certainty to have x_0 patients at the beginning of the study, then

we have the following fundamental relationship:

$$\mu_t = \mu_0 P^t,$$

where P^t is the matrix P raised to the power t (Zukerman, 2013). Hence, iterating the multiplication by the transition matrix P gives the distribution of the chain for each consecutive moment. For instance, $\mu_{1000,10} = \mathbb{P}(X_{1000} = 10)$ is the probability of having $i = 10$ patients in ICU at time $t = 1000$ and it can be computed by $\mu_{1000,10} = \mu_0 \times P^{1000}$.

Definition 2. *If it exists and is unique, the solution $\pi \in \mathbb{R}^{Q+1}$ of the system $\pi P = \pi$ is called the stationary distribution or invariant distribution of the chain X, where P is the transition matrix of the Markov chain $(X_t)_t$.*

This particular distribution is of interest since the process remains stable once it is nearly reached. This is what we could call an attractive point for the sequence $(X_t)_t$. In case of an irreducible chain with finite space states (and hence positive recurrent), π exists and is unique, which is true in our case. Let us then compute it.

Lemma 1. *The stationary distribution π of the chain $(X_t)_t$ exists and is unique. Moreover, we have $\pi_i = \frac{1}{Q+1}$ for all $i \in E$.*

Proof. See Appendix E □

This means that π follows a uniform distribution over space states $E = \{0, ..., Q\}$. Furthermore, the chain distribution of $(X_t)_t$ converges to π: for all $i \in E$,

$$\lim_{t \to \infty} \mathbb{P}(X_t = i) = \pi_i = \frac{1}{Q+1},$$

and this is true independently from μ_0, the initial distribution of X_0. In other words, no matter how the chain starts, if we wait an infinite amount of time, we have the same probability of having any number of patients between 0 and Q. Hence, after sufficient time has passed, there is no reason to believe that the situation will tilt towards a brighter or a darker side. A last, but important, result regarding the stationary distribution is the following.

Definition 3. *The first hitting time to state $i \in E$ is defined as $T_i := \min\{t \in \mathbb{N}^* \ / X_t = i\} \in \mathbb{N}^*$. We furthermore denote by $\mathbb{E}_i(Y) := \mathbb{E}(Y|X_0 = i)$ the expectancy of a random variable Y knowing that $X_0 = i$, i.e., starting from state i at initial time 0.*

The following important result holds.

Proposition 2. *For all $i \in E$, $\mathbb{E}_i(T_i) = Q + 1$. In particular, $\mathbb{E}_Q(T_Q) = Q + 1$.*

Proof. Since all states $i \in E$ of the chain are positive recurrent, $T_i < +\infty$, and we have the following link between the stationary distribution and the expected return time in this case (Zukerman, 2013): $\mathbb{E}_i(T_i) = \frac{1}{\pi_i}$. The result is immediate by Lemma 1. □

This proposition implies that returning to a previously visited state takes $Q+1$ time steps on average. So, if a critical situation occurs, we should expect it to occur again $Q+1$ time steps later.

2.1.3 Time-shortage t^*

Now that the whole framework is set, we can finally introduce the principal random variable targeted by of this model: the time-shortage.

Definition 4. *We call, "time-shortage", the random variable* $t^* = \min\{t \in \mathbb{N}^* \ / X_t = Q\}$.

This variable corresponds to the first moment all of the Q ventilators are occupied and not one is left for eventual future incoming patients. We provide in Figure 1 a random realization of a trajectory $(X_t)_{0\leq t\leq 100}$ for the set of parameters $p = 0.5$, $q = 0.4$ and $Q = 5$. Since $p > q$, X_t tends to increase. At time $t^* = 21$, the process for the first reaches the maximum possible number of patients $Q = 5$. In theory, the sequence continues to evolve and on average, keeps increasing, but in reality the threshold Q cannot be exceeded. This is the case because we posit that a patient is refused for treatment if he/she arrives when the service is already full. Thus it was common practice in France during the Covid-19 crisis to reject patients[8] or at least to impose more restrictive conditions of admission under that circumstance. One can see from Figure 1 that the situation eases for a little while before reaching new critical points.

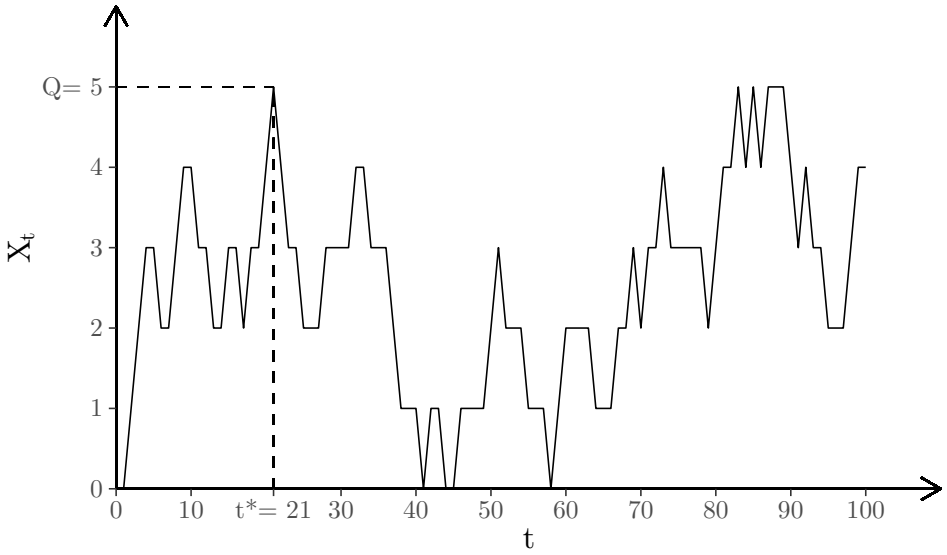


Figure 1: X_t as a function of t with $p = 0.5$, $q = 0.4$, and $Q = 5$.

[8] Zagdoun, B. and Lecouvé, P. (2022) Covid-19: peut-on dire qu'il n'y a pas eu de tri des patients à l'hôpital, comme l'affirme Emmanuel Macron ?
https://www.francetvinfo.fr/sante/maladie/coronavirus/vrai-ou-fake-covid-19-peut-on-dire-qu-il-n-y-a-pas-eu-de-tri-des-patients-a-l-hopital-comme-l-affirme-Emmanuel-macron_4908113.html.

Running this simulation many times would compute many trajectories for $(X_t)_t$ and hence many realizations for t^*. The histogram of these realizations should give us an idea of the probability distribution of t^*. This would be the basic standard approach to working out the distribution by means of simulations. The choice made in our paper, however, is to leave this method aside and focus on a more accurate approach, which is to find and compute the exact distribution of t^*, at least numerically speaking if no explicit formulas can be proved. If this can be done, the calculations and graphs made will be more reliable, and they will also be quicker to obtain numerically than via repeated simulations. When the process being studied is not too complicated and is not marked by too much uncertainty or too many random phenomena, an exact computation of the distribution should always be privileged over a simulation approach. The definition of t^* requires some mathematical explanations regarding its existence, which is not obvious.

Proposition 3. *The random time-shortage t^* exists almost surely.*

Proof. The chain $(X_t)_t$ is aperiodic, irreducible and with finite state space,[9] so that all spaces are positive recurrent. Hence, all states are visited infinitely many times with a non-zero probability, in particular state Q is visited infinitely many times and there exists $t \in \mathbb{N}^*$ such that $\mathbb{P}(X_t = Q) > 0$. □

This proposition ensures that all states are visited an infinitely many times with a non zero probability, and in particular that state Q is visited infinitely many times. However, this is not necessarily the case for a finite time horizon $\{0, \dots, N\}$. That is the reason why choosing a sufficiently large N is essential to be able to witness a shortage event when applying the model to concrete situations. Note that the existence of the time-shortage is probabilistic: it has a non-zero chance to exist at every moment, although this can be very small. In practice, we will see later that when $q > p$, it is the case.

Definition 5. *For all $i, j \in E$, the random variable $T_{i,j} := \min\{t \in \mathbb{N}^* \ / X_t = j, X_0 = i\}$ is the first time state j is reached given that the chain was in state i in the past. We denote $f_{i,j}^{(t)} := \mathbb{P}(T_{i,j} = t) = \mathbb{P}(X_t = j, X_{t-1} \neq j, \dots, X_1 \neq j \mid X_0 = i)$ the associated probabilities for all $i, j \in E$ and all $t \in \mathbb{N}^*$ and we call these numbers "first passage times" .*

In our case, the first passage time we are interested in is $t^* = T_{x_0,Q}$, that is the first time the total number of ventilators Q is reached knowing that there were x_0 patients in ICU at the initial time. Hence, for all $t \in \mathbb{N}^*$, $\mathbb{P}(t^* = t) := f_{x_0,Q}^{(t)}$. To achieve our goal, we just need to compute these probabilities, which can be done thanks to the following recursive relationship due to Neuts in 1973 (see, for instance, Alfa, 2016; Hillier and Lieberman, 2001). For all $i, j \in E$, we have:

$$\begin{cases} f_{i,j}^{(1)} = p_{i,j}, \\ f_{i,j}^{(t)} = \sum_{k \neq j} p_{i,k} f_{k,j}^{(t-1)} \ \forall t \in \mathbb{N}^*. \end{cases} \tag{1}$$

[9] We refer the reader to, for instance, Zukerman (2013).

The value $f_{x_0,Q}^{(t)}$ is easy to evaluate once this calculation is carried out, as well as the mean time-shortage given by $\mathbb{E}(t^*) = \sum_{t=1}^{\infty} t\mathbb{P}(t^* = t) = \sum_{t=1}^{\infty} t f_{x_0,Q}^{(t)}$. In Appendix A we provide the detailed pseudo-codes implemented to compute these time-shortage probabilities. Figure 2 displays the probability densities of the variable t^* as a function of time, with parameter q fixed for each graph, and p varying. All densities are increasing, reaching a mode and then decreasing, but the higher p, the more likely we are to observe a shortage happening very soon and the less likely to face it late. For lower values of p, the opposite, but but analogous reasoning still stands, since the density function flattens softly, meaning that we have lesser but still equal chances to face a shortage in the long term. For $p < q$, the chances of shortage are so small that the densities are far from complete and we only see partial and low-valued distribution functions.

