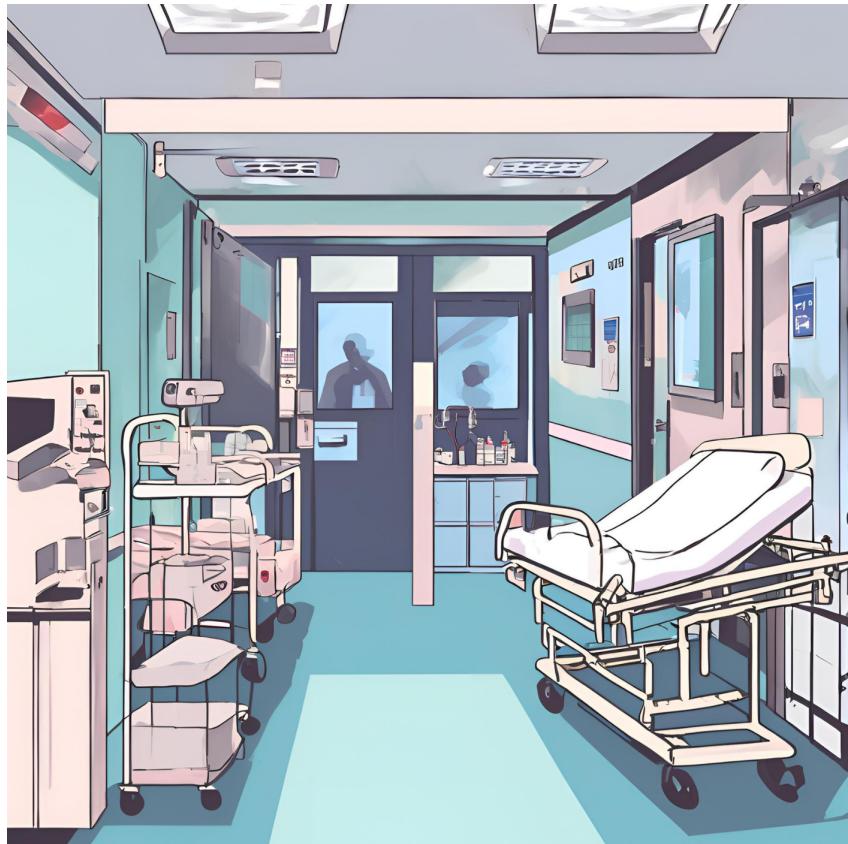


---

# DIMENSIONNEMENT ET STAFFING D'UNE UNITÉ D'URGENCE HOSPITALIÈRE

---

EA initiation à la Recherche Opérationnelle



Stéphane EILLES-CHAN WAY et Hugo PERCOT

Encadrés par :

Xavier ALLAMIGEON

Stéphane GAUBERT

December 15, 2024

# Introduction

Ce travail s'inscrit dans le cadre du projet de recherche URGE, collaboration entre Inria et l'AP-HP (Assistance Publique-Hôpitaux de Paris). Co-piloté par Xavier ALLAMIGEON, ce programme vise à analyser et optimiser les parcours patients aux urgences. Il s'agit d'un enjeu crucial, d'autant plus que "*la fréquentation des services d'accueil des urgences (SAU) a presque doublé au cours des vingt dernières années*"<sup>1</sup>.

Notre sujet s'articule plus particulièrement autour du dimensionnement et de la confection des emplois du temps du personnel (infirmiers, médecins). Nous nous sommes en partie appuyés sur les travaux réalisés pour les centres d'appels d'urgences (PFAU). Cependant, l'approche pour les services des urgences implique une plus grande complexité et une diversité de parcours qu'il convient de maîtriser. Pour cette raison, ce rapport présente différentes tentatives de modélisation et d'analyse de la dynamique de tels services. Nous reconnaissions ne pas avoir pu obtenir de résultats concrets, mais nous estimons avoir pu mettre en lumière diverses méthodes pour aborder ce sujet complexe.

Ainsi, une première partie fera un état des lieux des différents parcours des patients dans les services d'urgence. Puis, nous présenterons les modélisations utilisées, les résultats établis, les limites rencontrées et les démarches que nous avons mises en place pour les contourner. Enfin, nous discuterons de perspectives de poursuite et des opportunités qu'offre ce travail.

---

<sup>1</sup><https://www.inria.fr/fr/ap-hp-inria-projet-urge-optimisation-parcours-soins>

# Contents

<b>1 Description des pratiques et de l'organisation générale d'un service d'urgence</b>	<b>4</b>
1.1 Acteurs . . . . .	4
1.2 Lieux et fonctionnement . . . . .	4
<b>2 Première approche - Régime stationnaire</b>	<b>5</b>
2.1 Réseau de Petri temporel . . . . .	6
2.2 Mise en équation . . . . .	7
2.3 Résolution du régime stationnaire . . . . .	9
2.4 Diagramme de phase . . . . .	12
2.5 Problème d'optimisation . . . . .	12
2.6 Planning admissible . . . . .	14
<b>3 Optimisation du stock</b>	<b>15</b>
3.1 Réseau de Petri . . . . .	15
3.2 Mise en équation . . . . .	15
<b>4 Autre approche de la modélisation du stock</b>	<b>19</b>
4.1 Notations . . . . .	19
4.2 Modélisation d'une crise coup de bélier . . . . .	20
4.2.1 Sans stratégie d'allocation . . . . .	20
4.2.2 Avec une stratégie d'allocation qui favorise le traitement des patients en 1 . . . . .	21
4.3 Modélisation d'une crise longue . . . . .	23
4.3.1 Modélisation . . . . .	23
4.3.2 Remarques préliminaires sur la gestion d'une crise longue . . . . .	23
4.3.3 Gestion de la crise moyenne . . . . .	23
<b>5 Retour au problème de staffing</b>	<b>26</b>

# 1 Description des pratiques et de l'organisation générale d'un service d'urgence

Dans cette partie, nous décrivons la variété de parcours de soin au sein des services d'urgences hospitalières. On s'attachera à rester succinct tout en montrant la grande complexité qui caractérise ce type de système. Pour cela, nous nous appuyons dans une large mesure sur le rapport d'observation de Pascal BENCHIMOL réalisé à l'Hôtel-Dieu [4].

## 1.1 Acteurs

Tout d'abord, un service d'urgence hospitalière fait intervenir une multitude d'acteurs.

Les médecins seniors ont le plus haut niveau de responsabilité. Ils sont habilités à effectuer des consultations seuls et décident du traitement à appliquer au patient (soin, examen complémentaire, transfert, etc). Dans le cadre de l'Hôtel-Dieu, qui est aussi un hôpital universitaire, ils ont une mission éducative auprès des internes et des externes.

Les internes sont des étudiants en médecine. Ils ont passé le concours de l'internat et donc sont en pratique souvent compétents pour effectuer des consultations en autonomie. Dans le cadre légal, ils doivent faire valider leur diagnostic par un médecin senior, ce qui n'est pas toujours réalisé en pratique.

Les externes sont également des étudiants en médecine, mais moins formés. C'est souvent à eux que les médecins accordent le plus de temps de formation. Il peut effectuer des consultations, mais elles seront systématiquement revues par un médecin senior. En particulier, ce dernier prend aussi le temps de l'initier à la saisie des dossiers et des comptes-rendus.

Les infirmiers diplômés d'État sont en charge des actes médicaux simples, mais aussi du triage des patients et de leur répartition dans les boxes.

## 1.2 Lieux et fonctionnement

Le parcours des patients est caractérisé par un ensemble d'étapes successives et de lieux associés. Ce parcours est représenté dans la figure 1.

A son arrivée, le patient est inscrit dans la base administrative puis est reçu par un infirmier dans une salle de tri IAO (Infirmier d'Accueil et d'Orientation). L'objectif à ce stade est de déterminer le degré d'urgence du patient, noté de 1 (cas les plus graves) à 4 (cas les moins urgents). Le patient est ensuite placé en salle d'attente.

Ils sont ensuite pris en charge par un infirmier qui les installera dans un box. Avant d'être pris en charge, l'attente peut être longue, notamment parce que l'ordre de prise en charge est déterminé par le degré d'urgence établi au préalable. Un box est une petite salle équipée de matériel médical basique, isolée du reste du service, utilisée pour réaliser les consultations et les soins.

Lorsqu'un médecin (ou interne ou externe) est disponible, il vient dans un box pour effectuer une première consultation (encore une fois selon l'ordre des degrés d'urgences). Trois solutions s'offrent alors à lui : sortie du service ; soins par un infirmier (puis sortie ou analyse complémentaire) ; analyse complémentaire. Dans le dernier cas, le patient est revu lors d'une deuxième consultation, qui suit globalement les mêmes règles que la première.

Enfin, un patient peut être envoyé dans un service hospitalier classique si besoin, auquel cas il est placé aux portes, en attendant de lui trouver une place.

En parallèle de ce parcours, les médecins, internes et externes doivent remplir des compte-rendus à destination du dossier du patient.

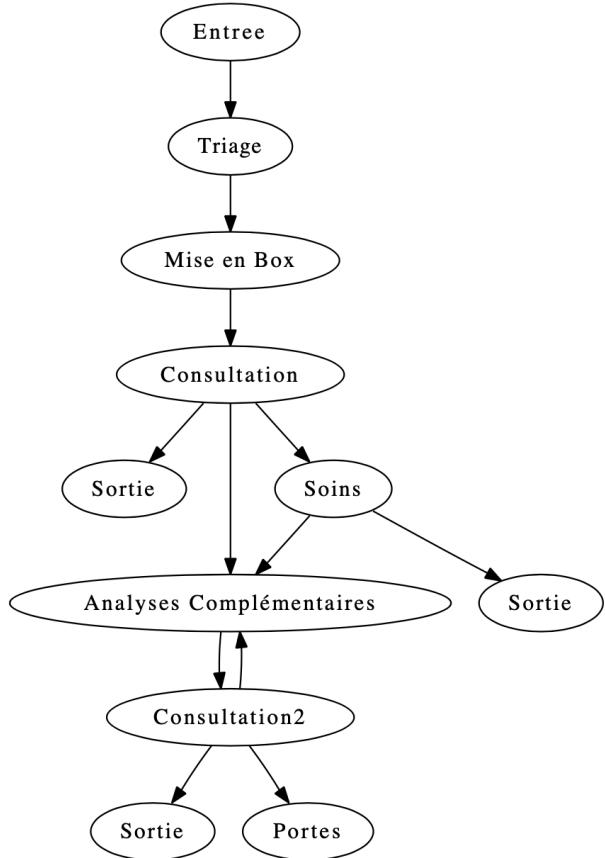


Figure 1: Graphe de parcours des patients dans le service d'urgence.  
Figure tirée du rapport de Pascal BENCHIMOL

## 2 Première approche - Régime stationnaire

Dans cette section, nous tentons de construire une première modélisation de l'unité d'urgence hospitalière. L'objectif est à la fois de pouvoir décrire sa dynamique, mais aussi de réfléchir à un critère et des contraintes pour l'optimisation qui suivra.

Nous commençons par faire volontairement des hypothèses simplificatrices pour travailler sur un modèle abordable. Tout d'abord, concernant les acteurs :

**Hypothèse.** *On suppose que les infirmiers sont parfaitement fluides et qu'il n'y a pas d'externe, de sorte que l'on ne s'intéressera qu'aux internes et médecins seniors.*

Puis, pour ce qui est du parcours des patients :

**Hypothèse.** *On suppose que les consultations mènent toutes à une sortie du système, de sorte que l'on ne modélise pas les soins et examens complémentaires.*

Enfin, on fait une dernière hypothèse sur le triage des patients :

**Hypothèse.** *On suppose qu'il existe seulement un degré d'urgence, de sorte que les patients sont traités dans leur ordre d'arrivée (First In First Out)*

À partir de ces hypothèses, on se retrouve à devoir modéliser uniquement la primo-consultation. On prend alors le fonctionnement suivant :

- Un patient qui arrive peut être pris en charge par un interne ou un (médecin) senior.

- Si la consultation est réalisée par un interne, alors le diagnostic doit être validé par un senior.
- Étant donné que les seniors ont deux tâches (consultation et validation), on décide de considérer les priorités suivantes :
  - Un patient sera prioritaire pris en charge par un interne s'il y a au moins un interne disponible, pour libérer du temps au senior pour la validation
  - Un senior ira prioritairement faire des validations s'il y a au moins un diagnostic d'interne en attente. Ainsi, il cherchera d'abord à faire sortir un patient du système avant d'en faire rentrer un nouveau.

Dans la suite de cette section, on présente l'outil utilisé pour représenter cette modélisation, et on résout les équations obtenues.

## 2.1 Réseau de Petri temporel

En s'inspirant des travaux de Martin Boyet [2] sur les centres d'appels d'urgences, nous avons d'abord choisi de modéliser le service par un réseau de Petri temporel.

**Définition 2.1** (Réseau de Petri). Un **réseau de Petri** est un triplet  $(\mathcal{P}, \mathcal{Q}, \mathcal{E})$ , où  $\mathcal{P}$  est un ensemble fini d'états,  $\mathcal{Q}$  un ensemble fini de transitions, et  $\mathcal{E} \subset (\mathcal{P} \times \mathcal{Q}) \cup (\mathcal{Q} \times \mathcal{P})$  représentant les relations possibles entre des états et des transitions.

Le réseau que nous avons alors construit est représenté à la figure 2. Les états sont symbolisés par des cercles et les transitions par des rectangles.

**Principe.** *Les agents interagissant avec le système représenté par le réseau de Petri seront représentés par des jetons (Tokens) pouvant passer d'un état à un autre par le biais des transitions. Plus d'informations sur le comportement d'un réseau de Petri peuvent être retrouvées dans la thèse de BOYER.*

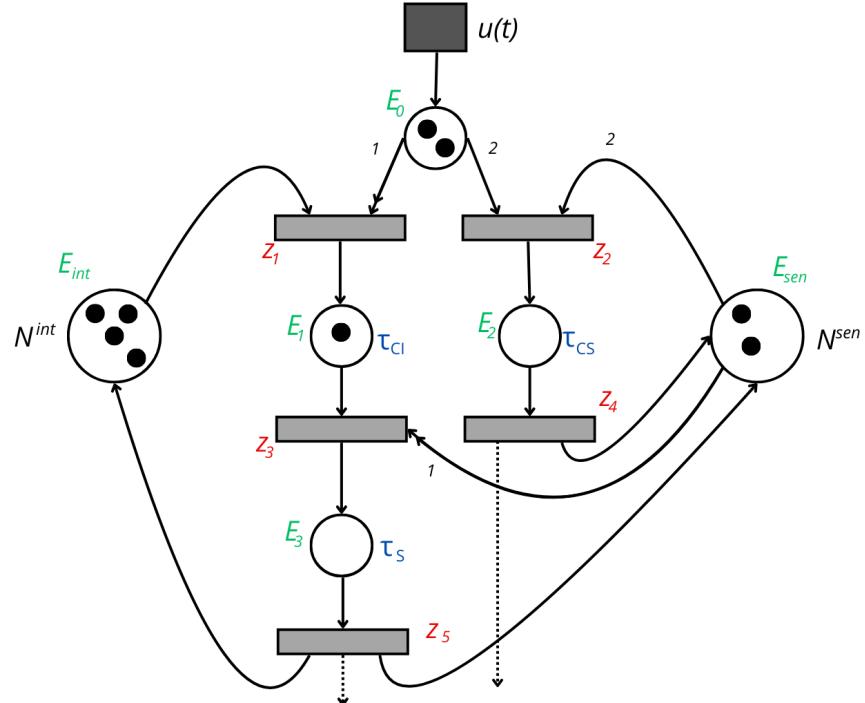


Figure 2: Réseau de Pétri du problème

Les patients qui arrivent dans le service par la transition  $u(t)$  qui génère des tokens et les place dans l'état  $E_0$ . La fonction  $u(t)$  renseigne du nombre de tokens produits sur la période  $[0, t]$ . Les tokens patients peuvent alors activer la transition  $Z_1$  (prise en charge par un interne) ou  $Z_2$  (prise en charge par un senior), selon si un token est disponible dans l'état  $E_{int}$  ou  $E_{sen}$ . Comme expliqué plus haut, et symbolisé par une double flèche sur le schéma, s'il y a un token disponible en  $E_{int}$  ET en  $E_{sen}$ , alors la transition  $Z_1$  est activée en priorité. Le système comprend  $N^{int}$  internes et  $N^{sen}$  seniors, qui sont au départ sous la forme de token en  $E_{int}$  et  $E_{sen}$ , puis peuvent circuler dans le système. Ces deux variables sont exactement l'effectif que l'on fait travailler à ce moment. Pour simplifier, on peut considérer pour l'instant que ces quantités sont constantes, i.e. qu'il n'y a pas de rotation d'équipe sur la période étudiée.

La temporalité du réseau de Petri est donnée par des temps caractéristiques associés à chaque état. Ils sont représentés en bleu sur la figure. Ces temps représentent un temps de maturité à partir duquel le token est disponible pour activer une transition en aval. Par exemple, un token issu de la transition  $Z_1$  ne pourra pas activer la transition  $Z_3$  avant un temps  $\tau_{CI}$  (qui correspond à la durée d'une consultation par un interne). Après  $\tau_{CI}$ , la transition  $Z_3$  reste inactivée tant qu'un token n'est pas disponible dans l'état  $E_{sen}$ .

Enfin, on supposera que toute transition est activée dès qu'au moins un token est disponible dans tous les états amonts.

Il n'est pas raisonnable de suivre chaque token individuellement, notamment parce qu'on peut imaginer qu'un grand nombre de patients va passer dans le système. On va plutôt se placer au niveau des transitions et compter le nombre d'activations au cours du temps. Pour cela, on introduit les fonctions suivantes :

**Définition 2.2** (Fonction compteur). Pour  $i \in \{1, \dots, 5\}$ , on note  $z_i$  la **fonction compteur de transition** qui, pour  $t > 0$ , donne le nombre d'activations de la transition  $Z_i$  sur la période  $[0, t]$ . Ces fonctions sont caractérisées par le fait que  $z_i(0) = 0$  et que  $z_i$  est croissante, constante par morceaux.

Soit  $t > 0$ , on notera  $z_i(t^-)$  le nombre d'activations de la transition  $Z_i$  sur la période  $[0, t]$ . En particulier, si  $Z_i$  est activé à  $t$ , alors  $z_i(t) = z_i(t^-) + 1$ .

Pour les calculs, on a aussi besoin d'introduire des fonctions similaires pour les états :

**Définition 2.3** (Fonction compteur). Pour  $i \in \{0, 1, 2, 3, int, sen\}$ , on note  $e_i$  la **fonction compteur d'état** qui, pour  $t > 0$ , donne le nombre de tokens ayant la maturité dans l'état  $E_i$  sur la période  $[0, t]$ . On remarque également que  $e_i(0) = 0$  et que  $e_i$  est croissante, constante par morceaux.

**Remarque.** Pour les états qui ne sont pas sujets à des temps de maturité, alors cette fonction compteur d'état correspond aussi au nombre de tokens arrivés en  $E_i$  sur la période  $[0, t]$ , puisque la maturité est immédiate ( $\tau = 0$ ).

**Proposition 2.1.** L'espace des fonctions compteurs est un cône convexe de l'espace vectoriel des fonctions de  $\mathbf{R}_+$  dans  $\mathbf{R}$

*Proof.* On voit immédiatement que la caractérisation des fonctions compteurs est stable par combinaison linéaire positive.  $\square$

## 2.2 Mise en équation

À partir du réseau de Petri que l'on a construit, on veut pouvoir étudier la dynamique du système. Comme expliqué plus haut, on cherche alors à avoir des expressions de  $z_i(t)$  en fonction des paramètres du système ( $N^{int}$  et  $N^{sen}$ ) et des données externes ( $u(t)$ ). Pour cela, on va déterminer le système d'équations que vérifie ce réseau.

On utilise le fait que les transitions sont activées dès que les conditions en amont sont remplies (i.e. présence de tokens). On distingue alors deux types de transitions : les transitions simples ( $Z_4$  et  $Z_5$ ) et les transitions avec priorité ( $Z_1$ ,  $Z_2$  et  $Z_3$ ). Les premières n'ont qu'un seul état amont, ce qui ne demande pas d'attention particulière : un simple retard permet d'obtenir l'équation de ces transitions :

$$\begin{aligned} z_4(t) &= z_2(t - \tau_{CS}) \\ z_5(t) &= z_3(t - \tau_S) \end{aligned}$$

Les secondes dépendent de deux états en amont et surtout font l'objet de règles de priorité sur le choix des agents. Leur gestion est plus complexe, nous détaillons donc l'obtention de l'équation de  $Z_1$ .

Cette transition est en aval de deux états :  $E_{int}$  et  $E_0$ . Si elle était l'unique aval de ces états, alors on aurait :

$$\forall t > 0, \quad z_1(t) = \min(e_{int}(t), e_0(t)) \quad (1)$$

En effet, le nombre d'activations de  $Z_1$  est limité par la pénurie de l'un de ses deux états amonts. Ensuite, on a :

$$\forall t > 0, \quad e_0(t) = u(t)$$

par définition de  $u$  et

$$\forall t > 0, \quad e_{int}(t) = N^{int} + z_5(t)$$

puisque  $N^{int}$  est le nombre de tokens présents initialement en  $E_{int}$  et  $z_5(t)$  est le nombre de tokens internes qui sont revenus en  $E_{int}$  sur  $[0, t]$ .

Cependant, l'état  $E_0$  est aussi en amont de la transition  $Z_2$ , donc il faut corriger l'équation 1 :

$$\forall t > 0, \quad z_1(t) = \min(e_{int}(t), e_0(t) - z_2(t^-))$$

car pour obtenir le nombre de tokens en  $E_0$  utilisés pour activer  $Z_1$ , on doit retrancher ceux utilisés pour activer  $Z_2$ . Étant donné la priorité mise sur  $Z_1$  par rapport à  $Z_2$ , on utilise  $z_2(t^-)$  au lieu de  $z_2(t)$  car si en  $t$  les deux transitions peuvent être activées, c'est bien  $Z_1$  que l'on choisira.

De même, par priorité de  $Z_3$  sur  $Z_2$  pour l'état  $E_{sen}$  on a :

$$\forall t > 0, \quad z_3(t) = \min(e_1(t), e_{sen}(t) - z_2(t^-))$$

A l'inverse, pour  $Z_2$ , les deux règles de priorité étant en défaveur de la transition, on a simplement :

$$\forall t > 0, \quad z_2(t) = \min(e_0(t) - z_1(t), e_{sen}(t) - z_3(t))$$

avec

$$\forall t > 0, \quad e_{sen}(t) = N^{sen} + z_4(t) + z_5(t)$$

et

$$\forall t > 0, \quad e_1(t) = z_1(t - \tau_{CI})$$

Finalement, on obtient le système d'équations régissant le réseau de Petri, ne dépendant que des fonctions compteurs de transition :

$$\forall t > 0, \quad \begin{cases} z_1(t) = \min(N^{int} + z_5(t), u(t) - z_2(t^-)) \\ z_2(t) = \min(N^{sen} + z_4(t) + z_5(t) - z_3(t), u(t) - z_1(t)) \\ z_3(t) = \min(z_1(t - \tau_{CI}), N^{sen} + z_4(t) + z_5(t) - z_2(t^-)) \\ z_4(t) = z_2(t - \tau_{CS}) \\ z_5(t) = z_3(t - \tau_S) \end{cases}$$

On peut réduire le nombre de variables en observant que  $z_4$  et  $z_5$  sont simplement un retard de  $z_2$  et  $z_3$  :

$$\forall t > 0, \quad \begin{cases} z_1(t) = \min(N^{int} + z_3(t - \tau_S), u(t) - z_2(t^-)) \\ z_2(t) = \min(N^{sen} + z_2(t - \tau_{CS}) + z_3(t - \tau_S) - z_3(t), u(t) - z_1(t)) \\ z_3(t) = \min(z_1(t - \tau_{CI}), N^{sen} + z_2(t - \tau_{CS}) + z_3(t - \tau_S) - z_2(t^-)) \end{cases} \quad (2)$$

### 2.3 Résolution du régime stationnaire

En l'état, il est difficile d'obtenir des formules analytiques de ces fonctions compteurs, et il faut gérer les termes en  $t^-$ . Pour cela, on va d'abord regarder le comportement du système en régime stationnaire. Ce dernier revient entre autres à considérer les fonctions compteurs non plus comme constantes par morceaux, mais continues. Cela se justifie dans le cadre d'un service à grand effectif et sur des périodes de temps macroscopiques.