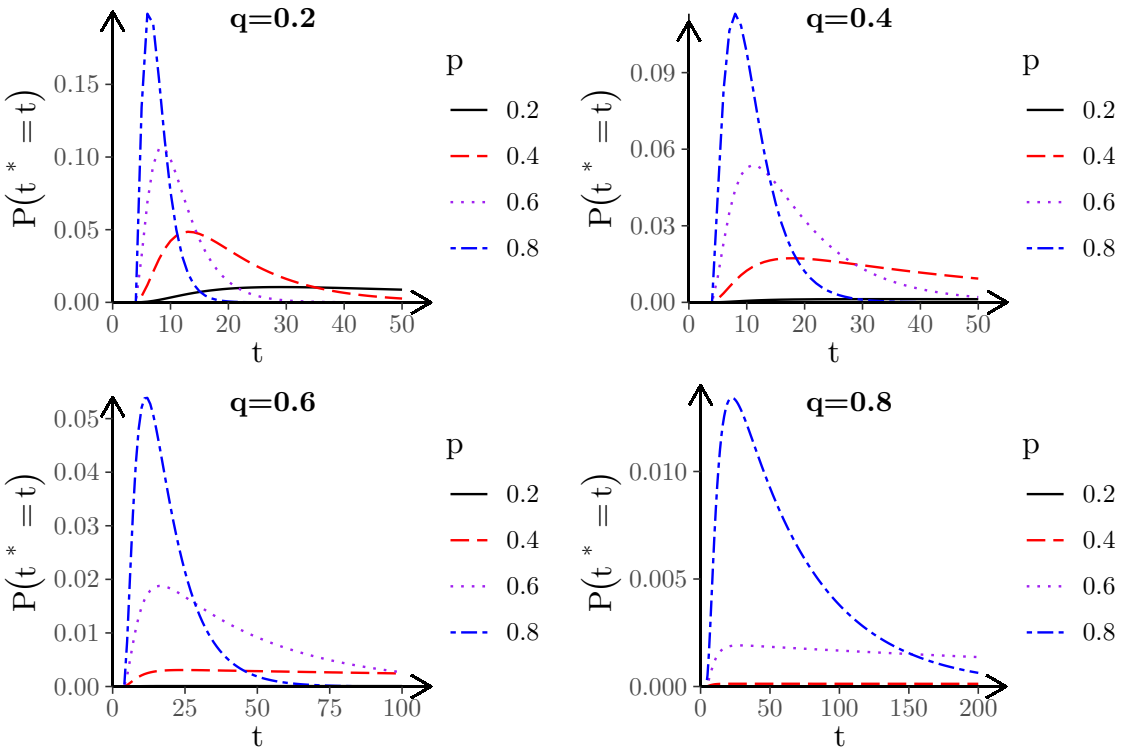


Figure 2: Probability density function of t^* as a function of t with $Q = 5$.

We can see from Figure 3 that, naturally, the fewer ventilators there are, the more critical the situation is and the more likely that a shortage will occur soon. However, it is interesting to note that after some time (after the modes), the tendencies tend to swap levels of severity and, with fewer ventilators, we have no further chance of experiencing a shortage (because it probably already happened), whereas with more ventilators the risk remains higher and non-negligible, slowly fading away with time. From one plot to another, we can see that the shapes of the density functions do not change much, essentially because the stress that the ICU undergoes is more or less the same with the ratio $\frac{p}{q}$ being

constant.

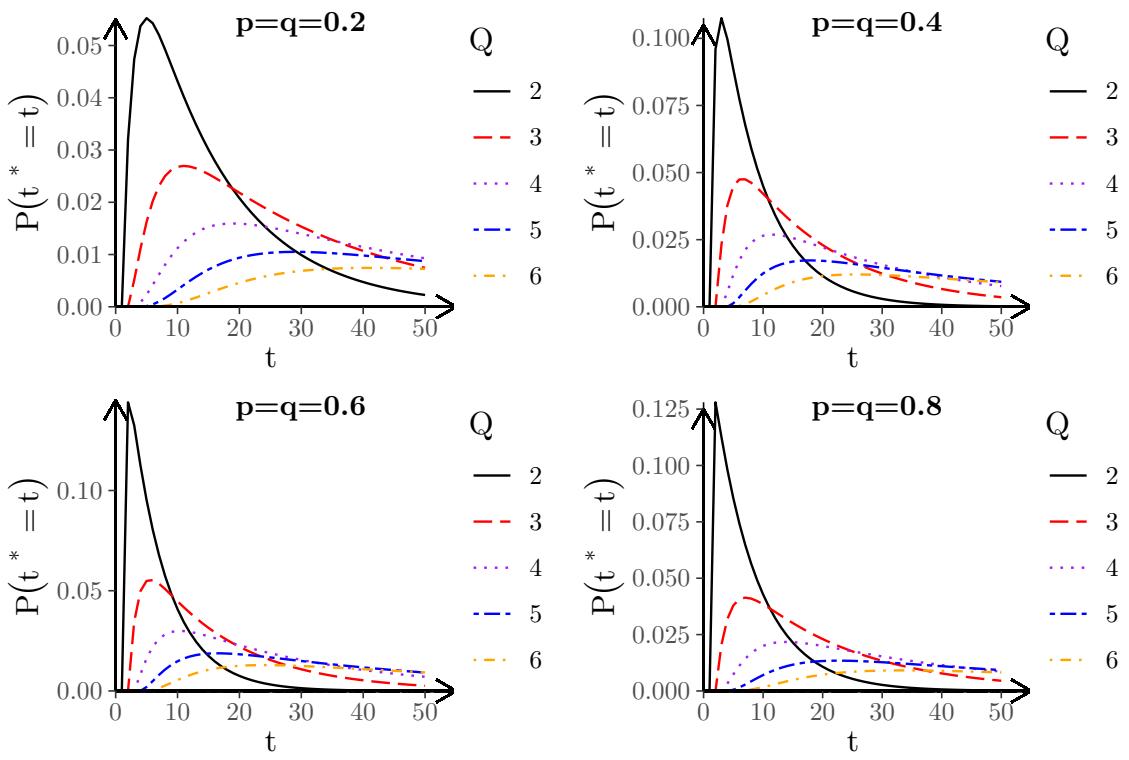


Figure 3: Probability density function of t^* as a function of t.

2.1.4 Further results

Proposition 4 (Sufficient condition for no shortage)**.** *For all $x_0, Q, t \in \mathbb{N}^*$, we have*

$$Q - x_0 > t \implies f_{x_0,Q}^{(t)} = 0.$$

Proof. We adopt the more convenient notations $i := x_0$ and $j := Q$. We want to show that $j - i > t \implies f_{i,j}^{(t)} = 0$. Let us proceed by inductive reasoning on integer t.
For $t = 1$, $j > i + 1 \implies f_{i,j}^{(1)} = p_{i,j} = 0$ by assumption on transition matrix P. Now, suppose that the result holds for a fixed $t \in \mathbb{N}^*$. Let us take $j - i > t + 1$. By the recursive formula defining the sequence $(f^{(t)})_t$,

$$f_{i,j}^{(t+1)} = p_{i,i-1} f_{i-1,j}^{(t)} + p_{i,i} f_{i,j}^{(t)} + p_{i,i+1} f_{i+1,j}^{(t)}$$

The induction hypothesis ensures that

- $j - (i-1) = j - i + 1 > t + 1 + 1 = t + 2 > t \implies f_{i-1,j}^{(t)} = 0$,
- $j - i > t + 1 > t \implies f_{i,j}^{(t)} = 0$, and
- $j - (i+1) = j - i - 1 > t + 1 - 1 = t \implies f_{i+1,j}^{(t)} = 0$.

Hence, $f_{i,j}^{(t+1)} = 0$ which ends the inductive reasoning and the proof. □

Hence, if the number of ventilators Q is high enough, or if the initial number of patients x_0 is small enough, or if the horizon of time considered is short enough, then the probability of shortage is null. The intuitive reason is that it is impossible, when transiting from one to one, to transit from x_0 to Q in less than $Q - x_0$ periods.

Proposition 5 (Maximum stress)**.** *For $Q > x_0$, we have*

$$f_{x_0,Q}^{(Q-x_0)} = \prod_{k=x_0}^{Q-1} p_{k,k+1} = (p(1-q))^{Q-x_0}$$

Proof. We adopt the more convenient notations $i := x_0$ and $j := Q$. We want to show that $j - i > 0 \implies f_{i,j}^{(j-i)} = \prod_{k=i}^{j-1} p_{k,k+1}$. Let us proceed by inductive reasoning on integer $j - i$.
For $j - i = 1$, $f_{i,i+1}^{(1)} = p_{i,i+1} = p(1-q)$ by assumption on transition matrix P. Now, suppose that the result holds for a fixed integer $j - i$. By the recursive formula defining the sequence $(f^{(t)})_t$,

$$f_{i,j+1}^{(j-i+1)} = p_{i,i-1} f_{i-1,j+1}^{(j-i)} + p_{i,i} f_{i,j+1}^{(j-i)} + p_{i,i+1} f_{i+1,j+1}^{(j-i)}$$

Previous proposition 4 ensures that

- $j + 1 - (i - 1) = j + 1 - i + 1 = j - i + 2 > j - i \implies f_{i-1,j+1}^{(j-i)} = 0$,
- $j + 1 - i > j - i \implies f_{i,j+1}^{(j-i)} = 0$

, and the induction hypothesis states that $j + 1 - (i + 1) = j - i \implies f_{i+1,j+1}^{(j-i)} = \prod_{k=i+1}^{j} p_{k,k+1}$. Hence,

$$f_{i,j+1}^{(j-i+1)} = p_{i,i+1} \prod_{k=i+1}^{j} p_{k,k+1} = \prod_{k=i}^{j} p_{k,k+1}$$

which ends the inductive reasoning. Since $p_{i,i+1} = p(1-q)$, we have the second equality and the proof is complete. □

This means that to switch from x_0 patients to Q patients in exactly $Q - x_0$ time steps, there is no choice but to gain one patient at each time step which corresponds to a situation of constant maximum tension.

Theorem 1 (Shortage time as a function of Q)**.** *Suppose there exists some time $T \in \mathbb{N}^*$ such that $\left(f_{x_0,Q}^{(T)}\right)_{Q>x_0}$ is decreasing. Then, for all $t \in \mathbb{N}$, the sequence $\left(f_{x_0,Q}^{(t+T)}\right)_{Q>x_0+t}$ is decreasing.*

Proof. We adopt the more convenient notation $i := x_0$ and $j := Q$. We want to show $j - i > t \implies f_{i,j+1}^{(t+T)} \leq f_{i,j}^{(t+T)}$. Let us proceed by inductive reasoning on integer t.

For $t = 0$, this is obviously true because of the hypothesis made.
Assume now that the result holds for a fixed $t \in \mathbb{N}$. Let us take $j - i > t + 1$. We have

$$\begin{aligned} f_{i,j+1}^{(t+T+1)} &= p_{i,i-1} f_{i-1,j+1}^{(t+T)} + p_{i,i} f_{i,j+1}^{(t+T)} + p_{i,i+1} f_{i+1,j+1}^{(t+T)} \\ &\leq p_{i,i-1} f_{i-1,j}^{(t+T)} + p_{i,i} f_{i,j}^{(t+T)} + p_{i,i+1} f_{i+1,j}^{(t+T)} \\ &= f_{i,j}^{(t+T+1)}. \end{aligned}$$

By the induction hypothesis, we have

- $j - (i-1) = j - i + 1 > t + 1 + 1 = t + 2 > t \implies f_{i-1,j+1}^{(t+T)} \leq f_{i-1,j}^{(t+T)}$,
- $j - i > t + 1 = t + 1 > t \implies f_{i,j+1}^{(t+T)} \leq f_{i,j}^{(t+T)}$, and
- $j - (i+1) = j - i - 1 > t + 1 - 1 = t \implies f_{i+1,j+1}^{(t+T)} \leq f_{i+1,j}^{(t+T)}$.