**Définition 2.4** (Régime stationnaire). Le **régime stationnaire** du réseau de Petri est un régime dans lequel toutes les fonctions compteurs sont affines du temps. Autrement dit, le réseau fonctionne à débit constant. Ainsi le cône des fonctions compteurs devient alors isométrique au cône convexe  $(\mathbf{R}_+, \mathbf{R})$ , par l'application :

$$\begin{aligned} Z &\rightarrow (\mathbf{R}_+, \mathbf{R}) \\ z_i &\mapsto (\rho_i, u_i) \end{aligned}$$

Enfin, on posera pour  $t > 0$ ,  $u(t) = \lambda t$ , ce qui revient à associer  $u$  à  $(\lambda, 0)$  dans l'isométrie précédente.

On veut alors travailler dans ce nouvel espace, mais on a besoin d'une relation d'ordre pour prendre le  $\min$  :

**Proposition 2.2.** *L'ordre lexicographique sur  $(\mathbf{R}_+, \mathbf{R})$  est une relation d'ordre :*

$$(\rho_i, u_i) > (\rho_j, u_j) \iff \begin{cases} \rho_i > \rho_j \\ u_i > u_j \text{ si } \rho_i = \rho_j \end{cases}$$

On s'appuie alors sur [1] pour gérer les termes en  $t^-$  :

$$\begin{aligned} (\rho_1, u_1) &= \begin{cases} (\lambda, -u_2) \wedge (\rho_3, N^{int} + u_3 - \rho_3 \tau_S) \text{ si } \rho_2 = 0 \\ (\rho_3, N^{int} + u_3 - \rho_3 \tau_S) \text{ si } \rho_2 > 0 \end{cases} \\ (\rho_2, u_2) &= (\lambda - \rho_1, -u_1) \wedge (\rho_2, N^{sen} + u_2 - \rho_2 \tau_{CS} - \rho_3 \tau_S) \\ (\rho_3, u_3) &= \begin{cases} (\rho_3, N^{sen} + u_3 - \rho_3 \tau_S) \wedge (\rho_1, u_1 - \rho_1 \tau_{CI}) \text{ si } \rho_2 = 0 \\ (\rho_1, u_1 - \rho_1 \tau_{CI}) \text{ si } \rho_2 > 0 \end{cases} \end{aligned} \quad (3)$$

La technique pour les équations de  $Z_1$  et  $Z_3$  est de distinguer les cas selon la valeur de  $\rho_2$ . Si cette dernière est non nulle, du fait de la règle de priorité, c'est que l'autre état amont limite ( $E_{int}$  pour  $Z_1$  et  $E_1$  pour  $Z_3$ ).

On pose  $\rho = \rho_1 + \rho_2$  le débit de patient qui est pris en charge. On note que  $\rho \leq \lambda$ . En effet, le débit de prise en charge est nécessairement limité par le débit d'arrivée des patients.

**1er cas :  $\rho_2 = 0$**  Dans ce cas, l'équation sur  $\rho_1$  donne  $\rho_1 = \lambda \wedge \rho_3$ , d'où  $\rho_1 \leq \rho_3$  et l'équation sur  $\rho_3$  donne  $\rho_3 = \lambda \wedge \rho_1$ , d'où  $\rho_3 \leq \rho_1$ . Ainsi  $\rho_1 = \rho_3 = \rho$  car  $\rho_2 = 0$ . Autrement dit, logiquement, tout le flux

de patients passe par la consultation d'interne en régime stationnaire. On cherche maintenant à calculer le débit de ce flux.

Supposons que  $\rho < \lambda$ . Alors d'une part, l'équation de  $u_1$ , c'est le terme de droite qui réalise le minimum, donc  $u_1 = N^{int} + u_3 - \rho\tau_S$ . D'autre part, l'équation sur  $u_3$  donne :

$$u_3 = (N^{sen} + u_3 - \rho\tau_S) \wedge (u_1 - \rho\tau_{CI})$$

d'où d'une part :

$$N^{sen} \geq \rho\tau_S$$

et d'autre part :

$$N^{int} = u_1 - u_3 + \rho\tau_S \geq \rho(\tau_S + \tau_{CI})$$

On en conclut, en prenant en compte le cas  $\rho = \lambda$ , que  $\rho = \min\left(\lambda, \frac{N^{sen}}{\tau_S}, \frac{N^{int}}{\tau_S + \tau_{CI}}\right)$ .

On remarque que selon le terme qui réalise ce minimum, la nature du facteur limitant est différente :

- Si  $\rho = \lambda$ , c'est le nombre de patients qui arrivent aux urgences qui limite le débit de patient traité.
- Si  $\rho = \frac{N^{sen}}{\tau_S}$ , c'est le nombre de médecins seniors qui limite le débit de patient traité.
- Si  $\rho = \frac{N^{int}}{\tau_S + \tau_{CI}}$ , c'est le nombre d'internes qui limite le débit de patient traité. En fait, ce cas n'apparaît que si on a aussi  $\rho = \frac{N^{sen}}{\tau_S}$ . Car si au contraire  $\rho < \frac{N^{sen}}{\tau_S}$  les médecins seniors ne sont pas tous occupés et ils prendront en charge le surplus de patients, on aurait alors  $\rho_2 > 0$ .

On en déduit alors que le cas  $\rho_2 = 0$  est réalisé si et seulement si  $\frac{N^{int}}{\tau_S + \tau_{CI}} \geq \min(\lambda, \frac{N^{sen}}{\tau_S})$ , donc si et seulement si  $\frac{N^{int}}{\tau_S + \tau_{CI}} > \lambda$  (les internes sont assez nombreux pour faire toutes les primo-consultations eux-mêmes) ou si  $\frac{N^{int}}{\tau_S + \tau_{CI}} > \frac{N^{sen}}{\tau_S}$  (les internes ne sont peut-être pas assez nombreux, mais de toute façon les seniors ne sont pas assez nombreux pour valider leurs diagnostics).

**2eme cas :  $\rho_2 > 0$**  D'après la conclusion du cas précédent, ce cas est équivalent à

$$\min\left(\lambda, \frac{N^{sen}}{\tau_S}\right) > \frac{N^{int}}{\tau_S + \tau_{CI}}. \quad (4)$$

On a, par l'équation sur  $\rho_1$ , que  $\rho_1 = \rho_3$ . Puis par l'équation sur  $u_1$  :

$$u_1 = N^{int} + u_3 - \rho_3\tau_S$$

et par l'équation sur  $u_3$  :

$$u_3 = u_1 - \rho_1\tau_{CI}$$

En sommant les deux équations, on obtient :

$$\rho_3 = \rho_1 = \frac{N^{int}}{\tau_S + \tau_{CI}}$$

Naturellement, on remarque que cette valeur de  $\rho_1$  correspond à la situation des internes limitants dans le cas  $\rho_2 = 0$ . Plus précisément, le débit de patient par la voie de primo-consultation par un interne est limité par le nombre d'internes. Cela est par ailleurs cohérent avec l'équivalence annoncée au début du cas, par  $\rho_2 > 0$  ainsi que le fait que  $\rho = \rho_1 + \rho_2 \leq \lambda$ , on a bien  $\frac{N^{int}}{\tau_S + \tau_{CI}} < \lambda$ .

Déterminons  $\rho_2$  désormais. Par l'équation de  $u_2$ , on a :

$$u_2 \leq N^{sen} + u_2 - \rho_2\tau_{CS} - \rho_3\tau_S$$

avec égalité si  $\rho_1 + \rho_2 < \lambda$ .

On obtient alors :

$$\rho_2 \leq \frac{N^{sen} - \rho_3 \tau_S}{\tau_{CS}} = \frac{N^{sen} - N^{int} \frac{\tau_S}{\tau_S + \tau_{CI}}}{\tau_{CS}}$$

avec égalité si  $\rho_1 + \rho_2 < \lambda$ .

Retrouve cette fois que la condition  $\rho_2 > 0$  implique que  $\frac{N^{sen}}{\tau_S} > \frac{N^{int}}{\tau_S + \tau_{CI}}$ , c'est-à-dire que la ligne de traitement  $\rho_1$  n'est pas limitée par le nombre de médecins seniors, mais donc par le nombre d'internes, d'où l'équivalence 4.

La quantité  $N^{int} \frac{\tau_S}{\tau_S + \tau_{CI}}$  correspond à la quantité de médecins pris par la validation du diagnostic des internes. Elle augmente si le nombre d'internes  $N^{int}$  augmente, si le temps de diagnostic des internes  $\tau_{CI}$  diminue, ou si le temps de validation du diagnostic  $\tau_S$  augmente. Ainsi, le terme au numérateur dans l'expression de  $\rho_2$  correspond au nombre de médecins seniors restant pour traiter directement les nouveaux patients.

Étant données les deux bornes supérieures de  $\rho_2$ , on a :

$$\rho_2 = \min \left( \lambda - \rho_1, \frac{N^{sen} - N^{int} \frac{\tau_S}{\tau_S + \tau_{CI}}}{\tau_{CS}} \right)$$

Ainsi dans le cas  $\rho_2 > 0$ ,

$$\rho = \min \left( \lambda, \frac{N^{sen} - N^{int} \frac{\tau_S}{\tau_S + \tau_{CI}}}{\tau_{CS}} + \rho_1 \right) = \min \left( \lambda, \frac{N^{sen}}{\tau_{CS}} + \frac{N^{int}}{\tau_S + \tau_{CI}} \left( 1 - \frac{\tau_S}{\tau_{CS}} \right) \right)$$

En liant les deux cas, notamment par leur équivalence sur l'ordre des termes impliqués, on en conclut que de manière générale.

$$\rho = \begin{cases} \min \left( \lambda, \frac{N^{sen}}{\tau_S} \right) & \text{si } \frac{N^{sen}}{\tau_S} \leq \frac{N^{int}}{\tau_S + \tau_{CI}} \\ \min \left( \lambda, \frac{N^{sen}}{\tau_{CS}} + \frac{N^{int}}{\tau_S + \tau_{CI}} \left( 1 - \frac{\tau_S}{\tau_{CS}} \right) \right) & \text{sinon} \end{cases} \quad (5)$$

Dans le cas où  $\tau_S \leq \tau_{CS}$ , on a que  $(1 - \frac{\tau_S}{\tau_{CS}}) \geq 0$ . Ainsi, si  $\frac{N^{sen}}{\tau_S} \leq \frac{N^{int}}{\tau_S + \tau_{CI}}$ , on a :

$$\frac{N^{sen}}{\tau_{CS}} + \frac{N^{int}}{\tau_S + \tau_{CI}} \left( 1 - \frac{\tau_S}{\tau_{CS}} \right) \geq \frac{N^{sen}}{\tau_S}$$

et si  $\frac{N^{sen}}{\tau_S} > \frac{N^{int}}{\tau_S + \tau_{CI}}$ , on a :

$$\frac{N^{sen}}{\tau_{CS}} + \frac{N^{int}}{\tau_S + \tau_{CI}} \left( 1 - \frac{\tau_S}{\tau_{CS}} \right) \leq \frac{N^{sen}}{\tau_S}$$

d'où, toujours dans le cas :

$$\rho = \min \left( \lambda, \frac{N^{sen}}{\tau_S}, \frac{N^{sen}}{\tau_{CS}} + \frac{N^{int}}{\tau_S + \tau_{CI}} \left( 1 - \frac{\tau_S}{\tau_{CS}} \right) \right) \quad (6)$$

auquel cas  $\rho$  est une fonction concave de  $(N^{sen}, N^{int})$ . Intuitivement, ce cas est acceptable dans le sens où l'on s'attend à ce qu'un senior soit plus rapide pour valider le diagnostic d'un interne que pour mener la consultation de lui-même. Dans la pratique, les seniors ne font presque pas de validation sur les internes expérimentés, en revanche, ils passent un temps long avec les externes (non modélisés ici), à cause de leur mission de formation.

## 2.4 Diagramme de phase

Dans le cas  $\tau_S \leq \tau_{CS}$ , on obtient une surface tropicale concave, majorée par  $\lambda$ . Si on affiche  $\rho$  selon  $N^{sen}$  et  $N^{int}$ , on obtient la figure 3.

$$\rho = \min\left(\lambda, \frac{N^{sen}}{\tau_S}, \frac{N^{sen}}{\tau_{CS}} + \frac{N^{int}}{\tau_S + \tau_{CI}}(1 - \frac{\tau_S}{\tau_{CS}})\right)$$

$\lambda = 2; \tau_S = 2; \tau_{CS} = 3; \tau_{CI} = 4$

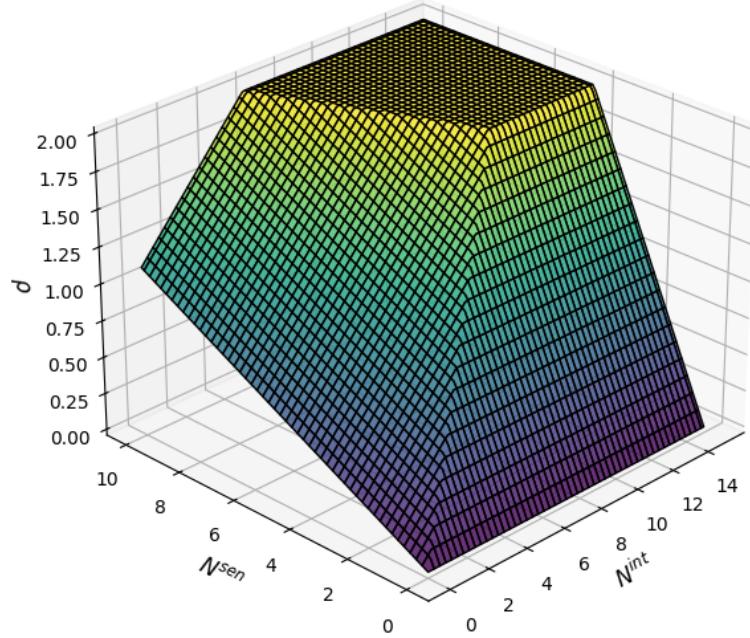


Figure 3: Représentation graphique de  $\rho$  avec des paramètres d'exemple

## 2.5 Problème d'optimisation

Maintenant que l'on a obtenu une expression du débit de traitement des patients en fonction du nombre d'internes et de seniors, on veut aboutir à un problème d'optimisation sur le staffing. On se place dans le cas  $\tau_S \leq \tau_{CS}$ , et on suppose que le régime quasi-stationnaire s'applique, c'est-à-dire que le temps caractéristique de variation du débit d'arrivée des patients est grand devant les temps de maturité  $\tau_{CI}$ ,  $\tau_{CS}$  et  $\tau_S$ .

Pour traiter le problème d'optimisation, il faut introduire un horizon temporel  $T$ . On peut par exemple prendre une semaine. L'intérêt est de ne pas prendre un horizon trop court (un jour) car les règles de planning s'étendent généralement sur plusieurs jours consécutifs, ni trop long (un mois) car le problème devient alors trop difficile à tracter.

**Définition 2.5** (Planning admissible). Un **planning admissible** est un planning de travail respectant un certain nombre de règles (Cf. section suivante au sujet des plannings admissibles).

On note  $\mathcal{P}^{sen} = \{p_1^{sen}, \dots, p_{n_P^{sen}}^{sen}\}$  l'ensemble des plannings admissibles pour les seniors, et  $\mathcal{P}^{int} = \{p_1^{int}, \dots, p_{n_P^{int}}^{int}\}$  l'ensemble des plannings admissibles pour les internes.

Formellement, les éléments de  $\mathcal{P}^{int/sen}$  sont des fonctions de  $[0, T]$  dans  $\{0, 1\}$ , qui valent 1 sur les intervalles de travail, et 0 pendant le repos.

Pour pouvoir traiter le problème, on pourra vectoriser les plannings par une discréétisation du temps  $0 = t_0 < t_1 < \dots < t_{n_T} = T$ . Alors les éléments  $\mathcal{P}^{int/sen}$  sont des vecteurs des  $\{0, 1\}^{n_T}$  dont la  $k$ -eme coordonnée vaut 1 si le planning prévoit du travail sur  $[t_{k-1}, t_k]$ , et 0 sinon. On peut alors voir  $\mathcal{P}^{int/sen}$  comme la matrice :

$$\begin{pmatrix} (p_1^{int/sen})^T \\ (p_2^{int/sen})^T \\ \vdots \\ (p_{n_P^{int/sen}}^{int/sen})^T \end{pmatrix}$$

**Définition 2.6** (Staffing). Un **staffing** est le fait d'attribuer à chaque séniors et interne un planning admissible. On note  $N_{tot}^{sen}$  et  $N_{tot}^{int}$  respectivement le nombre total de seniors et d'internes disponibles dans le service (en poste ou au repos).

Alors un staffing est composé de deux matrices  $X^\bullet$  de taille  $N_{tot}^\bullet$  par  $n_P^\bullet$  pour  $\bullet \in \{sen, int\}$ , qui traduisent l'attribution des plannings. Par exemple, pour  $i \in \{1, \dots, N_{tot}^{int}\}$ ,  $j \in \{1, \dots, n_P^{int}\}$ ,  $X_{ij}^{int}$  vaut 1 si le  $i$ -eme interne du service suit le  $j$ -eme planning, 0 sinon.

**Proposition 2.3.** *Étant donné sur le personnel ne peut se voir attribué un seul planing, Les matrices de staffing vérifient*

$$\forall i \in \{1, \dots, N_{tot}^{int}\}, \sum_j X_{ij}^{int} = 1 \quad (7)$$

$$\forall i \in \{1, \dots, N_{tot}^{sen}\}, \sum_j X_{ij}^{sen} = 1. \quad (8)$$

**Proposition 2.4.** *Étant donné sur un staffing  $(X^{sen}, X^{int})$ , on pose  $N^{sen}(t)$  et  $N^{int}(t)$  sont les nombre de seniors et d'internes en poste à la date  $t \in [0, T]$ . Alors, pour tout  $t \in [0, T]$ , si  $t \in [t_{k-1}, t_k]$ , on a les formules suivantes:*

$$N^{sen}(t) = \sum_i^{N_{tot}^{sen}} \sum_j^{n_P^{sen}} \mathcal{P}_{j,k}^{sen} X_{ij}^{sen} \quad (9)$$

$$N^{int}(t) = \sum_i^{N_{tot}^{int}} \sum_j^{n_P^{int}} \mathcal{P}_{j,k}^{int} X_{ij}^{int} \quad (10)$$

(11)

**Définition 2.7** ( $\alpha$ -couverture). Un staffing réalise une  **$\alpha$ -couverture** de  $t \mapsto \lambda(t)$  si pour tout  $t \in [0, T]$ ,  $\rho(t) \geq \alpha \lambda(t)$ , où  $\rho(t) = \min \left( \frac{N^{sen}(t)}{\tau_S}, \frac{N^{sen}(t)}{\tau_{CS}} + \frac{N^{int}(t)}{\tau_S + \tau_{CI}} \left(1 - \frac{\tau_S}{\tau_{CS}}\right) \right)$ .

**Remarque.** *Dans la définition précédente, le débit de traitement des patients est légèrement différent de celui établi précédemment. Cela se justifie par le fait que le concept de  $\alpha$ -couverture nous permet de nous assurer un marge de sécurité par rapport au débit nominal d'arrivé de patients (avec  $\alpha > 1$ ).*

En somme, il nous reste à déterminer un critère à optimiser. Si on considère la  $\alpha$ -couverture comme une contrainte (on veut garantir une certaine qualité de service), alors on peut vouloir minimiser le coût

du staffing. Le problème est alors :

$$\begin{aligned}
\text{Minimiser : } & \sum_i^{N_{tot}^{sen}} \sum_j^{n_P^{sen}} X_{ij}^{sen} \Pi_{ij}^{sen} + \sum_i^{N_{tot}^{int}} \sum_j^{n_P^{int}} X_{ij}^{int} \Pi_{ij}^{int} \\
\text{Sous les contraintes : } & \forall t \in [0, T], \quad \rho(t) \geq \alpha \lambda(t) \\
& \forall t \in [0, T], \quad \rho(t) \leq \frac{N^{sen}(t)}{\tau_S} \\
& \forall t \in [0, T], \quad \rho(t) \leq \frac{N^{sen}(t)}{\tau_{CS}} + \frac{N^{int}(t)}{\tau_S + \tau_{CI}} \left(1 - \frac{\tau_S}{\tau_{CS}}\right) \\
& \forall i \in \{1, \dots, N_{tot}^{int}\}, \quad \sum_j^{n_P^{int}} X_{ij}^{int} = 1 \\
& \forall i \in \{1, \dots, N_{tot}^{sen}\}, \quad \sum_j^{n_P^{sen}} X_{ij}^{sen} = 1 \\
& X^{sen} \in \{0, 1\}^{N_{tot}^{sen} \times n_P^{sen}} \\
& X^{int} \in \{0, 1\}^{N_{tot}^{int} \times n_P^{int}}
\end{aligned} \tag{12}$$

où  $\Pi_{ij}^{sen/int}$  est la matrice qui donne le coût d'attribuer le planning  $j$  au senior/interne  $i$ . On peut raisonnablement supposer que les seniors/internes sont payés au même prix, ce qui amène à voir les prix comme un vecteur  $\Pi_j^{sen/int}$ .

Le problème 12 est un programme linéaire en nombre entier. Il possède un nombre très important de contraintes. Nous n'avons pas cherché à résoudre un tel problème, et nous avons décidé d'essayer d'autres approches pour aboutir à des problèmes différents.

## 2.6 Planning admissible

Dans cette sous-section, nous évoquons rapidement la question des plannings de travail pour les internes et les seniors.

Ces derniers travaillent sur des intervalles de temps (des shifts). Plusieurs règles régissent les horaires et les combinaisons possibles de ces shifts, mais nous n'en ferons qu'une présentation simplifiée et évoquant les règles de planning des internes <sup>2</sup>.

Pour faire simple, les médecins peuvent travailler sur deux intervalles : garde de jour (8h - 18h) et garde de nuit (18h - 8h). La règle principale est qu'une garde de jour est interdite juste après une garde de nuit (repos obligatoire). L'enchaînement garde de jour → garde de nuit est quant à lui autorisé (garde de 24h).

Les gardes de nuit sont mieux rémunérés et donc implique un coût supplémentaire. De plus, les gardes le week-end sont rémunérées davantage, et le fait de faire plusieurs gardes de nuit dans la même semaine déclenche l'obtention d'un bonus.