This ends the induction and the proof. □

This shows that the more ventilators we have at disposal, the more unlikely the ICU will face a shortage situation. This result is to be tempered with the fact it requires this phenomenon to happen at some particular time T to be true for all time greater than T. Obviously, this result holds for $T = 1$ since $(f_{x_0,Q}^{(1)})_{Q>x_0} = (p_{x_0,Q})_{Q>x_0}$ but mainly states that the null sequence is decreasing thanks to Proposition 4. That is why a sufficiently large T will be wanted to guarantee the relevance of this property.

Without any other hypothesis, it is difficult to prove any stronger result. For example, we can show that if $p_{i+1,i+2} > p_{i,i}$, then $f_{i,i+1}^{(2)} < f_{i,i+2}^{(2)}$ such that the sequence $(f_{i,j}^{(2)})_{j>i}$ is no longer decreasing.

Theorem 2 (Shortage time as a function of x_0)**.** *Suppose there exists some time $T \in \mathbb{N}^*$ such that $\left(f_{x_0,Q}^{(T)}\right)_{x_0<Q}$ is increasing. Then, for all $t \in \mathbb{N}$, the sequence $\left(f_{x_0,Q}^{(t+T)}\right)_{x_0<Q-t}$ is increasing.*

Proof. We adopt the more convenient notation $i := x_0$ and $j := Q$. We want to show that $j - i > t \implies f_{i,j}^{(t+T)} \leq f_{i+1,j}^{(t+T)}$. Let us proceed by inductive reasoning on integer t.
The case $t = 0$ is obvious because of the hypothesis made. Assume now that the result holds for a fixed $t \in \mathbb{N}$. Let us take $j - i > t + 1$. We have

$$\begin{aligned} f_{i,j}^{(t+T+1)} &= p_{i,i-1} f_{i-1,j}^{(t+T)} + p_{i,i} f_{i,j}^{(t+T)} + p_{i,i+1} f_{i+1,j}^{(t+T)} \\ &\leq p_{i,i-1} f_{i,j}^{(t+T)} + p_{i,i} f_{i+1,j}^{(t+T)} + p_{i,i+1} f_{i+2,j}^{(t+T)} \\ &= p_{i+1,i} f_{i,j}^{(t+T)} + p_{i+1,i+1} f_{i+1,j}^{(t+T)} + p_{i+1,i+2} f_{i+2,j}^{(t+T)} \\ &= f_{i+1,j}^{(t+T+1)}. \end{aligned}$$

Indeed, we were able to apply the induction hypothesis since

- $j - (i-1) = j - i + 1 > t + 2 > t \implies f_{i-1,j}^{(t+T)} \leq f_{i,j}^{(t+T)}$,

- $j - i > t + 1 > t \implies f_{i,j}^{(t+T)} \leq f_{i+1,j}^{(t+T)}$, and
- $j - (i+1) = j - i - 1 > t + 1 - 1 = t \implies f_{i+1,j}^{(t+T)} \leq f_{i+2,j}^{(t+T)}$.

We also used for the last equality the fact that the sequences $(p_{i,i-1})_i$, $(p_{i,i})_i$ and $(p_{i,i+1})_i$ are constant. This ends the induction and the proof. □

Hence, the higher the number of patients in the ICU at the beginning of the study, the greater the chances to expect a shortage. This result should be moderated by the fact that it holds until a certain rank decreasing as t grows. The value of T, once again, needs to be large enough to bring substantial information. The case $T = 1$ is true but irrelevant.

Proposition 6 (Shortage time as a function of p)**.** *Suppose there exists some time $T \in \mathbb{N}^*$ such that $\left(f_{x_0,Q}^{(T)}\right)_{x_0<Q}$ is increasing and that the function $p \mapsto f_{x_0,Q}^{(T)}$ is increasing for $x_0 < Q$. Then, the function $p \mapsto f_{x_0,Q}^{(t+T)}$ is increasing for $Q - x_0 > t$.*

Proof. We adopt the more convenient notation $i := x_0$ and $j := Q$. We want to show that $j - i > t \implies \dfrac{\partial f_{i,j}^{(t+T)}}{\partial p} \geq 0$. Let us proceed by induction reasoning on integer t.
For $t = 0$, this is the hypothesis made. Assume now that the result holds for a fixed $t \in \mathbb{N}$. Let us take $j - i > t + 1$. We have

$$\begin{aligned} f_{i,j}^{(t+T+1)} &= p_{i,i-1} f_{i-1,j}^{(t+T)} + p_{i,i} f_{i,j}^{(t+T)} + p_{i,i+1} f_{i+1,j}^{(t+T)} \\ &= q(1-p) f_{i-1,j}^{(t+T)} + [(1-p)(1-q) + pq]\, f_{i,j}^{(t+T)} + p(1-q) f_{i+1,j}^{(t+T)} \end{aligned}$$

so that

$$\begin{aligned} \frac{\partial f_{i,j}^{(t+T+1)}}{\partial p} &= -q f_{i-1,j}^{(t+T)} + q(1-p) \frac{\partial f_{i-1,j}^{(t+T)}}{\partial p} \\ &\quad + (2q-1) f_{i,j}^{(t+T)} + [(1-p)(1-q) + pq] \frac{\partial f_{i,j}^{(t+T)}}{\partial p} \\ &\quad + (1-q) f_{i+1,j}^{(t+T)} + p(1-q) \frac{\partial f_{i+1,j}^{(t+T)}}{\partial p} \end{aligned}$$

The induction hypothesis applies and shows that

- $j - (i-1) = j - i + 1 > t + 1 + 1 = t + 2 > t \implies \dfrac{\partial f_{i-1,j}^{(t+T)}}{\partial p} \geq 0$,
- $j - i > t + 1 > t \implies \dfrac{\partial f_{i,j}^{(t+T)}}{\partial p} \geq 0$ and
- $j - (i+1) = j - i - 1 > t + 1 - 1 = t \implies \dfrac{\partial f_{i+1,j}^{(t+T)}}{\partial p} \geq 0$

hence, the derived terms at the end of each line are positive. For the remaining terms (at the start of each line), notice that, thanks to the previous proposition 2 regarding the growth in x_0 of the shortage time, we have

- $j-(i-1)=j-i+1>t+2+1=t+3>t+1 \implies f_{i-1,j}^{(t+T)} \leq f_{i,j}^{(t+T)}$ and
- $j-(i+1)=j-i-1>t+1-1=t>t-1 \implies f_{i,j}^{(t+T)} \leq f_{i+1,j}^{(t+T)}$

That way,

$$-qf_{i-1,j}^{(t+T)}+(2q-1)f_{i,j}^{(t+T)}+(1-q)f_{i+1,j}^{(t+T)} \geq -qf_{i,j}^{(t+T)}+(2q-1)f_{i,j}^{(t+T)}+(1-q)f_{i,j}^{(t+T)}=0$$

We can conclude that $\dfrac{\partial f_{i,j}^{(t+T+1)}}{\partial p} \geq 0$ which ends the induction and the proof. □

We have the similar analogous proposition regarding the exit flow q:

Proposition 7 (Shortage time as a function of q)**.** *Suppose there exists some time $T \in \mathbb{N}^*$ such that $\left(f_{x_0,Q}^{(T)}\right)_{x_0<Q}$ is increasing and that the function $q \mapsto f_{x_0,Q}^{(T)}$ is decreasing for $x_0 < Q$. Then, the function $q \mapsto f_{x_0,Q}^{(t+T)}$ is decreasing for $Q - x_0 > t$.*

Proof. A simple copy of the previous proof swapping the roles of p and q, changing the signs, the inequalities and the words "increasing" by "decreasing" yields the claimed result. □

Hence, the higher the inflow of patients, the greater the chances of shortage. Similarly, the higher the outgoing rate, the lower the risks of shortage.

Proposition 8. *For $x_0, Q \in \mathbb{N}^*$ and all $k \in \mathbb{Z}$ such that $x_0 + k \geq 0$ and $Q + k > 0$, we have*

$$f_{x_0+k,Q+k}^{(t)} = f_{x_0,Q}^{(t)}$$

Proof. The process $W_t = X_t + k$ is also a homogeneous Markov chain since it verifies the same recursive relationship as $(X_t)_t$: $W_{t+1} = W_t + Y_t - Z_t$. Henceforth, W has the same distribution as X with translated support $[|k, Q+k|]$ and its transition matrix is identical with translated space states. Since first passage times depend only on transition probabilities, we have the result. □

That means that if the initial number of patients and the total number of ventilators are shifted by the same amount (whether it is a gain or a loss), then the shortage time remains exactly the same (and its distribution as well).

We want now to relax the strong assumption of homogeneity of the Markov chain. Suppose that the process is no longer homogeneous and that the rates of flows in the ICU change over time. We adopt the following notations.

Definition 6. *Denote by*

- *$p(t)$ the probability of arrival of a patient at time t*
- *$q(t)$ the probability of discharge of a patient at time t*

- $p_{i,j}(t) = \mathbb{P}(X_{t+1} = j \mid X_t = i)$, *the analogous notations of the transition probabilities at time t for the non homogeneous case*
- $P(t) := (p_{i,j}(t))_{i,j}$ *the transition matrix at time t*
- $f_{i,j}(n,m) = \mathbb{P}(X_n = j, X_{n-1} \neq j, ..., X_{m+1} \neq j \mid X_m = i)$, *the analogous notation for the first passage time for the non homogeneous case.*
- $F(n,m) := (f_{i,j}(n,m))_{i,j}$ *the first passage time matrix*

For a homogeneous Markov chain, we have the particular properties:

- $f_{i,j}(n,0) = f_{i,j}^{(n)}$, the classical first passage time or in a matrix formulation, $F(n,0) = F(n)$
- $f_{i,j}(m+1,m) = p_{i,j}(m)$ or in a matrix formulation, $F(m+1,m) = P(m)$
- $p_{i,j}(0) = p_{i,j}$, the classical transition probabilities or in a matrix formulation, $P(0) = P$

Proposition 9. *We have the following recursive relationship:*

$$f_{ij}(n,m) = \sum_{k \neq j} f_{kj}(n, m+1) p_{ik}(m)$$

Proof. Thanks to the law of total probability:

$$\begin{aligned}
f_{ij}(n,m) &= \mathbb{P}(X_n = j, X_{n-1} \neq j, ..., X_{m+1} \neq j \mid X_m = i) \\
&= \sum_{k \neq j} \mathbb{P}(X_n = j, X_{n-1} \neq j, ..., X_{m+1} = k \mid X_m = i) \\
&= \sum_{k \neq j} \mathbb{P}(X_n = j, X_{n-1} \neq j, ..., X_{m+2} \neq j \mid X_{m+1} = k, X_m = i) \mathbb{P}(X_{m+1} = k \mid X_m = i) \\
&= \sum_{k \neq j} \mathbb{P}(X_n = j, X_{n-1} \neq j, ..., X_{m+2} \neq j \mid X_{m+1} = k) p_{ik}(m) \\
&= \sum_{k \neq j} f_{kj}(n, m+1) p_{ik}(m)
\end{aligned}$$

where we used the relation

$$\mathbb{P}(A \cap B | C) = \mathbb{P}(A | B \cap C) \mathbb{P}(B | C)$$

□

Proposition 10. *In this framework, we can compute the shortage time distribution through this relation:*

$$\mathbb{P}(t^* = t) = f_{x_0,Q}(t, 0)$$

We can notice that computing $F(t,0)$ requires to compute $F(t,1)$,..., and so on until matrix $F(t,t-1) = P(t-1)$. If we suppose we have full knowledge regarding the future (but changing) probabilities of arrival and discharge, the whole sequence $(P(t))_{t\in\mathbb{N}}$ is known, there is no problem for assessing quickly $F(t,0)$ and the distribution of the shortage time.