Enfin, pour visualiser concrètement cette organisation du staffing en shift, on peut se référer à la figure 4

---

<sup>2</sup>Les planning des seniors sont plus flexibles, mais donc plus difficiles à définir en pratique

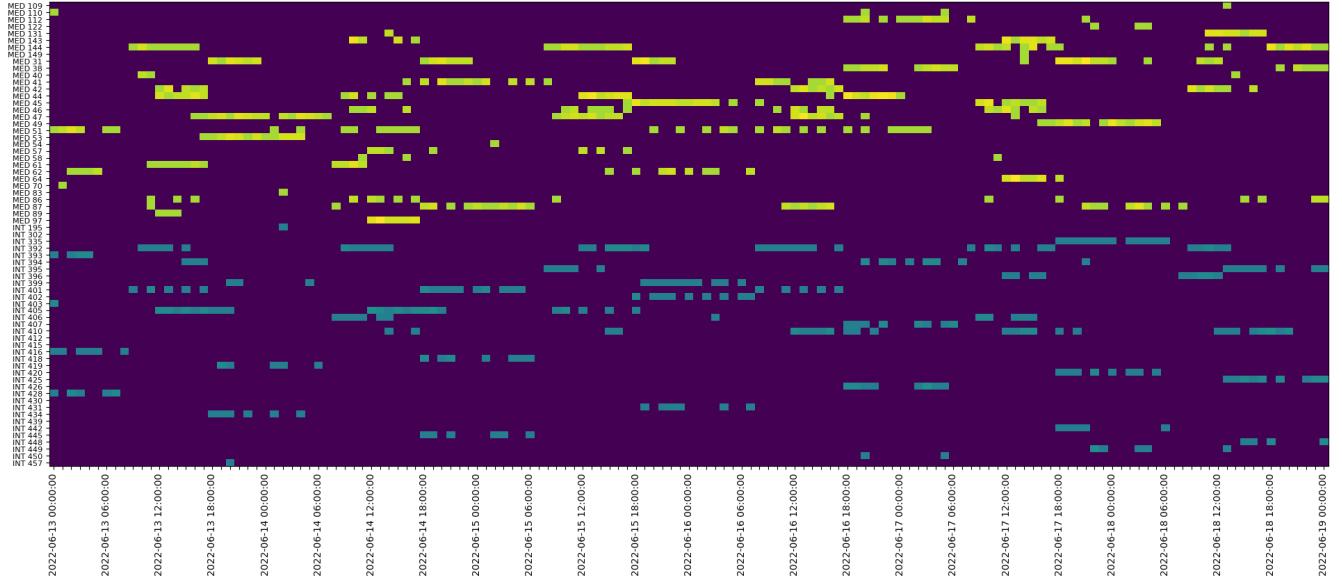


Figure 4: Activités enregistrées pour tout le personnel. On observe bien les gardes de nuits et de jour pour les internes (en bleu). L’organisation du travail pour les seniors est plus difficile à caractériser, mais on observe tout de même des gardes.

### 3 Optimisation du stock

Les résultats obtenus dans la partie précédente sont intéressants mais ne s’appliquent que dans le cadre limité du régime stationnaire. Ils ne permettent pas de rendre compte de phénomènes plus microscopiques, comme l’accumulation de ”stock de patient” ou encore le temps d’attente de chaque patient.

Pour étudier ces questions, nous choisissons de repartir d’un réseau de Petri, mais de le simplifier pour ne pas avoir à faire l’hypothèse du régime stationnaire.

#### 3.1 Réseau de Petri

Le nouveau réseau de Petri présenté sur la figure 5 ne présente plus qu’un seul type de médecin. En revanche, on est cette fois-ci plus exhaustif sur le parcours de soin en intégrant un état  $E_2$  de soin ou d’examen complémentaire (cf Partie 1 - Description des pratiques). Pendant que le patient est dans cet état, le médecin est libéré et peut s’occuper de nouveaux patients.

#### 3.2 Mise en équation

Le réseau de Pétri mène aux équations suivantes :

$$\begin{aligned}
 z_1(t) &= \min(z_{M_1}(t), u(t)) \\
 z_2(t) &= z_1(t - \tau_1) \\
 z_3(t) &= \min(z_1(t - \tau_1 - \tau_3), z_{M_3}(t)) \\
 z_4(t) &= z_3(t - \tau_3) \\
 z_{M_1}(t) + z_{M_3}(t) &= N^{med} + z_1(t - \tau_1) + z_3(t - \tau_3)
 \end{aligned}$$

On aimeraient minimiser (selon un critère à définir) le temps d’attente des patients. Dans notre modèle, deux points de congestion peuvent se créer :

- Point 1 : Avant la transition  $z_1$ , les patients s’accumulent s’il n’y pas assez de médecin pour prendre en charge le flux entrant.

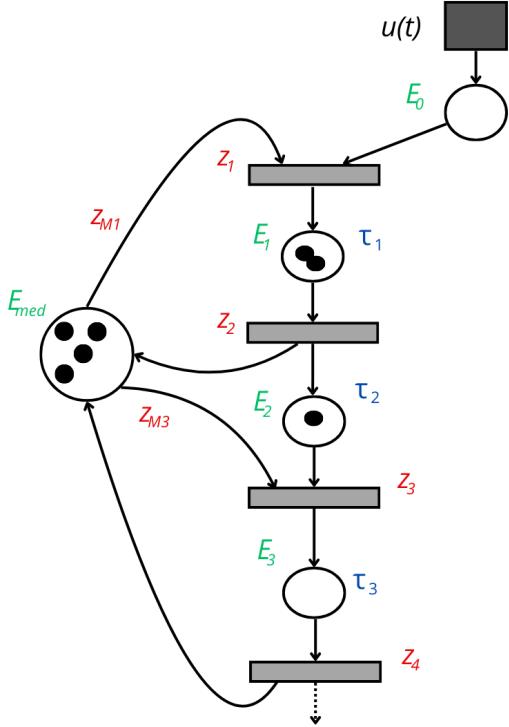


Figure 5: Réseau de Pétri avec examens complémentaires

- Point 3 : Avant la transition  $z_3$ , les patients s'accumulent après avoir fait leur examen complémentaire s'il n'y pas assez de médecin pour prendre effectuer la contre-visite.

Ces deux points de congestions n'ont pas la même gravité. On aimerait limiter le temps d'attente dans le premier point pour être sûr de prendre en charge rapidement les éventuels cas urgents (on suppose que l'on ne trie pas les types d'urgence en amont). L'attente après l'examen complémentaire (point 3) est moins critique, normalement les cas les plus urgents ont déjà été décelés à ce stade. Toutefois, afin de ne pas permettre une accumulation infinie de patients à ce stade du parcours, on veut tout de même contraindre le temps d'attente en ce point.

Le problème est un problème de contrôle : on veut trouver la meilleure fonction  $z_{M1}(\cdot)$  qui minimise l'attente maximale des patients, afin de garantir une certaine qualité de soin.

Ainsi, en indiquant par  $i \in \{1, \dots, N_{pat}\}$  les  $N_{pat}$  patients qui vont se présenter aux urgences entre  $t = 0$  et  $t = T$ , et en notant  $T_a^j(i)$  le temps d'attente du patient  $i$  au point  $j \in \{1, 3\}$ , on cherche à résoudre le problème suivant :

$$\begin{aligned} & \min_{z_{M1}(\cdot)} \left\{ \max_{i \in \{1, \dots, N_{pat}\}} T_a^1(i) \right\} \\ & \text{s.c. } \max_{i \in \{1, \dots, N_{pat}\}} T_a^2(i) \leq \bar{T} \end{aligned} \quad (13)$$

Il est difficile de travailler sur les variables dateurs. On va utiliser la dualité dateur-compteur. On introduit alors les variables de stocks au point  $j \in \{1, 3\}$  :  $S^j(\cdot)$ . Le temps d'attente et le stock au point 1 dépendent des compteurs  $u(t)$  et  $z_1(t)$ . Alors qu'il est difficile de trouver explicitement la relation entre ces derniers et  $T_a^1(i)$ , la variable compteur  $S^1$  s'exprime par :

$$\forall t \in [0, T], S^1(t) = u(t) - z_1(t)$$

. De même, pour  $S^2$ , on a :

$$\forall t \in [0, T], S^2(t) = z_1(t - \tau_1 - \tau_2) - z_3(t)$$

Cela se comprend bien à l'aide de la figure 6.

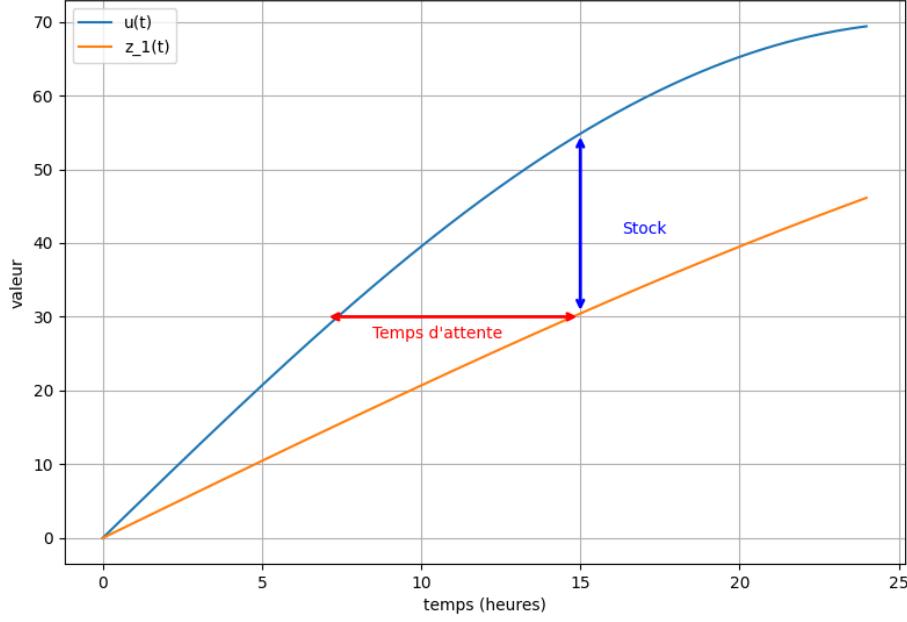


Figure 6: Représentation du temps d'attente et du stock à partir des courbes d'exemple du flux d'arrivée et du nombre de prises en charge

Alors, pour traiter le problème 13, on supposera que l'on peut plutôt regarder ce nouveau problèmes :

$$\begin{aligned} & \min_{z_{M_1}(\cdot)} \left\{ \max_{t \in [0, T]} S^1(t) \right\} \\ & \text{s.c.} \quad \max_{t \in [0, T]} S^2(t) \leq \bar{S} \end{aligned} \tag{14}$$

On veut résoudre ce problème par programmation dynamique. Pour simplifier la complexité du problème on fait l'hypothèse suivante :  $\forall t \in [0, T], z_{M_1}(t) \leq u(t)$  et  $z_{M_3}(t) \leq z_1(t - \tau_1 - \tau_3)$ . Sous cette hypothèse, qui revient à dire que les médecins ont toujours du travail à faire, on a les équations simplifiées suivantes :

$$\begin{aligned} z_1(t) &= z_{M_1}(t) \\ z_3(t) &= z_{M_3}(t) \\ z_{M_1}(t) + z_{M_3}(t) &= N^{med} + z_{M_1}(t - \tau_1) + z_{M_3}(t - \tau_3) \end{aligned}$$

et

$$\begin{aligned} \forall t \in [0, T], S^1(t) &= u(t) - z_{M_1}(t) \\ \forall t \in [0, T], S^2(t) &= z_{M_1}(t - \tau_1 - \tau_2) - z_{M_3}(t) \end{aligned}$$

La troisième équation va poser un problème pour la programmation dynamique : si l'on contrôle

$z_{M_1}(t)$  alors pour en déduire  $z_{M_3}(t)$ , il faut avoir accès à l'historique de  $z_{M_1}$  avant l'instant  $t$ .

Soit  $z_{M_1}$  un contrôle défini jusqu'à  $t \in [0, T]$ , on pose la quantité

$$Z_{M_1}(t) = \begin{pmatrix} z_{M_1}(t-1) \\ z_{M_1}(t-2) \\ \vdots \\ z_{M_1}(0) \end{pmatrix}$$

, avec la convention que  $Z_{M_1}(0)$  est le vecteur vide. Par ailleurs, on note

$$Z_{M_1}(t) + \tilde{z}_{M_1}(t) = \begin{pmatrix} \tilde{z}_{M_1}(t) \\ z_{M_1}(t-1) \\ z_{M_1}(t-2) \\ \vdots \\ z_{M_1}(0) \end{pmatrix}$$

On introduit alors la variable  $S_{Z_{M_1}(t)}^t$ , définie pour  $t \in [0, T]$  par :

$$S_{Z_{M_1}(t)}^t = \begin{cases} \min_{z_{M_1}(t) \in [0, u(t)]} \left\{ \max \left( u(t) - z_{M_1}(t), S_{Z_{M_1}(t)+z_{M_1}(t)}^{t+1} \right) \right\} & \text{si } (\mathcal{C}_{Z_{M_1}(t)}^t) \\ +\infty & \text{sinon.} \end{cases}$$

où la condition  $(\mathcal{C}_{Z_{M_1}(t)}^t)$  est liée à la contrainte  $\max_{t \in [0, T]} S^2(t) \leq \bar{S}$  :

$$z_{M_1}((t-1) - \tau_1 - \tau_2) + \sum_{k=0}^{k^*} z_{M_1}((t-1) - k\tau_3) - z_{M_1}((t-1) - \tau_1 - k\tau_3) \leq \bar{S} + (k^* + 1)N^{med}$$

et

$$k^* = \left\lfloor \frac{(t-1) - \tau_1 - \tau_2}{\tau_3} \right\rfloor$$

On impose comme condition limite à  $t = T + 1$ , pour tout  $Z_{M_1}(T + 1)$ :

$$S_{Z_{M_1}(T+1)}^{T+1} = \begin{cases} 0 & \text{si } (\mathcal{C}_{Z_{M_1}(T+1)}^{T+1}) \\ +\infty & \text{sinon.} \end{cases}$$

En réalité, ce problème est difficile à traiter avec de la programmation dynamique, car les choix au temps  $t$  ont des conséquences au temps  $t + \tau_1 + /tau_2$  et  $t + \tau_3$ . Cela demande donc de stocker beaucoup trop d'informations, ou alors de restreindre les temps caractéristiques à des petites valeurs.

## 4 Autre approche de la modélisation du stock

### 4.1 Notations

Le but de cette section est de proposer une autre modélisation du stock dans le réseau de la figure 5 et d'étudier l'évolution de ceux-ci lors de crises, elle s'inspire des travaux réalisés par Skandère SAHLI et Alban ZAMMIT [3] sur les centres d'appels d'urgence.

**Hypothèse.** *On fait l'approximation fluide pour le traitement des patients : on suppose que les patients arrivent en nombre suffisant pour assimiler le nombre de patients qui arrivent par unité de temps à un débit.*

On note alors :

- $\lambda$  le flux nominal d'arrivée des patients
- $\Delta$  un pic de patients qui arrive en  $t = 0$
- $\Lambda$  un flux de patients en situation de crise
- $T_c$  une durée de crise pendant laquelle le flux de patients devient  $\Lambda$
- $S_1(t)$  le nombre de patients en attente avant la première consultation
- $S_3(t)$  le nombre de patients en attente avant la consultation finale
- $N_1(t)$  le nombre de médecins alloués aux premières consultations
- $N_3(t)$  le nombre de médecins alloués aux dernières consultations
- $N_1^0$  le nombre de médecins alloués aux premières consultations en situation nominale
- $N_3^0$  le nombre de médecins alloués aux dernières consultations en situation nominale
- $N$  le nombre total de médecins dans le service
- $\alpha$  la proportion de patients qui doivent faire des examens complémentaires

Une stratégie d'allocation consiste en la donnée des fonctions  $N_1(t)$  et  $N_3(t)$  (une seule fonction suffit puisque  $N_1(t) + N_3(t) = N = \text{cste}$ ).

En situation nominale, il n'y a pas de patients qui s'accumulent (pas de file d'attente), ce qui se traduit dans l'approximation fluide par aucune accumulation de masse. Autrement dit, en situation nominale on a :

$$\forall t, \begin{cases} S_1(t) = 0 \\ S_3(t) = 0 \end{cases}$$

Pour qu'une telle situation puisse avoir lieu, il faut que le flux d'arrivée des patients soit suffisamment petit, il faut :

$$\lambda \leq \frac{N_1^0}{\tau_1} \quad (15)$$

De plus, la répartition des médecins entre la première consultation et la contre-visite doit être telle qu'aucun patient ne se retrouve en attente avant la contre-visite. Cette condition impose que le flux pouvant être traité par les médecins en 3 soit supérieur au maximum du flux pouvant arriver en 3 :

$$\alpha \frac{N_1^0}{\tau_1} \leq \frac{N_3^0}{\tau_3}$$

Puisque l'on cherche une répartition optimale, on prend le cas d'égalité de cette condition, sinon il y aurait des médecins qui ne feraient en permanence rien 3 et on devrait les affecter en 1, ce qui revient à rechercher le cas d'égalité. On a donc :

$$N_3^0 = \alpha \frac{\tau_3}{\tau_1} N_1^0 \text{ et } N_1^0 = \frac{\tau_1}{\tau_1 + \alpha \tau_3} N \quad (16)$$

La condition (15) devient alors :

$$\lambda \leq \frac{N}{\tau_1 + \alpha \tau_3}$$

Dans toutes la suite du rapport, le flux nominal d'arrivée des patients  $\lambda$  vérifiera donc cette condition.

**Hypothèse.** *On suppose que les internes ne peuvent faire que les premières consultations et qu'en situation nominale, les médecins qui font la première visite ne sont que des internes. Ainsi :*

$$N_1^0 = N^{int}$$

avec  $N^{int}$  le nombre d'internes

Concernant la stratégie d'allocation, les médecins seniors peuvent donc venir aider pour faire des premières visites, alors que les internes ne peuvent pas faire de contre visite.

## 4.2 Modélisation d'une crise coup de bélier

Dans cette section, on considère un crise dite en **coup de bélier**. Cette crise peut être modélisée par une arrivée instantanée de patients, identique à une fonction de Dirac. On note  $\Delta$  le nombre de patients supplémentaires survenant en  $t = 0$ , et on cherche à modéliser les stocks jusqu'au retour à la normale.

### 4.2.1 Sans stratégie d'allocation

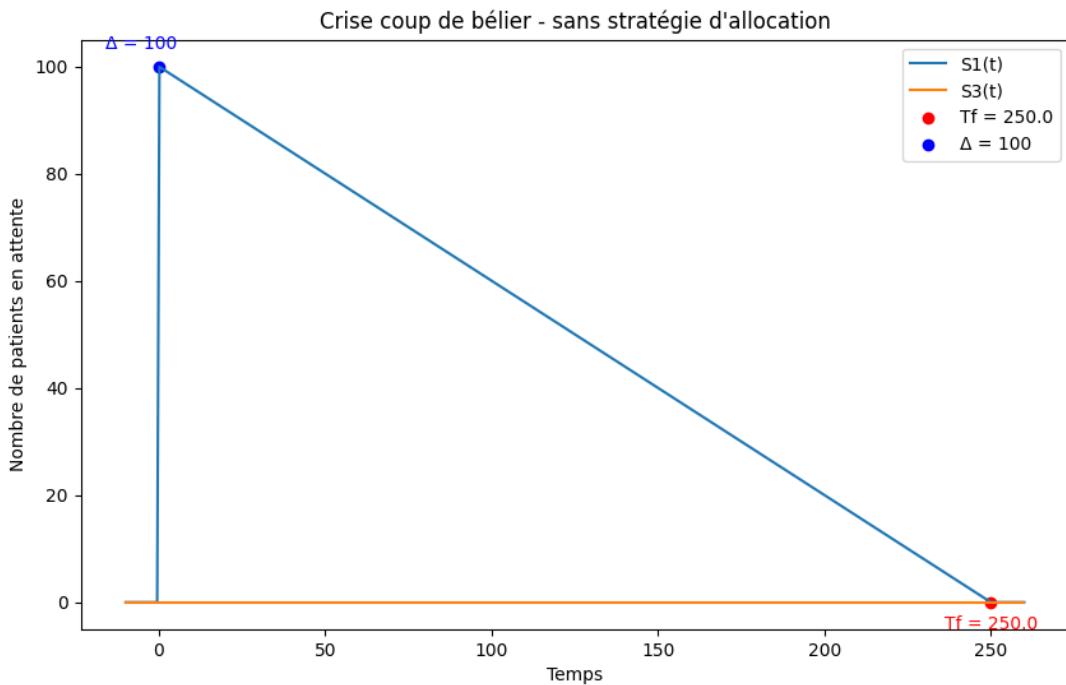


Figure 7: Représentation des stocks lors de la crise

On traite la crise en gardant  $N_1^0 = N_1^0$  et  $N_3(t) = N_3^0$ .

On écrit la variation de  $S_1$  qui est autre que le flux de patients entrants moins le flux de patients sortants :

$$\dot{S}_1 = \lambda - \frac{N_1}{\tau_1} \leq 0 \text{ d'après la condition (15)}$$

Comme  $S_1(t = 0) = \Delta$ , on a:

$$S_1(t) = \Delta + (\lambda - \frac{N_1}{\tau_1})t$$

Le temps  $T_f$  de retour à la normale (temps mis pour "absorber la crise") vérifie alors  $S_1(t = T_f) = 0$ , c'est-à-dire :

$$T_f = \frac{\Delta}{\frac{N_1}{\tau_1} - \lambda}$$

La figure 7 résume la dynamique que nous avons mis en évidence pour une crise coup de bélier en configuration frontale.

#### 4.2.2 Avec une stratégie d'allocation qui favorise le traitement des patients en 1

On cherche une stratégie qui permet de traiter le plus vite possible les patients en 1 sans que le nombre de patients qui attendent en 3 ne dépasse un certain  $\bar{S}$  fixé par la dimension du service, c'est-à-dire  $S_3(t) \leq \bar{S}$ . La stratégie d'allocation est la suivante :

- Dans une première phase, jusqu'au temps  $T_1$  à déterminer, on déplace tous les médecins de 3 vers 1 pour absorber le pic de patients  $\Delta$ .  $T_1$  est déterminé de manière à être le plus grand possible tout gardant que  $S_3(t)$  reste tout le temps inférieur à  $\bar{S}$ .
- Dans une seconde phase, les médecins séniors reviennent en 3 pour les patients qui ont nécessité des examens complémentaires pour éviter que le stock  $S_3$  ne dépasse  $\bar{S}$ .

Comme le flux de patient traité en 1 à l'instant  $t$  n'est ressenti en 3 qu'à l'instant  $t + \tau_1 + \tau_2$ , les médecins doivent anticiper cela pour leur retour en 3. C'est pourquoi, ils ne peuvent pas simplement retourner en 3 lorsque  $S_3(T_1) = \bar{S}$  car ils ne seraient pas assez nombreux pour traiter le flux qui arrive car  $\frac{N_3^0}{\tau_3} < \alpha \frac{N_1^0 + N_3^0}{\tau_1}$  et donc juste après  $T_1$  on aurait  $S_3 > \bar{S}$ . Ainsi la condition sur  $T_1$  est :

$$S_3(T_1 + \tau_1 + \tau_2) = \bar{S} \quad (17)$$

**Proposition 4.1.** *Cette stratégie est celle qui minimise le temps de retour à la normale  $T_f$  sous la contrainte  $S_3 \leq \bar{S}$*

Pour  $t \in [0, \tau_1 + \tau_2]$ , on a :

$$\begin{cases} S_1(t) = \Delta + (\lambda - \frac{N_1^0 + N_3^0}{\tau_1})t \\ S_3(t) = \alpha \lambda t \end{cases}$$

On remarque alors qu'en fonction des valeurs numériques, on peut déjà avoir besoin que les médecins reviennent en 3 dès cet intervalle de temps, c'est-à-dire  $T_1 \leq \tau_1 + \tau_2$ .