2.2 The case of two non-simultaneous profiles

In this section, we provide a technical discussion of how to adapt this model to the case of two profiles (the most general case with n profiles is described in Appendix B). Suppose now we wish to add some diversity regarding the types of patients. For example, among Covid-19 patients, some of them were more vulnerable than others. Of course, considering only two types of patients might seem unrealistic, if the only features considered are the admission and leaving rates of these patients regardless of their age, their sex, their comorbidities, etc. However, it is not false that among general population, a few patients are severely diseased and a lot are slightly infected or even asymptomatic. Considering these commonly known health facts, we can imagine having two profiles of patients, one type corresponding to severely diseased patients, in high need of assistance whose life is in real danger, whereas we have a secondary type of patient who is ill enough to be admitted to ICU but still in a reasonable shape with significant chances of survival compared to the first profile.

Hence, let us consider two profiles indexed by 1 and 2. The admission rates are respectively denoted by p_1 and p_2 for profiles 1 and 2, the leaving rates by q_1 and q_2 respectively, the arrivals by Y^1 and Y^2 respectively, the discharges by Z^1 and Z^2 respectively and, last, the total number of patients by X^1 and X^2 respectively. We still set classical Bernoulli distributions as follows: $Y^1 \sim \mathcal{B}(p_1)$, $Y^2 \sim \mathcal{B}(p_2)$, $Z^1 \sim \mathcal{B}(q_1)$ and $Z^2 \sim \mathcal{B}(q_2)$ but, now, we need to condition by the events $Y^1Y^2 = 0$ and $Z^1Z^2 = 0$, which means we cannot at the same time have $Y_1 = 1$ and $Y_2 = 1$ (entries of both types), and the same holds for exits. The bivariate chain $Y = (Y^1, Y^2)$ has a joint distribution given in Table 1.

$Y^{(1)} \backslash\ Y^{(2)}$	0	1
0	$\alpha_0 := \frac{(1-p_1)(1-p_2)}{1-p_1p_2}$	$\alpha_1 := \frac{(1-p_1)p_2}{1-p_1p_2}$
1	$\alpha_2 := \frac{p_1(1-p_2)}{1-p_1p_2}$	0

Table 1: The joint distribution of $Y = (Y^1, Y^2)$.

Analogous reasoning stands for the bivariate chain $Z = (Z^1, Z^2)$ representing the exits which has the joint distribution given in Table 2.

$Z^{(1)} \setminus Z^{(2)}$	0	1
0	$\beta_0 := \frac{(1-q_1)(1-q_2)}{1-q_1 q_2}$	$\beta_1 := \frac{(1-q_1)q_2}{1-q_1 q_2}$
1	$\beta_2 := \frac{q_1(1-q_2)}{1-q_1 q_2}$	0

Table 2: The joint distribution of $Z = (Z^1, Z^2)$.

We can interpret the probabilities as follows: α_0 is the probability to admit no new patient, β_0 the probability to release no patient, α_1 is the probability to admit a patient of type 1 but not type 2, α_2 is the probability to admit a patient of type 2 but not type 1, β_1 is the probability to release a patient of type 1 but not type 2, and β_2 is the probability to admit a patient of type 2 but not type 1. The Markov chain X has two components $X = (X^1, X^2)$ accounting for the total number of patients of type 1 and 2. To summarize, the bi-dimensional model can be written as follows:

$$\forall t \in \mathbb{N}, \begin{cases} X_t = (X_t^1, X_t^2), \\ X_0 = (x_{0,1}, x_{0,2}), \\ X_{t+1} = X_t - Z_t + Y_t, \\ 0 \leq X_t^1 \text{ and } 0 \leq X_t^2 \text{ and } X_t^1 + X_t^2 \leq Q. \end{cases} \tag{2}$$

Figure 4 is a realization of a couple trajectory with $Q = 8$ ventilators and $x_{0,1} = x_{0,2} = 0$ patients at the start. The profile 1 contains severely diseased patients ($p_1 = 0.6$ and $q_1 = 0.4$) with a tendency to appear in large surges and to remain for a long period, whereas profile 2 contains patients with the opposite behavior ($p_2 = 0.4$ and $q_2 = 0.6$). The patients of type 2 fail to increase so that the total number of patients mainly comes from type 1, which is the more problematic group. At time $t^* = 18$, all of the 8 ventilators of the service are used and shortage cannot be avoided.

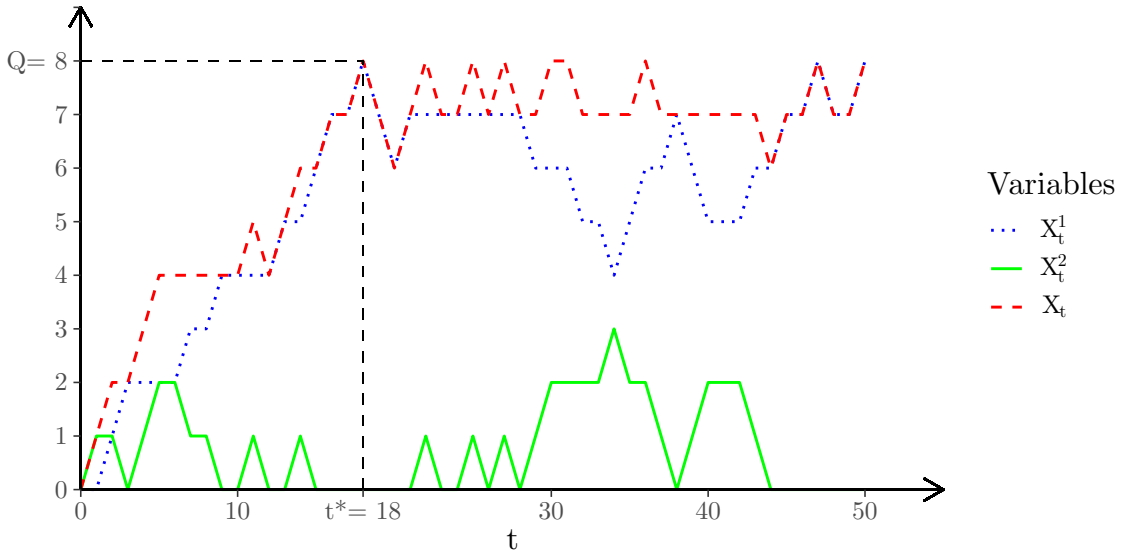
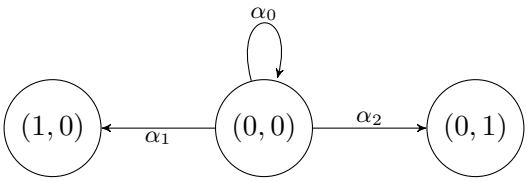
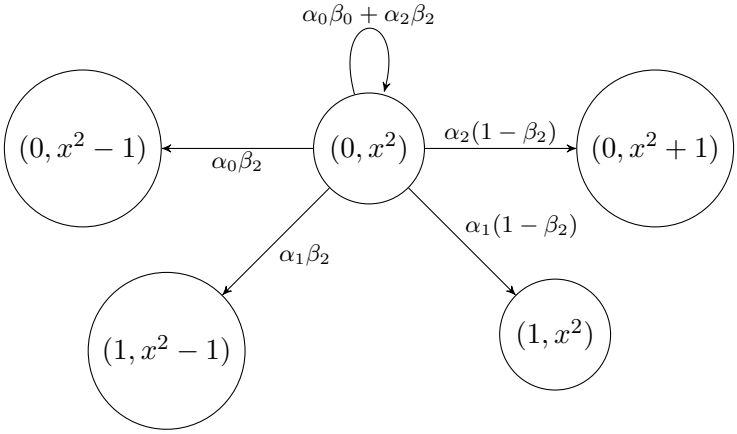


Figure 4: A 2-profile trajectory as a function of t with $p_1 = q_2 = 0.6$, $p_2 = q_1 = 0.4$, and $Q = 8$.

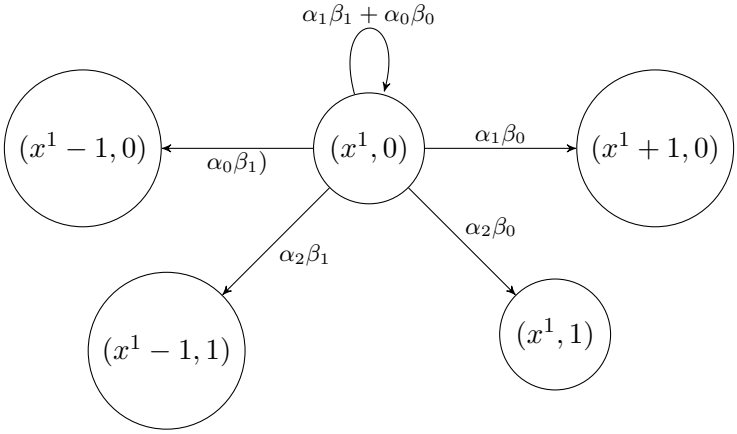
Here are the transition graphs of initialization. Starting from state $(0, 0)$, it is impossible to release patients since the service is empty, hence $\beta_1 = \beta_2 = 0$ and $\beta_0 = 1$.



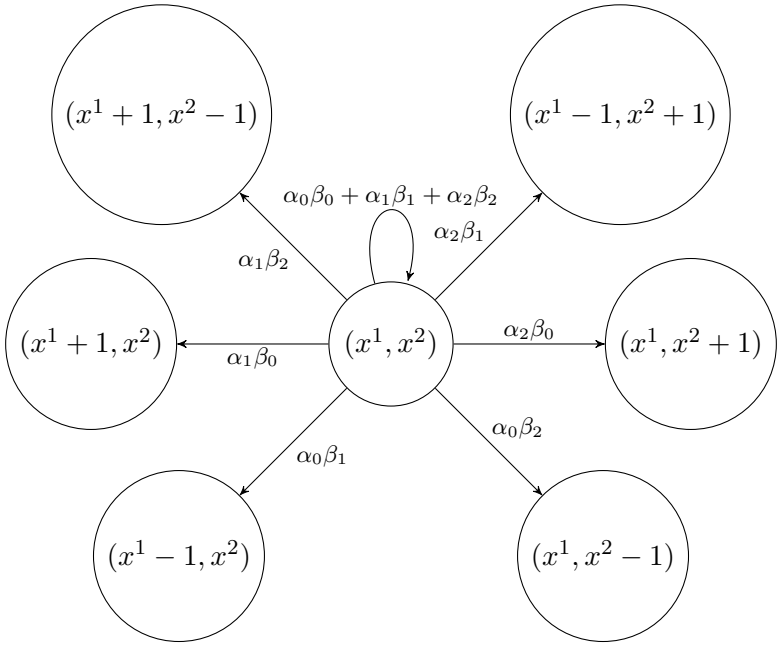
Starting from $(0, x^2)$, it is impossible to release patients of type 1, which means that $\beta_1 = 0$. Let us consider a few cases for the sake of better understanding. To move from situation $(0, x^2)$ to $(0, x^2 - 1)$, for instance, we need to admit no patient of type 1 (with probability α_0) and to release a patient of type 2 (with probability β_2) so that the overall probability is $\alpha_0\beta_2$. To move to $(1, x^2)$, we need to gain one patient of type 1 (with probability α_1) and not to release a patient of type 2 (with probability $1 - \beta_2$) so that the overall probability is $\alpha_1(1 - \beta_2)$.



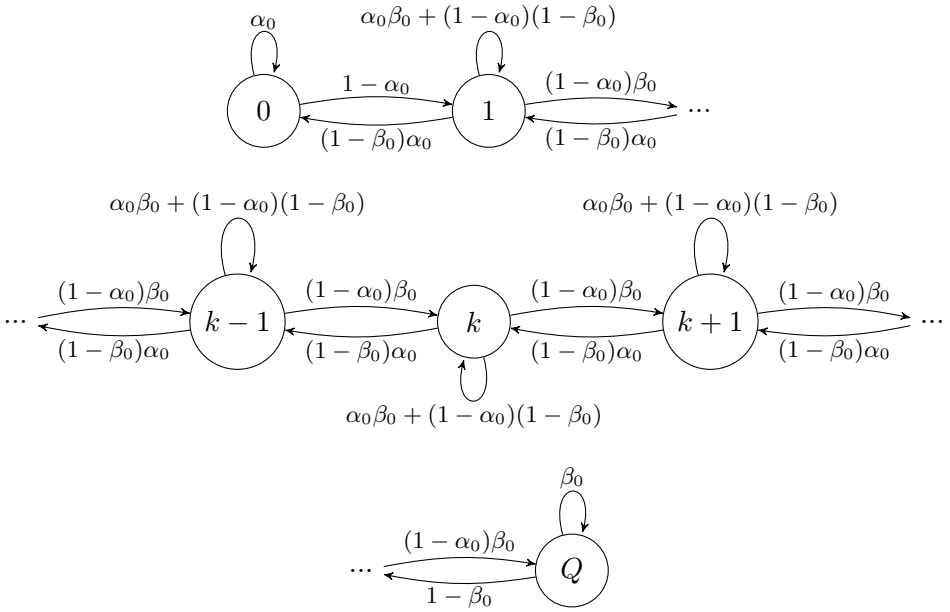
The reasoning is analogous when starting from $(x^1, 0)$. It is impossible to release patients of type 2, which means that $\beta_2 = 0$.



And here is the graph from a certain state (x^1, x^2) with $x^1 + x^2 \leq Q$.



We are now able to define the chain $(S_t)_t$ by $S_t = X_t^1 + X_t^2$, accounting for the total number of patients, regardless of their types. One can easily show that the transition graph and the transition matrix are exactly the same as with one profile, only replacing the coupled values (p, q) with $(1 - \alpha_0, 1 - \beta_0)$. Hence, we obtain the following transition diagram and transition matrix.



$$P(1-\alpha_0, 1-\beta_0) = \begin{pmatrix} \alpha_0 & 1-\alpha_0 & & & \\ (1-\beta_0)\alpha_0 & (1-\beta_0)(1-\alpha_0)+\beta_0\alpha_0 & \beta_0(1-\alpha_0) & & (0) \\ & \ddots & \ddots & \ddots & \\ & & (1-\beta_0)\alpha_0 & (1-\beta_0)(1-\alpha_0)+\beta_0\alpha_0 & \beta_0(1-\alpha_0) \\ (0) & & & 1-\beta_0 & \beta_0 \end{pmatrix}$$

The time-shortage t^* is obtained in exactly the same way as before, computing the first passage time from x_0 to Q on the previous matrix: $\mathbb{P}(t^* = t) = f_{x_0,Q}^{(t)}$.

3 A simple operational forecasting tool: a traffic signal

In this section, we want to propose an application of our model that at first sight would seem promising, that is a traffic signal.

3.1 Traffic signal

The first type of study that can be quickly and immediately carried out is a simple computation of the time of shortage for a given set of parameters for each simulation. health-care professionals can decide to run various studies of this kind to get a glimpse of how complicated the situation is or is about to become.

Let us consider, for instance, a situation with $Q = 6$ ventilators and a daily time step in order to make our interpretations. We can plot the density distribution function of

the random time-shortage t^* as shown in Figure 5. A continuous multi-color gradient, assessing the risk of shortage, helps us to illustrate, in a quite explicit way, the urgency of the situation. The color denotes the value of the cumulative distribution function (from 0 to 1) of t^* such that with the passing of time, the probability that a shortage event has happened by that time increases. For example, on the left-hand side graph of Figure 5, the chance of having a new patient every day is $p = 60\%$ and the chance of releasing one patient is $q = 40\%$. Until day 10, all seems to be under control with a low probability of shortage ($\leq 25\%$). But it rapidly gets worse up until day 20 where all signals start triggering before meeting complete disaster around day 30 when it is probably already too late to react. Once the overall distribution is computed, any value of interest related to it can be calculated such as the mean time-shortage, the median time, the mode, etc. For instance, we obtain $\mathbb{E}(t^*) \approx 23$ days which means that a shortage is to be expected on average on the twenty-third day. Actually, the ratio $\frac{p}{q}$ can be seen as the mean number of patients arriving for each patient that leaves, which means that if this value is too big the outcome of a shortage is likely to be swift and brutal. The density function is heavily concentrated on the left and does not spread uniformly along the whole graph. This can be seen as a graphical sign that the situation is critical. Conversely, the right-hand side graph of Figure 5 (with $q > p$) shows a situation with low tension turning orange at day 500 and red at day 1000 after the beginning of the study.

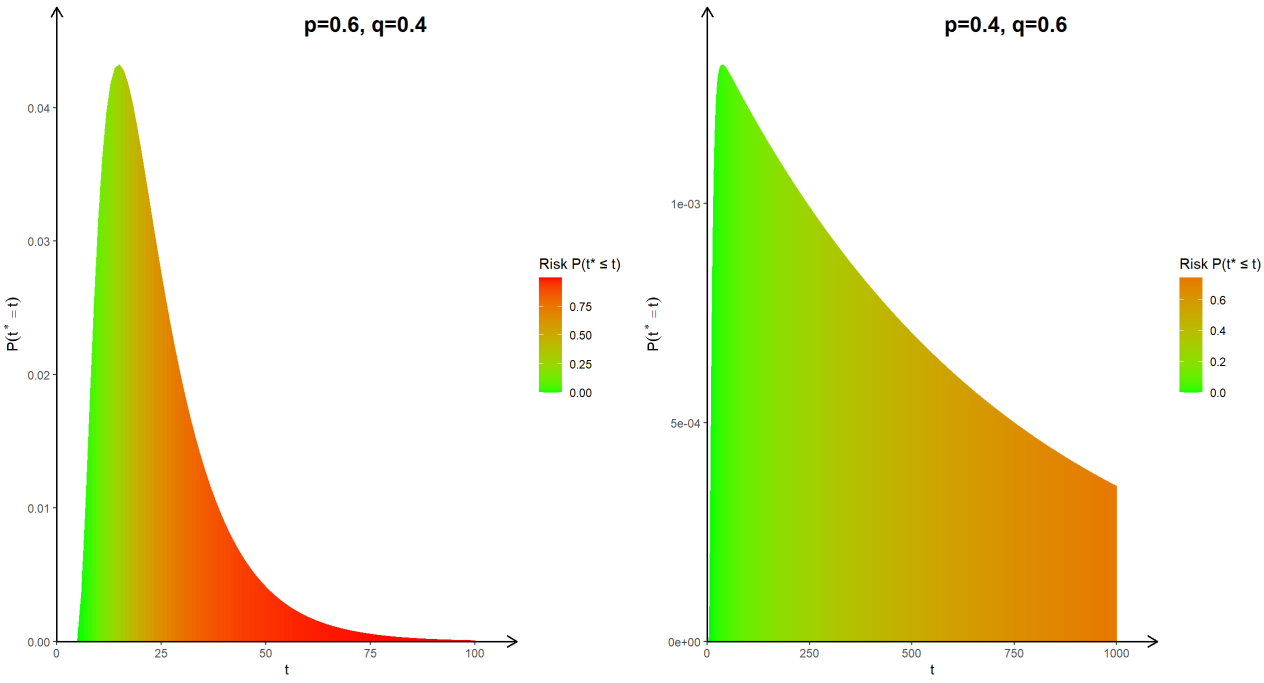


Figure 5: Traffic signal with $Q = 6$.

This kind of graphical traffic signal can be freely handled online on the web application developed by the first author through the link https://shortage.shinyapps.io/shortage/ with the possibility of choosing a single or a multiple profiles model, the values

of the three parameters p, q, and Q, and the duration of the study (on the x-axis).

3.2 Application on data for a French ICU

In this section, we aim to apply this model on real data. We have at our disposal data from a French ICU that can be described as follows. From 01/01/2010 to 31/12/2019 (before the Covid-19 pandemic started), we have all patients who went through this service during the 10 years under consideration, which represents not less than 8,600 patients. More particularly, for each patient who stayed in ICU, we have his/her date and time of admission, date and time of discharge, the number of hospital beds with ventilators in the service officially available at the moment of entry, and the Simplified Acute Physiology Score (SAPS). This last feature is a point score that is calculated from 12 physiological variables and 3 disease-related variables during the first 24 hours, along with information about previous health status and some information obtained at admission. In other words, the higher this score is, the lower the survival expectancy of the patient. This score can be useful to split our sample of patients into multiple groups that correspond to what we call "profiles" of patients. Let us now consider technically how to apply our model to these data. The number of beds in the service is in general constantly equal to 15, which will be our reference value for Q, considering that one bed occupied is equivalent to one ventilator occupied. The time step to choose would be a time duration small enough so that there is no more than one entry and one exit per time unit. A quick study shows that a 2 hour time step enables us to make this hypothesis valid on average. The hardest part consists in being able to find a period during these 10 years of data where the estimations of parameters p and q do not vary much over time. Indeed, the assumption according to which p and q must be constant with time is paramount. Let us choose, for example, the month of January in 2012. On this period, we assess p as

$$\hat{p} = \frac{\text{Number of patients admitted in January 2012}}{\text{Number of 2-hours time steps in January 2012}}.$$

Since 60 patients were admitted and since we have $12 \times 31 = 372$ time steps of 2 hours, we get the estimated admission probability $\hat{p} \approx 0.161$. Similarly, the estimated discharge probability is $\hat{q} \approx 0.147$. We can now plot the density distribution of the time-shortage for this month. The x-axis is labeled with calendar dates so that the reader does not have to convert time steps into dates.