On se place par exemple dans le cas  $T_1 \geq \tau_1 + \tau_2$ :

Pour  $t \in [\tau_1 + \tau_2, T_1]$ , on a :

$$\begin{cases} S_1(t) = \Delta + (\lambda - \frac{N_1^0 + N_3^0}{\tau_1})t \\ S_3(t) = \alpha \lambda (\tau_1 + \tau_2) + \alpha \frac{N_1^0 + N_3^0}{\tau_1} (t - (\tau_1 + \tau_2)) \end{cases}$$

Pour  $t \in [T_1, T_1 + \tau_1 + \tau_2]$ , on a :

$$\begin{cases} S_1(t) = \Delta + (\lambda - \frac{N_1^0 + N_3^0}{\tau_1})T_1 + (\lambda - \frac{N_1^0}{\tau_1})(t - T_1) \\ S_3(t) = \alpha \lambda (\tau_1 + \tau_2) + \alpha \frac{N_1^0 + N_3^0}{\tau_1} (T_1 - (\tau_1 + \tau_2)) + (\alpha \frac{N_1^0 + N_3^0}{\tau_1} - \frac{N_3^0}{\tau_3})(t - T_1) \end{cases}$$

En résolvant l'équation (17), on trouve :

$$T_1 = \left( \bar{S} + \frac{\tau_1 + \tau_2}{\tau_3} N_3^0 - \alpha \lambda (\tau_1 + \tau_2) \right) \frac{\tau_1}{\alpha (N_1^0 + N_3^0)}$$

Pour  $t \in [T_1 + \tau_1 + \tau_2, T_f]$ , on a :

$$\begin{cases} S_1(t) = \Delta + (\lambda - \frac{N_1^0 + N_3^0}{\tau_1}) T_1 + (\lambda - \frac{N_1^0}{\tau_1})(t - T_1) \\ S_3(t) = S_3(T_1 + \tau_1 + \tau_2) + (\alpha \frac{N_1^0}{\tau_1} - \frac{N_3^0}{\tau_3})(t - (T_1 + \tau_1 + \tau_2)) = \bar{S} \end{cases}$$

En résolvant l'équation  $S_1(T_f) = 0$ , on trouve :

$$T_f = T_1 + \frac{\Delta + (\lambda - \frac{N_1^0 + N_3^0}{\tau_1}) T_1}{\frac{N_1^0}{\tau_1} - \lambda} = \frac{\Delta - \frac{N_3^0}{\tau_1} T_1}{\frac{N_1^0}{\tau_1} - \lambda}$$

La figure 8 représente l'évolution de  $S_1(t)$  et  $S_3(t)$  avec les pentes (qui ont donc la dimension de flux) indiquée au dessus des courbes.

## Crise coup de bélier

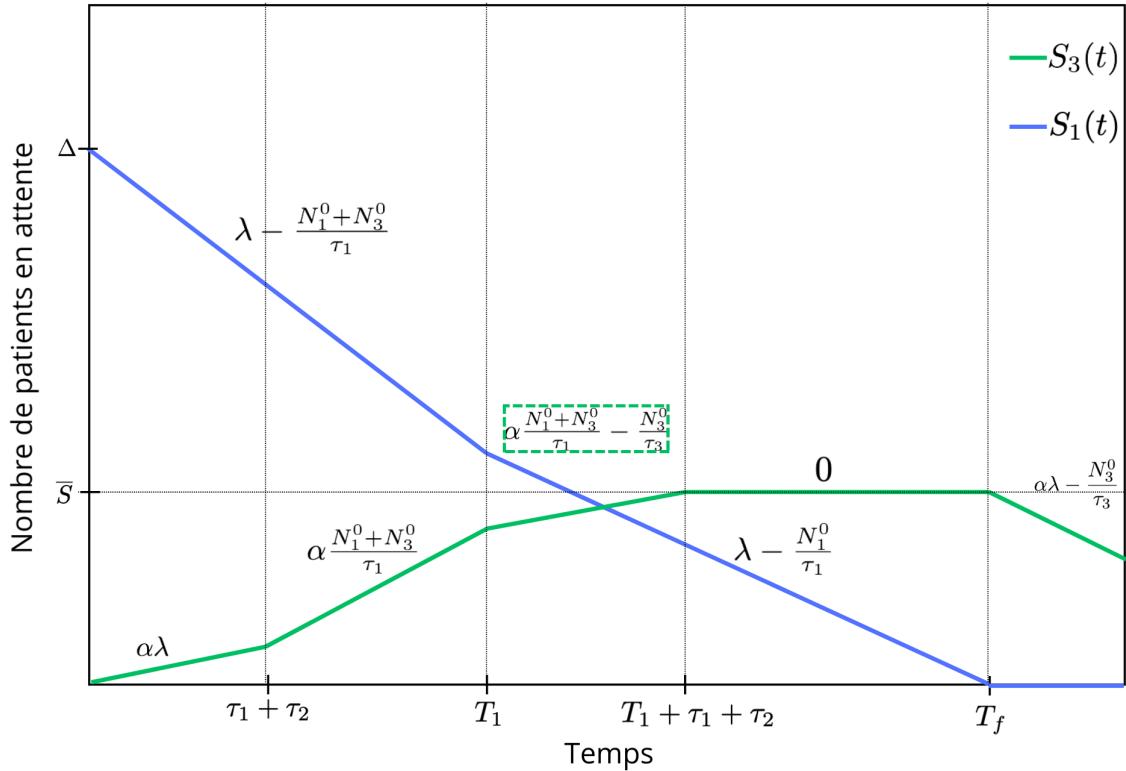


Figure 8: Représentation des stocks lors de la crise

Il y a de plus un cas où  $T_1$ , n'arrive jamais ce qui correspond à la situation où tous les personnes arrivées pendant la crise ont été traitées avec  $N_1^0 + N_3^0$  médecins sans que le stock  $S_3(t)$  n'ait pu atteindre  $\bar{S}$ . Ce dernier ainsi que le cas  $T_1 \leq \tau_1 + \tau_2$  ne seront pas détaillés mais se traitent de la même manière que le cas d'exemple ci-dessus.

## 4.3 Modélisation d'une crise longue

### 4.3.1 Modélisation

La modélisation précédente d'une crise comme un pic de patient instantané est adéquate pour les crises très courtes cependant elle ne convient pas à des crises plus longues. C'est pourquoi, nous modélisons maintenant une crise comme un **créneau**. Nous avons toujours en temps normal un flux d'appel nominal égal à  $\lambda \leq \frac{N_1^0}{\tau_1}$ , mais pendant un temps de crise  $T_c$ , le flux nominal d'appels entrant devient  $\Lambda > \frac{N_1^0}{\tau_1}$  (sinon le service est capable de gérer cette crise de manière normale). La figure 9 illustre une telle situation.

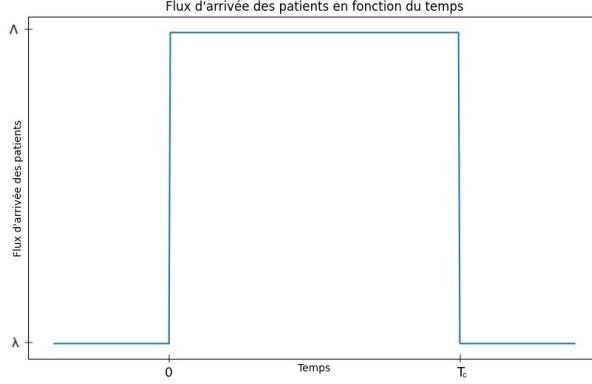


Figure 9: Modélisation d'une crise longue par un créneau

### 4.3.2 Remarques préliminaires sur la gestion d'une crise longue

On adopte une stratégie similaire à celle que nous avons développé pour la crise coup de bétier. On envoie toujours un certain nombre de médecins en 1 jusqu'à un temps  $T_1$  pour privilégier le traitement des nouveaux patients tout en maintenant le fait que  $S_3 \leq \bar{S}$ , ensuite on fait revenir les médecins en 3.  $T_1$  est comme précédemment calculé de manière à ce que  $S_1(T_1 + \tau_1 + \tau_2) = \bar{S}$ .

On est alors amené à séparer plusieurs cas de figure (qui sont illustrés par la figure 10) :

- Le premier cas est le cas d'une crise majeure vérifiant  $\frac{N_1^0 + N_3^0}{\tau_1} \leq \Lambda$ . On doit alors envoyer tous les médecins en 1 et il ne reste plus de médecin en 3.
- Le deuxième cas est le cas d'une crise moyenne vérifiant  $\frac{N_1^0}{\tau_1} \leq \Lambda \leq \frac{N_1^0 + N_3^0}{\tau_1}$ . On envoie alors le bon nombre de médecin en 1 pour traiter le flux de patients arrivant  $\Lambda$  et il reste des médecins en 3 qui peuvent faire des contre visites.

Seule l'étude de la crise moyenne sera détaillée dans la sous-partie suivante car c'est dans ce cas que nous nous plaçons pour revenir à notre problème de staffing dans la partie 5. Cependant, le cas de la crise majeure se traite de façon similaire.

### 4.3.3 Gestion de la crise moyenne

Comme décrit précédemment, on envoie  $(\Lambda - \frac{N_1^0}{\tau_1})\tau_1 = \Lambda\tau_1 - N_1^0$  en 1 jusqu'à  $T_1$ .

Pour  $t \in [0, \tau_1 + \tau_2]$  :

$$\begin{cases} S_1(t) = 0 \\ S_3(t) = (\alpha\lambda - \frac{N_1^0 + N_3^0 - \Lambda\tau_1}{\tau_3}) + t \end{cases}$$

2 types de crises longues - crise moyenne et crise majeure

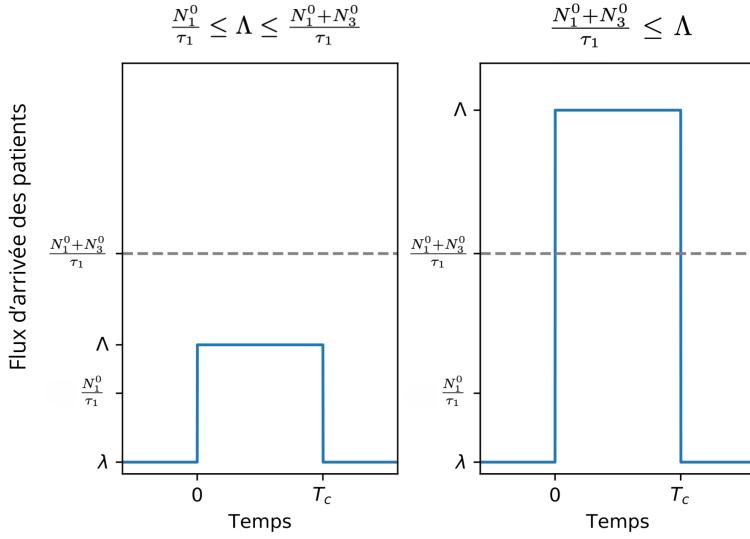


Figure 10: Illustration de deux types de crises longues

La partie positive viens provient de la disjonction de cas en fonction de si les médecins restant en 3 sont assez nombreux pour traiter les patients provenant de l'ancien flux  $\lambda$  ou non. On note maintenant  $m := (\alpha\lambda - \frac{N_3^0 + N_1^0 - \Lambda\tau_1}{\tau_3})^+$ .

Pour  $t \in [\tau_1 + \tau_2, T_1]$  :

$$\begin{cases} S_1(t) = 0 \\ S_3(t) = m(\tau_1 + \tau_2) + (\alpha\Lambda - \frac{N_3^0 + N_1^0 - \Lambda\tau_1}{\tau_3})(t - \tau_1 - \tau_2) \end{cases}$$

Pour  $t \in [T_1, T_1 + \tau_1 + \tau_2]$  :

$$\begin{cases} S_1(t) = (\Lambda - \frac{N_1^0}{\tau_1})(t - T_1) \\ S_3(t) = m(\tau_1 + \tau_2) + (\alpha\Lambda - \frac{N_3^0 + N_1^0 - \Lambda\tau_1}{\tau_3})(T_1 - \tau_1 - \tau_2) + (\alpha\Lambda - \frac{N_3^0}{\tau_3})(t - T_1) \end{cases}$$

En résolvant l'équation  $S_3(T_1 + \tau_1 + \tau_2) = \bar{S}$ , on trouve :

$$T_1 = \tau_1 + \tau_2 + \frac{\bar{S} - m(\tau_1 + \tau_2) - (\alpha\Lambda - \frac{N_3^0}{\tau_3})(\tau_1 + \tau_2)}{\alpha\Lambda - \frac{N_3^0 + N_1^0 - \Lambda\tau_1}{\tau_3}}$$

Pour  $t \in [T_1 + \tau_1 + \tau_2, T_c]$  :

$$\begin{cases} S_1(t) = (\Lambda - \frac{N_1^0}{\tau_1})(t - T_1) \\ S_3(t) = \bar{S} \end{cases}$$

Pour  $t \in [T_c, T_f]$  :

$$\begin{cases} S_1(t) = (\Lambda - \frac{N_1^0}{\tau_1})(T_c - T_1) + (\lambda - \frac{N_1^0}{\tau_1})(t - T_c) \\ S_3(t) = \bar{S} \end{cases}$$

En résolvant l'équation  $S_1(T_f) = 0$ , on trouve :

$$T_f = T_c + \frac{(\Lambda - \frac{N_1^0}{\tau_1})(T_c - T_1)}{\frac{N_1^0}{\tau_1} - \lambda}$$

La figure 11 représente l'évolution de  $S_1(t)$  et  $S_3(t)$  avec les pentes (qui ont donc la dimension de flux) indiquée au dessus des courbes.