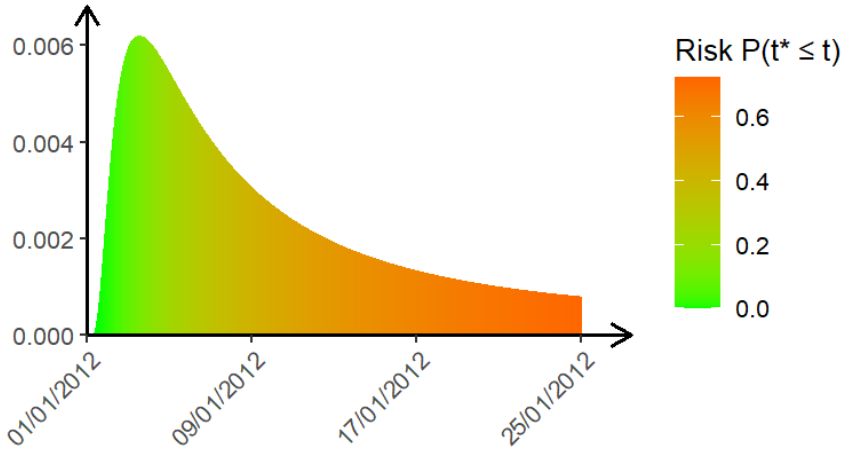


Figure 6: Risk of shortage for a French ICU in January 2012 with one single profile.

We see that the situation starts deteriorating around day 8. With a more than 60% chance a shortage will occur after 17 days. Since $\hat{p} > \hat{q}$, the shape of the distribution is very soon concentrated on the left. The expected time-shortage can also be computed, for which we get $\widehat{\mathbb{E}}(t^*) \approx 75$ steps. In other words it takes approximately 6 days on average to face a shortage. By extending this study beyond the date 25/01/2012, we could have a more complete curve and risks even higher than what we observe here.

If we wish to partition our sample into subgroups of patients with different characteristics, we can use the severity score given by the SAPS variable. A histogram regarding the distribution of this score over the 8,600 patients allows us to split them into 3 groups as shown in Table 3.

Profile	SAPS range	$\hat{p}$	$\hat{q}$
1	$[0, 25]$	0.018	0.017
2	$(25, 55)$	0.0847	0.0825
3	$[55, +\infty)$	0.0496	0.0455

Table 3: Calibration to 3 profiles for a French ICU according to SAPS values.

That way, we define a low/medium/high gravity type of patients, which allows us to estimate the admission rate and the discharging rate in January 2012 for each profile, as explained before for one profile. An analogous density plot can be drawn for these three profiles as shown in Figure 7.

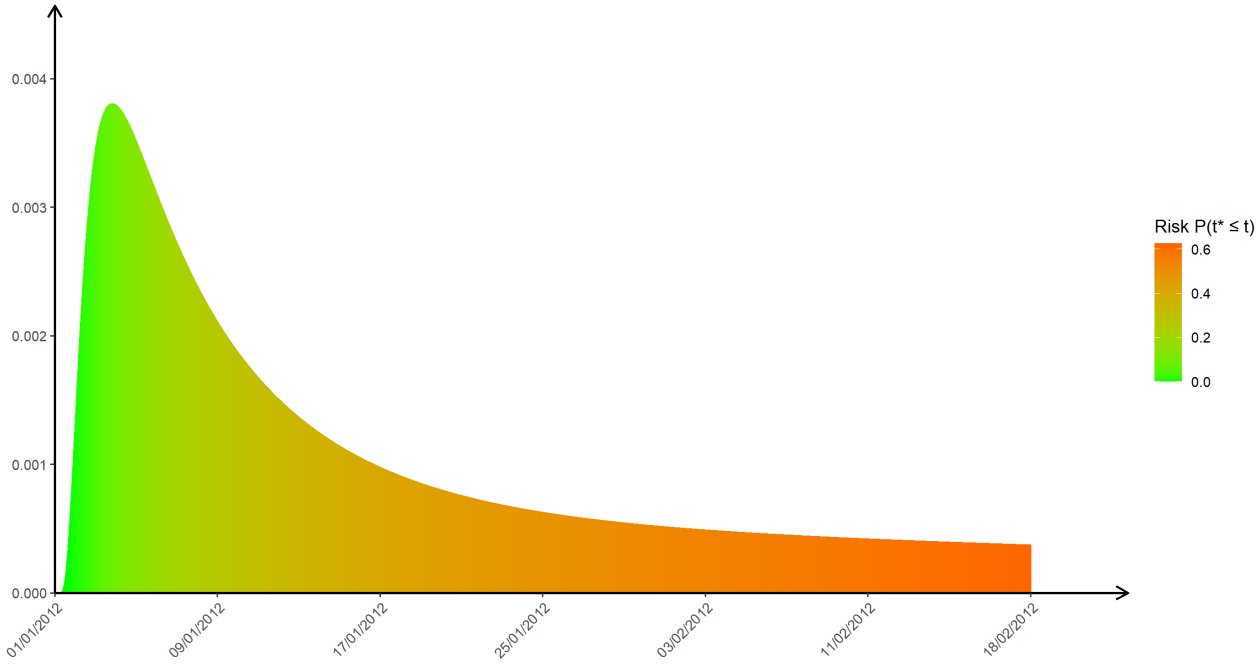


Figure 7: Risk of shortage for a French ICU in January 2012 with 3 profiles.

We can comment that we need to go further ahead in time to experience similar quantitative risks compared to the one-type case. On the date 18/02/2012, 50 days after the beginning of the study, we barely reach a 60% chance of shortage whereas after 25 days, with one single profile, we had a risk of more than 70% risk. Splitting patients into multiple profiles results, for this particular example, in a considerable gain of time and effort.

4 A management tool to evaluate purchasing decisions

4.1 Comparison of situations when adding one ventilator

Another application of our model that can be carried out is a comparison between two situations. We can imagine comparing two different intensive care units or a given unit with different parameters. The range of values taken by the four parameters (x_0, p, q, Q) can be freely chosen according to the situation of study, without any boundary conditions or restrictions. Figure 8 shows a type of situation that physicians could encounter. Imagine a service with $p = 0.6$ and $q = 0.5$, which means that shortage is likely to occur after some time. Physicians have 6 ventilators and and are hesitating about purchasing one more to cope with the situation and help them gain some time. Considering the actual cost of a ventilator,[10] this represents an investment, meaning that it is important to know if this decision is worth taking. We then introduce the two triplets of parameters (p_1, q_1, Q_1) and (p_2, q_2, Q_2) with $p_1 = p_2 = 0.6$, $q_1 = q_2 = 0.5$, $Q_1 = 6$ and $Q_2 = 7$ such that scenario 1 is

[10] Not to mention the commitment of the health care staff necessary to take charge of ventilated patients.

the current actual critical situation of the service and scenario 2 is the imaginary fictive situation, but more desirable one, with the additional ventilator. In other words, situation 2 is situation 1 with one more ventilator. We can now plot the cumulative distribution functions of respectively t_1^* and t_2^*, the times of shortage in situation 1 and 2. If we consider a daily time step, we can see, for instance, that the number of days saved is 8 (30 against 38) for a 50% risk of shortage.

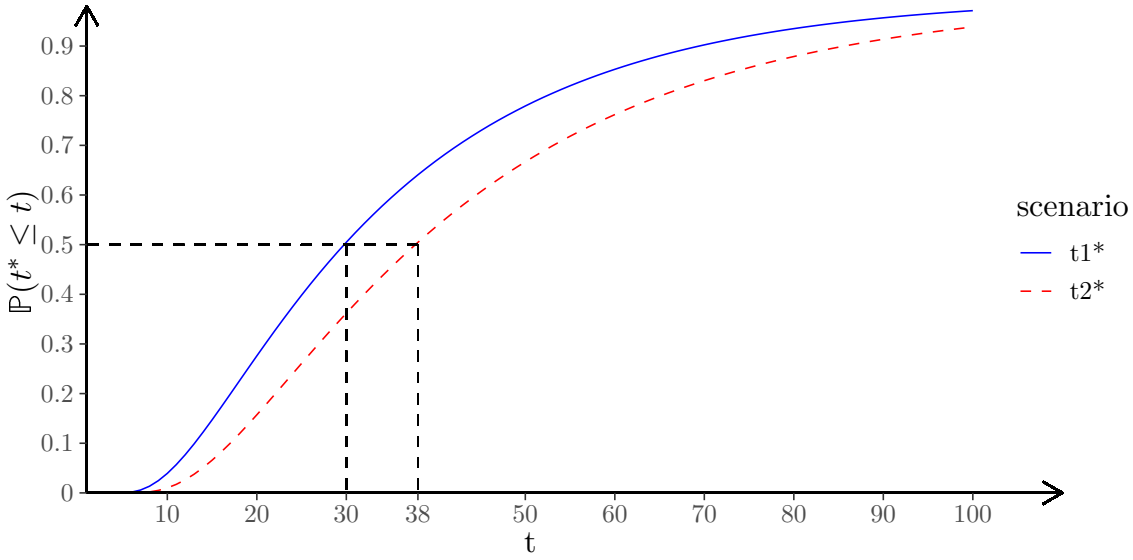


Figure 8: Risk of shortage as a function of time with $p_1 = p_2 = 0.6$, $q_1 = q_2 = 0.5$, $Q_1 = 6$, and $Q_2 = 7$.

This graphical comparison can also be carried out online with the freedom to choose any possible values for the parameters p, q, and Q, again through the same link as before: https://shortage.shinyapps.io/shortage/. One can compute the amount of time saved but also the value of risk avoided (in terms of probability). If we now plot the quantile difference $z_2(\alpha) - z_1(\alpha)$,[11] that is the number of time steps (days in our case) saved as a function of the risk taken, we see that this difference is positive (see Figure 9). This means that, no matter the level of risk that has to be assumed, scenario 2 is always more favorable since the shortage always occurs later with one more ventilator, but, in addition, the greater the level of risk that is assumed, the more we manage to save days since the represented function increases.

[11] For any risk $\alpha \in [0, 1]$, the quantile z_α is the unique value such that $\mathbb{P}(t^* \leq z_\alpha) = \alpha$. Hence, $\alpha \mapsto z_\alpha$ is the reciprocal function of the cumulative distribution function $t \mapsto \mathbb{P}(t^* \leq t)$.