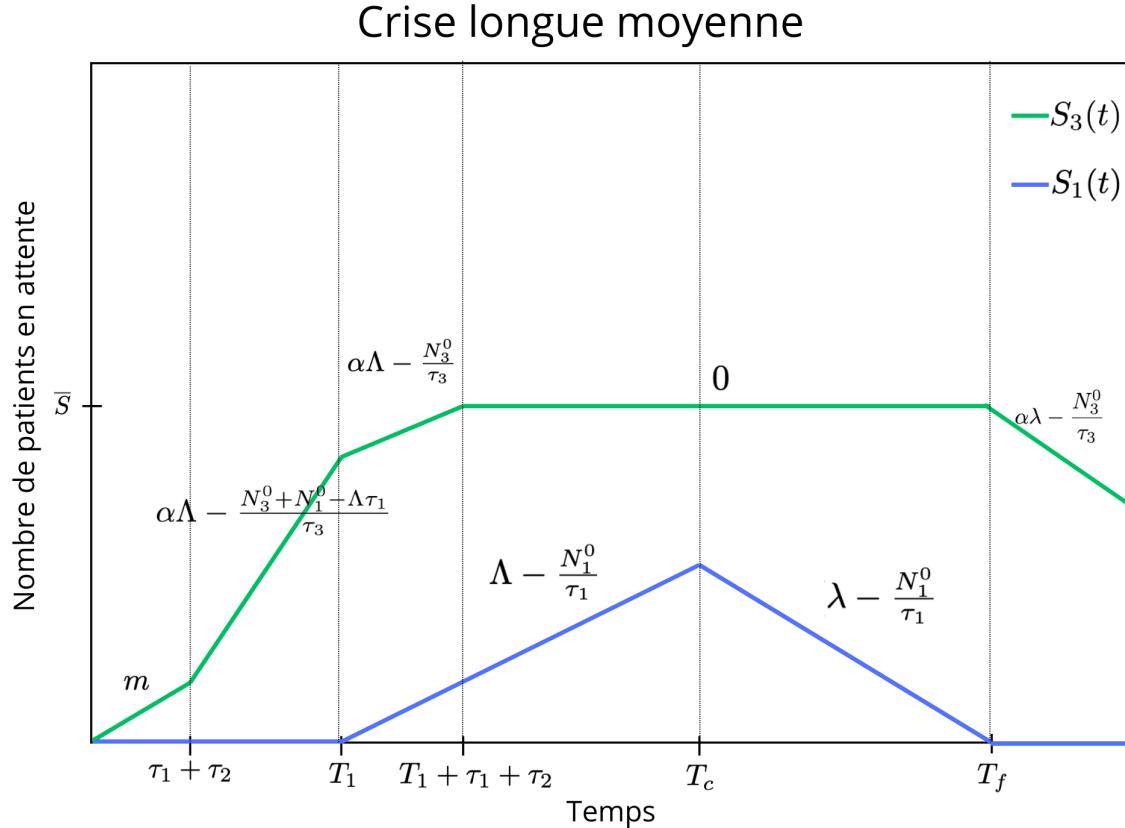


Figure 11: Représentation des stocks lors de la crise longue moyenne

Avec cette modélisation on peut alors obtenir le temps de retour à la normale ainsi que l'évolution du nombre de patient en attente.

## 5 Retour au problème de staffing

Pour revenir au problème de staffing, on se place tout d'abord dans un cadre simple.

**Hypothèse.** *Le nombre de médecins  $N_3^0$  et d'internes  $N_1^0$  sont constants sur une journée de 24h et définis par l'affectation des médecins aux différents plannings.*

**Hypothèse.** *Le flux de patients au cours d'un journée est constitué de deux phases, une phase durant laquelle le flux vaut  $\lambda \leq \frac{N_1^0}{\tau_1}$  et une autre durant laquelle elle vaut  $\Lambda$  vérifiant  $\frac{N_1^0}{\tau_1} \leq \Lambda \leq \frac{N_1^0 + N_3^0}{\tau_1}$ .*

La figure 12 représente les deux phases du flux de patients.

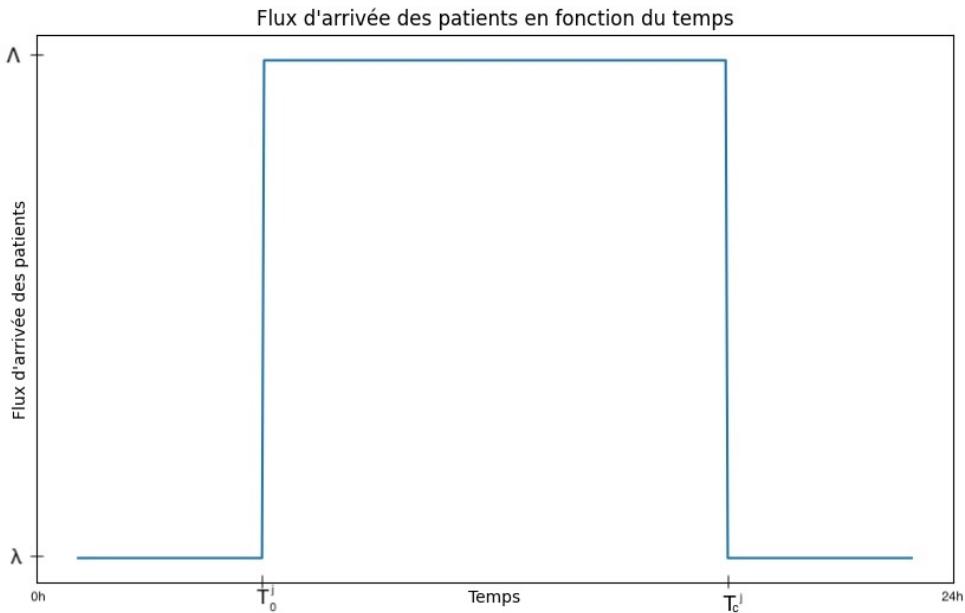


Figure 12: Flux d'arrivée des patients sur 24h lors du jour  $j$

On décide d'étudier une planification sur un horizon de une semaine. On considère ainsi un catalogue de plannings admissibles (respectant le cadre légal du travail des soignants) qu'on note  $P_i^{int}$  pour  $i \in \{1, \dots, M^{int}\}$  et  $P_i^{sen}$  pour  $i \in \{1, \dots, M^{sen}\}$ , et qui sont définis de la même manière que dans la partie 2.5.

Nos variables sont alors  $(x_{ij}^{int})$  pour  $(i, j) \in \{1, \dots, M^{int}\} \times \{1, \dots, N_{tot}^{int}\}$  et  $(x_{ij}^{sen})$  pour  $(i, j) \in \{1, \dots, M^{sen}\} \times \{1, \dots, N_{tot}^{sen}\}$  définis par :

$$x_{ij}^{int/sen} = \begin{cases} 1 & \text{si le médecin } j \text{ est affecté au planning } i \\ 0 & \text{sinon} \end{cases}$$

Comme indiqué dans la partie 2.5, la donnée de  $X = (x_{ij}^{int}, x_{i'j'}^{sen})$  définit alors les valeurs de  $N_1^0$  et  $N_3^0$ . On impose cependant la condition que pour chaque jour, on a :

$$N_3^0 \geq \alpha \frac{\tau_3}{\tau_1} N_1^0$$

Comme chaque médecin ou interne est affecté exactement à un planning, on a :

$$\forall j \in \{1, \dots, N_{tot}^{int}\}, \sum_i x_{ij}^{int} = 1$$

et

$$\forall j \in \{1, \dots, N_{tot}^{sen}\}, \sum_i x_{ij}^{sen} = 1$$

Concernant le critère il y a de nombreux choix possibles, on peut par exemple choisir de minimiser l'intégrale de  $S_1(t)$  sur tout la semaine, ou les maximums de  $S_1(t)$  sur chaque jour de la semaine.

Pour l'illustration on choisit de minimiser la somme sur chaque jour de la semaine du maximum de  $S_1(t)$  qui est atteint en  $T_c^j$  où  $j$  est le jour de la semaine. Cela s'interprète comme la minimisation du nombre de patients dans la salle d'attente avant d'être vu pour la première fois.

Le problème est alors :

$$\begin{aligned}
\text{Minimiser : } & \sum_{jour} S_1(T_c^{jour}) \\
\text{Sous les contraintes : } & \text{pour chaque jour, } N_3^0 \geq \alpha \frac{\tau_3}{\tau_1} N_1^0 \\
& \forall j \in \{1, \dots, N_{tot}^{int}\}, \sum_i x_{ij}^{int} = 1 \\
& \forall j \in \{1, \dots, N_{tot}^{sen}\}, \sum_i x_{ij}^{sen} = 1 \\
& \text{pour chaque jour, } \forall t, N_t^{int} = N_1^{0,jour} \\
& \text{pour chaque jour, } \forall t, N_t^{sen} = N_3^{0,jour}
\end{aligned} \tag{18}$$

Les deux dernières contraintes sont dues au fait que le nombre de médecins  $N_3^0$  et d'internes  $N_1^0$  sont constants sur une journée de 24h.

Pour chaque jour  $S_1(T_c^{jour})$  est une fonction de  $N_1^{0,jour}$  qui est lui même donné par le staffing.

## Conclusion

Ce travail a permis d'explorer pour la première fois le système complexe des services d'urgences, en proposant plusieurs modélisations du parcours des patients, tant dans un régime stationnaire que dans une gestion des patients en attente. À travers ces approches, nous avons étudié l'allocation des ressources humaines, les flux de patients et l'optimisation des temps d'attente, en fournissant des résultats intéressants pour la gestion dynamique de ces services essentiels.

Les modélisations réalisées ont permis de mettre en lumière des mécanismes clés du système des urgences, offrant ainsi une première base pour comprendre les dynamiques internes de ces services et les facteurs influençant leur performance. Nous avons montré que des stratégies d'allocation des médecins adaptées peuvent améliorer les temps d'attente et la gestion des flux de patients.

Cependant, plusieurs pistes demeurent à explorer pour affiner et concrétiser ces résultats. L'étude des plannings admissibles pour les personnels soignants et l'obtention de valeurs concrètes des paramètres du système permettraient de rendre ces modèles encore plus applicables dans des contextes réels. De plus, la prise en compte des différents degrés d'urgence constitue un axe d'enrichissement important pour mieux adapter le modèle aux réalités du terrain.

Le système des urgences demeure un problème complexe et dynamique, nécessitant une exploration continue et des ajustements méthodologiques pour aboutir à des solutions concrètes et efficaces, adaptées aux défis pratiques du terrain.

## References

- [1] Xavier ALLAMIGEON, Vianney BOEUF et Stéphane GAUBERT, Performance evaluation of an emergency call center: Tropical polynomial systems applied to timed Petri nets, 2015.
- [2] Marin BOYET, Piecewise affine dynamical systems applied to the performance evaluation of emergency call centers, 2022.
- [3] Skandère SAHLI et Alban ZAMMIT, Optimisation de la Réponse Coordonnée aux Appels d’Urgence 17-18-112 en Région Parisienne, *Projet de recherche de troisième année MAP511*, 2019.
- [4] Pascal BENCHIMOL, Modélisation du service des urgences de l'Hôtel-Dieu, 2009.