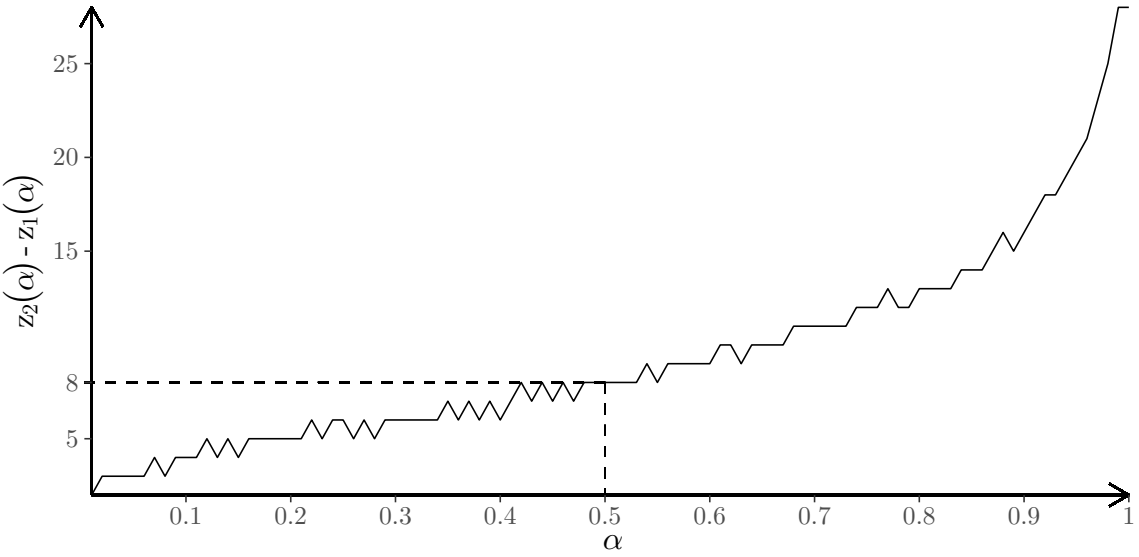


Figure 9: Quantile difference between scenarios 1 and 2 as a function of the risk α.

In order to propose relevant comparative situations, let us think about what parameters can be set up in reality. As we have just illustrated, the number of ventilators can change because the hospital can decide to borrow or buy ventilators, but some can also undergo maintenance or turn out to be broken. The parameter q, accounting for the probability of departure of the patients, is not really up to the therapists to choose. It is conditioned by the health state of patients and by the quality of work of the staff. The fact that a patient recovers or dies after a period of time in ICU does generally not depend on the good will of the staff. Hence, q is not really a parameter that can be modified according to what clinicians would want. As for parameter p, accounting for the probability of arrival of patients, it is mainly governed by epidemiological dynamics. However, in a situation of tension, doctors can decide to restrict the selection criteria for patient admissions, which was the easiest and most common practice adopted during the Covid-19 pandemic when ventilators were lacking. Hence, other fictive comparisons with the same number of ventilators but different values for p could also be relevant. To summarize, the only parameter we should not want to change too much relative to its actual value is parameter q.

4.2 Application on French ICU data

Applying the same methodology to our real data provides us with the following quantile curve:

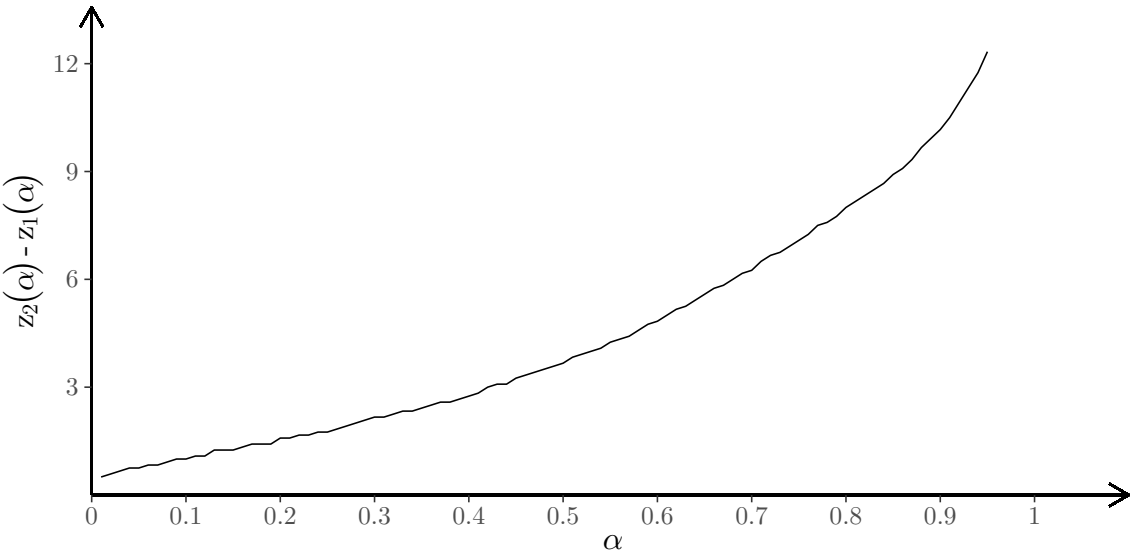


Figure 10: Amount of time saved as a function of the risk of adding one ventilator.

The delay is strictly increasing, meaning that the more risk practitioners are willing to take, the more time they will save. We see that for a 40% risk, 3 days can be gained with one more ventilator. t seems to us that it is not unreasonable to consider that this risk is already quite high, considering that (i) a ventilator is expensive and (ii) there is a time delay between the moment we purchase the additional ventilator and the moment it starts being operational on site, etc.

5 Conclusion

The ideal way to allocate scarce resources in times of crisis remains a delicate topic of research. We saw that it is possible to assess quickly, in terms of probabilities, how much time there is left before a shortage arises as long as we know the quantity of available resources and the rates of occupation and release of the scarce resource. We were able to obtain density, cumulative, and quantile plots for the time-shortage and deduce the risk of an impending shortage as a function of the time ahead. Comparison between situations of interest and an application on real ICU data have been provided to emphasize the pragmatic side of our work.

The calibration on real data also reveals some restrictive aspects of the model, especially the fact that parameters need to be constant over time during the whole study. These parameters are likely to be very volatile and manifest great variability, particularly in a period of tension. This could make the process difficult to apply in practice, and hence it should be exclusively dedicated to projections in the short term. The further ahead we look in time, the less the model is reliable since it is never updated with new data. A solution would be to search for periods when the moving side deviation of p and q are relatively low among all available data. In any case, for this main reason, applying

the model on real data and being able to formulate guidelines or medical protocols seems slightly ambitious. Being able to relax the hypotheses that p and q are constant would be a significant improvement, which can be done by carrying out simulations or even adapting the recursive formula 1 to a non-homogeneous case.

Another drawback which needs to be pointed out is the fact that, when making forecasts, we assume that the parameters calibrated on the previous period remain the same on the targeted future period in time. For instance, in Figure 6, the graph extends until the month of February but only the month of January was used to calibrate the parameters. Hence, what is represented from the date 01/02/2012 is pure prediction based on the model, and should be interpreted very carefully, or at least during only a very short horizon of time.

Note that the status dead/alive of a patient leaving the service is unknown in our model. Taking into account this information would be a possible improvement, introducing, for example, proportions of death and recovery after a period of time in ICU, the proportion of lethal cases growing with the stress in the service. The indicator that the service is under stress could be that $\frac{p}{q}$ is high, or that the total number of patients is high leading to greater loss of chance of survival. In that case, the natural variable to be minimized would be the proportion of deaths.

Other variations could be considered in order to improve our model. Recall that in practice one patient requires many nurse. So we could imagine a model where a patient uses more than one unit of the resource with a certain proportion. We could even think about a model with many types of resources being employed, dependent on various factors (1 hospital bed for 4 nurses for instance). We could also set, for each incoming patient, a random duration of stay in the service with a certain probability distribution, so that the probability of exiting the service depends only on the patient himself/herself.

Another side that could have been more deeply investigated from the economic point of view consists in introducing, for instance, the cost of a ventilator, the cost of a human life lost, and establishing some kind of "cost evaluation" for the decision maker. In this context, the economic problem amounts to an optimal arbitration between (i) a maximum number of ventilators that the hospital an purchase, thus saving maximum lives but at a greater cost to public health finances, and (ii) a minimum number of ventilators, allowing money to be saved but with the risk of high mortality rates.

References

Acemoglu, D., Chernozhukov, V., Werning, I., and Whinston M.D. (2020) A multi-risk SIR model with optimally targeted lockdown. Mimeo, National Bureau of Economic Research. Available at https://ifs.org.uk/publications/multi-risk-sir-model-optimally-targeted-lockdown.

Akbarpour, M., Budish, E., Dworczak, P., and Kominers, S.D (2021) An economic frame-

work for vaccine prioritization. Mimeo, *Available at SSRN 3846931.*

Alfa, A. (2016) *Applied discrete-time queues.* Springer.

Atkeson, A. (2020) What will be the economic impact of Covid-19 in the US? Rough estimates of disease scenarios. Mimeo, National Bureau of Economic Research. Available at http://www.nber.org/papers/w26867.

Bhaskar, S., Tan, J., Bogers, M., Minssen, T., Badaruddin, H., Israeli-Korn, S., and Chesbrough, H. (2020) At the epicenter of Covid-19, the tragic failure of the global supply chain for medical supplies. *Frontiers in Public Health*, 24(8):562882.

Bonneuil, N. (2021) Optimal age-and sex-based management of the queue to ventilators during the covid-19 crisis. *Journal of Mathematical Economics*, 93:102494.

Box, G., Jenkins, G., Reinsel, G., and Ljung, G. (2015) *Time series analysis: forecasting and control.* John Wiley & Sons.

Chen, W.C., Chen, L., and Kao, Y.C. (2021) Efficient mask allocation during a pandemic. Mimeo, *Available at SSRN 4006799.*

Dar, M., Swamy, L., Gavin, D., and Theodore, A. (2021) Mechanical-ventilation supply and options for the Covid-19 pandemic. Leveraging all available resources for a limited resource in a crisis. *Annals of the American Thoracic Society*, 18(3):408–416.

Eikenberry, S., Mancuso, M., Iboi, E., Phan, T. Eikenberry, K., Kuang, Y., Kostelich, E., and Gumel, A. (2020) To mask or not to mask: modeling the potential for face mask use by the general public to curtail the Covid-19 pandemic. *Infectious Disease Modelling*, 5:293–308.

Emanuel, E., Persad, G., Upshur, R., Thome, B. Parker, M., Glickman, A., Zhang, C., Boyle, C., Smith, M., and Phillips, J. (2020) Fair allocation of scarce medical resources in the time of Covid-19. *New England Journal of Medicine*, 382:2049-2055.

Gómez, S., Arenas, A., Borge-Holthoefer, J., Meloni, S., and Moreno, Y. (2010) Discrete-time Markov chain approach to contact-based disease spreading in complex networks. *Europhysics Letters*, 89(3):38009

Hassan, M.R. and Nath, B. (2005) Stock market forecasting using hidden Markov model: a new approach. *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*:192–196.

Hillier, G.J. and Lieberman, F.S. (2001) *Introduction to operations research.* McGraw-Hill, New York.

Lee, H.R. and Lee, T. (2018) Markov decision process model for patient admission decision at an emergency department under a surge demand. *Flexible Services and Manufacturing Journal*, 30:98-122.

Mayhew, L. and Smith, D. (2008) Using queuing theory to analyse the government's 4-h completion time target in accident and emergency departments. *Health Care Management Science*, 11(1):11-21.

Meisami, A., Deglise-Hawkinson, J., Cowen, M.E., and Van Oyen, M.P. (2019) Data-driven optimization methodology for admission control in critical care units. *Health Care Management Science*, 22:318–335.

Nganmeni, Z., Pongou, R., Tchantcho, B., and Tondji, J.B. (2022) Vaccine and inclusion. *Journal of Public Economic Theory*, 24(5):1101–1123.

Norris, J.R. (1998) *Markov chains.* Cambridge University Press.

Olmos, P.R. and Borzone, G.R. (2021) Stepwise Markov model: a good method for forecasting mechanical ventilator crisis in Covid-19 pandemic. *Epidemiologic Methods*, 10(1):2020-0021.

Pathak, P.A., Sönmez, T., Unver, M.U., and Yenmez, M.B. (2020) Leaving no ethical value behind: triage protocol design for pandemic rationing. National Bureau of Economic Research. Mimeo, Available at https://www.nber.org/papers/w26951.

Ranney, M.L., Griffeth, V., and Jha., A.K. (2020) Critical supply shortages—the need for ventilators and personal protective equipment during the Covid-19 pandemic. *New England Journal of Medicine*, 382(18):e41.

Rosenbaum., L. (2020) Facing Covid-19 in Italy—ethics, logistics, and therapeutics on the epidemic's front line. *New England Journal of Medicine*, 382(20):1873–1875.

Saha, E. and Ray., P.K. (2019) Patient condition-based medicine inventory management in health-care systems. *IISE Transactions on health-care Systems Engineering*, 9(3): 299–312.

Santini, A., Messina, A., Costantini, E., Protti, A., and Cecconi, M. (2022) Covid-19: dealing with ventilator shortage. *Current Opinion in Critical Care*, 28(6):652-659.

Truog, R.D., Mitchell, C., and Daley., G.Q. (2020) The toughest triage—allocating ventilators in a pandemic. *New England Journal of Medicine*, 382(21):1973–1975.

Verity, R., Okell, L.C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., Cuomo-Dannenburg, G., Thompson, H., Walker, P.G.T., and Fu, H. (2020) Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet Infectious Diseases*, 20(6):669–677.

Zukerman, M. (2013) Introduction to queueing theory and stochastic teletraffic models. Mimeo, arXiv:1307.2968.

Appendix

A

Here is a concise sketch of the algorithm. The function "diag" is the following:

Algorithm 1 "diag" function

Input: a transition matrix P

Output: diag(P), the same matrix as P on the diagonal with 0's everywhere else

$n \leftarrow \text{size}(P)$
for i, j *from* 1 *to* n **do**
  if $i \neq j$ **then**
    $P[i, j] \leftarrow 0$
  end
end
Return P

Algorithm 2 Computation of $\mathbb{P}(t^* = t)$ for $t = 1, \ldots, N$ and for parameters (p, q, Q, x_0)

Input: $p, q \in [0, 1]$, $Q \in \mathbb{N}^*$, N the duration of study, x_0 number of initial patients

Output: $\mathbb{P}(t^* = t)_{1 \leq t \leq N}$ as t^*

Initialize the transition matrix $P \leftarrow 0$ with size $Q + 1$
$P[0, 0] \leftarrow 1 - p$
$P[0, 1] \leftarrow p$
for k *from* 1 *to* Q **do**
  if $k < Q$ **then**
    $P[k, k + 1] \leftarrow p(1 - q)$
  end
  else
    if $k > 1$ **then**
      $P[k, k - 1] \leftarrow q(1 - p)$
    end
    else
      $P[k, k] \leftarrow pq + (1 - p)(1 - q)$
    end
  end
end
Initialize the matrix f of first passage times: $f \leftarrow P$
Initialize the vector of shortage probabilities: $t^* \leftarrow [0, ..., 0]$ of size N
Initialize time: $t \leftarrow 1$
while $t < N + 1$ **do**
  $t^*[t] \leftarrow f[x_0, Q]$ ▷ Adding up to vector t^* the newly computed shortage probability
  $f \leftarrow P \times (f - \text{diag}(f))$ ▷ Computing the first passage matrix for the following time
  $t \leftarrow t + 1$
end
Return t^*

B

In this section, we briefly sketch the values of probabilities in the case of n different profiles of patients, with the condition that maximum one type enters or leaves at the same time.

For $i = 1, ..., n$, let us consider the admissions $Y^i \sim \mathcal{B}(p_i)$ and the exits $Z^i \sim \mathcal{B}(q_i)$ knowing that $\sum_{i=1}^n Y^i \leq 1$ and $\sum_{i=1}^n Z^i \leq 1$, i.e., no more than one of the n variables can be equal to 1. The vector $X = (X^1, ..., X^n)$ gives the total number of patients of each type.

Denoting $Y = (Y^1, ..., Y^n)$ the total number of patients of each type, $r_i = \frac{p_i}{1-p_i}$ and $R = \sum_{i=1}^n r_i$, we have the probability distribution of Y given by

$$\mathbb{P}\left(Y = 0_n \middle| \sum_j Y^j \leq 1\right) = \frac{1}{1+R} := \alpha_0,$$

where $0_n := (0, ..., 0)$. If we introduce the vector $e_i = (0, ...0, 1, 0, ...0)$ where the 1 is in position i for $i \geq 1$, we can show that

$$\mathbb{P}\left(Y = e_i \middle| \sum_j Y^j \leq 1\right) = \frac{r_i}{1+R} := \alpha_i.$$

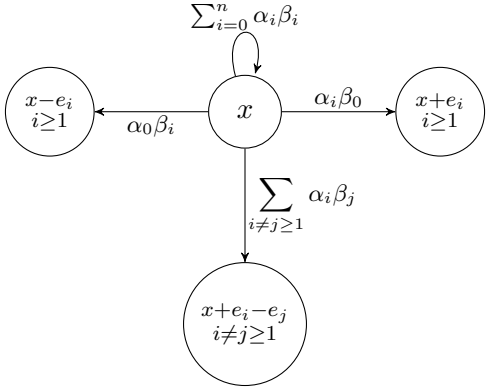
In a similar way, let us define $r'_i = \frac{q_i}{1-q_i}$, $R' = \sum_{i=1}^n r'_i$ so that

$$\mathbb{P}\left(Z = 0_n \middle| \sum_i Z^i \leq 1\right) = \frac{1}{1+R'} := \beta_0,$$

and

$$\mathbb{P}\left(Z = e_i \middle| \sum_j Z^j \leq 1\right) = \frac{r'_i}{1+R'} := \beta_i.$$

The complete graph of chain X is obviously not possible to draw. However, we can describe the situation starting from a certain state $x = (x^1, ..., x^n) \in \mathbb{N}^{*n}$ (the case with zeros would need another study). Then we have the following transitions and associated probabilities:



Nothing changes from the previous case for the rest. If we denote $S_t = \sum_{i=1}^{n} X_t^i$, then S has the same transition graph and matrix as before.

The end is exactly the same as before. The graph and the matrix of the chain $S = \sum_{i=1}^{n} X^i$ are unchanged and the first passage time is to be computed on the Markov chain S.

C

For instance, in the case of a single profile with entries Y and exits Z, we can check that transition probabilities are given by the formula

$$p_{i,j} = \sum_{k,k'=0}^{\infty} \mathbb{1}(s(i-k'+k)=j)\mathbb{P}(Y_t=k)\mathbb{P}(Z_t=k'),$$

where $s(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } 0 \leq x \leq Q \\ Q & \text{if } x \geq Q \end{cases}$ and

$\mathbb{1}(a=b) = \begin{cases} 1 & \text{if } a=b \\ 0 & \text{otherwise} \end{cases}$

D

For all $(i,j) \in E$, for all $t \in \mathbb{N}^*$,

$$\begin{aligned} p_{i,j} &= \mathbb{P}(X_{t+1}=j | X_t=i) \\ &= \mathbb{P}(X_t + Y_t - Z_t = j | X_t = i) \\ &= \mathbb{P}(i + Y_t - Z_t = j | X_t = i) \\ &= \mathbb{P}(Y_t - Z_t = j - i). \end{aligned}$$

Let $Y \sim \mathcal{B}(p)$, $Z \sim \mathcal{B}(q)$ be two Bernoulli independent variables. Since $Y - Z \in \{-1,0,1\}$, if $j - i \notin \{-1,0,1\}$, then $p_{i,j} = 0$ clearly. So, we have the 3 following remaining cases to treat:
(i) $\mathbb{P}(Y-Z=1) = \mathbb{P}(\{Y=1\} \cap \{Z=0\}) = \mathbb{P}(Y=1)\mathbb{P}(Z=0) = p(1-q)$
(ii) $\mathbb{P}(Y-Z=-1) = (1-p)q$ in a similar way
(iii)

$$\begin{aligned} \mathbb{P}(Y-Z=0) &= \mathbb{P}(\{Y=Z=1\} \cup \{Y=Z=0\}) \\ &= \mathbb{P}\left([\{Y=1\} \cap \{Z=1\}] \cup [\{Y=0\} \cap \{Z=0\}]\right) \\ &= \mathbb{P}(Y=1)\mathbb{P}(Z=1) + \mathbb{P}(Y=0)\mathbb{P}(Z=0) \\ &= pq + (1-p)(1-q), \end{aligned}$$

hence, the announced result holds. To finish, notice that $X_t = 0 \implies Z_t = 0$ such that $p_{0,0} = \mathbb{P}(X_{t+1} = 0 | X_t = 0) = \mathbb{P}(Y_t - Z_t = 0 | X_t = 0, Z_t = 0) = \mathbb{P}(Y_t = 0) = 1 - p$. Furthermore,
$p_{0,1} = \mathbb{P}(X_t + Y_t - Z_t = 1 | X_t = 0) = \mathbb{P}(Y_t = 1 + Z_t | X_t = 0, Z_t = 0) = \mathbb{P}(Y_t = 1) = p$. An analog reasoning can be made to establish $p_{Q,Q}$ and $p_{Q,Q-1}$.

E

Let P be the transition matrix and π be the unique stationary distribution. Then

$$\pi P = \pi \iff \begin{cases} (1-p)\pi_0 + p\pi_1 = \pi_0 \\ q(1-p)\pi_{k-1} + (pq + (1-p)(1-q))\pi_k + p(1-q)\pi_{k+1} = \pi_k, \ k = 1, \dots, Q-1 \\ q(1-p)\pi_{Q-1} + (1 - q(1-p))\pi_Q = \pi_Q \end{cases}$$

$$\iff \begin{cases} \pi_1 = \pi_0 \\ q(1-p)\pi_{k-1} + (2pq - p - q)\pi_k + p(1-q)\pi_{k+1} = 0, \ k = 1, \dots, Q-1 \\ \pi_{Q-1} = \pi_Q \end{cases}$$

We can then easily show, by inductive reasoning, that $(\pi_k)_{0 \leq k \leq Q}$ is constant, and since $\sum_{k=0}^{Q} \pi_k = 1$, we have $\pi_k = \frac{1}{Q+1}$ for all $k = 0, \dots, Q$.

Statements and Declarations

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.