Apprentissage par Renforcement

Charles Sutty
Martin Reix

Février 2026

Table des matières

1 Introduction **1**

2 Modélisation du problème **2**
2.1 États et actions 2
2.2 Explication des équations 3
2.3 Définition de la fonction de coûts 3
2.4 Valeur numérique des paramètres 4
2.4.1 Paramètres sur les états 4
2.4.2 Estimations économiques 4

3 Résolution du problème par renforcement **5**
3.1 Astuces de calcul 5
3.2 Algorithmes 5
3.3 Implémentation 7

4 Evaluations **8**
4.1 Politiques des heuristiques 9
4.2 Politiques des algorithmes 10
4.3 Evolution de l'épidémie 12

5 Considération de coûts économiques **12**
5.1 Conclusion 13

1 Introduction

Lors de la pandémie mondiale du Covid-19 de nombreuses mesures ont été prises par les différents gouvernements afin de réguler la propagation de l'épidémie. De nombreuses modélisations ont été faites afin d'appréhender l'évolution de la maladie. Ces simulations ont été une base qui a servi par la suite sur la détermination des politiques publiques prises par le gouvernement. Parmi ces mesures, nous étudions en particulier les effets du confinement et de la vaccination.

Ces deux mesures ont fait face à de nombreux débats sur leur impact, tant en termes de privation de liberté, que de leur effet sur l'épidémie, mais aussi sur l'économie française. Lors de la pandémie, le gouvernement a ainsi dû faire des compromis entre pertes économiques et vies humaines.

Dans ce projet, nous tentons d'étudier comment les algorithmes d'apprentissage par renforcement peuvent être utilisés pour mieux guider les politiques publiques afin de répondre aux différents enjeux provoqués par l'apparition d'une souche virale comme le coronavirus.

De nombreuses études et modèles ont été proposés suite à la crise de 2020 et nous nous inspirons de ces modélisations dans notre projet.

2 Modélisation du problème

2.1 États et actions

La modélisation la plus simple pour représenter notre problème est le format SIR. D'autres études proposent aussi le format SEIR. Enfin, nous nous proposons d'étudier la modélisation SEIRD, qui nous semble plus complète.

Ce modèle présente un état sous forme de 5-uplet avec les éléments :

— **S (Susceptible)** : La population susceptible d'être atteinte par la maladie.

— **E (Exposed)** : La population exposée à la maladie, qui en est porteuse mais qui n'a pas encore les symptômes. Cet élément nous a paru pertinent avec l'expérience du statut de 'cas contact' qui voulait prévenir la prolifération de personnes qui ne se savaient pas déjà malades.

— **I (Infected)** : La population infectée par la maladie et qui est susceptible de la transmettre.

— **R (Restablished)** : Le nombre de personnes rétablies, que l'on considère immunisées par la maladie. Cela peut alors aussi dépendre de la vaccination.

— **D (Deceased)** : Afin d'étudier une évolution de la population de manière dynamique, nous introduisons la population décédée depuis l'origine de l'étude.

La population totale à l'instant $t \in \mathbb{N}$ est notée $N(t)$ et suit :

$$N(t) = S(t) + E(t) + I(t) + R(t) + D(t)$$

avec $N(0) = N_0 = 1$ et $D(0) = D_0 = 0$.

Ces différents états, nous amènent à poser les équations suivantes :

$$\frac{dS}{dt} = -\beta(1 - u_{conf})SI - v(t) \tag{1}$$

$$\frac{dE}{dt} = \beta(1 - u_{conf})SI - \sigma E \tag{2}$$

$$\frac{dI}{dt} = \sigma E - \gamma I \tag{3}$$

$$\frac{dR}{dt} = \gamma(1 - \mu)I + v(t) \tag{4}$$

$$\frac{dD}{dt} = \gamma\mu I \tag{5}$$

avec la contrainte logistique sur la vaccination :

$$v(t) = \min\big(u_{\text{vacc}}(t) \cdot v_{\max},\ S(t)\big) \tag{6}$$

où

— β : Taux de transmission sans confinement.

— σ : Taux d'incubation = 1/ temps d'incubation

— γ : Taux de guérison = 1/ temps de guérison

— μ : Taux de mortalité

— R_0 : Taux de reproduction de base = β/γ.

Notre contrôle, ce sur quoi nous avons un impact est :

— u_{conf} : le niveau de confinement entre 0 et 1. 0 correspond au fonctionnement normal et 1 à un confinement total de la population. Nous le confondons avec le niveau d'intention proposé dans le sujet.

— $u_{vacc} \in [0, 1]$: la fraction de la capacité logistique maximale de vaccination $v_{\max}$ que l'on choisit de mobiliser. Le flux réel de vaccination $v(t)$ est donné par l'équation (6) : il est plafonné par la capacité logistique $v_{\max}$ et saturé par la population susceptible restante $S(t)$. On ne peut pas vacciner plus de personnes qu'il n'y en a de susceptibles.

2.2 Explication des équations

L'équation (1) indique l'évolution de la population saine. Nous voyons que le premier terme $\beta(1-u_{conf})SI$ correspond à la contamination par les individus infectés. Nous observons la conservation de ce flux vers la population E dans l'équation (2). Le deuxième terme $v(t)$ correspond au flux de vaccination (6), qui fait passer les éléments sains directement en rétablis (immunisés). Ce flux est borné par la capacité logistique journalière $v_{\max}$ et par la population susceptible restante $S(t)$, ce qui modélise correctement la contrainte matérielle : on ne peut vacciner qu'un nombre limité de personnes par jour, et on ne peut pas vacciner plus que les susceptibles disponibles. Nous retrouvons ce flux dans l'équation (4).

L'équation (3) montre aussi la transformation des exposés vers des infectés I à la suite d'un temps régi par σ. Nous retrouvons ce flux dans l'équation (3). Les infectés peuvent soit se rétablir avec un taux γI, soit décéder avec un taux μ et c'est ce que nous retrouvons en (5). Nous avons en (4) les rétablis qui ne sont pas décédés, ou qui ont été vaccinés.

2.3 Définition de la fonction de coûts

Afin de représenter des contraintes réalistes, nous devons considérer une fonction de coût qui rend compte des compromis nécessaires à l'élaboration d'une politique sanitaire. En prenant l'expérience du Covid-19, il y avait des considérations de liberté, de protection des personnes mais également des coûts logistiques et un impact immense sur l'économie. La France a fait des choix de confinements forts et de vaccination massive impliquant un « quoi qu'il en coûte » qui a encore des impacts sur l'économie. Bien que cynique, il y a une nécessité de quantifier économiquement les pertes de vies humaines. De plus, nous devons prendre en compte les coûts financiers de soins des personnes malades qui, dans le cas français, sont des coûts pour l'État, ainsi que les arrêts maladies impliqués.

La fonction de coût $\mathcal{L}(s_t, a_t)$ qui dépend de l'état s_t et le niveau d'action a_t à l'instant t que nous voulons minimiser est définie par l'équation suivante :

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{eco} + \mathcal{L}_{log} + \mathcal{L}_{deaths} + \mathcal{L}_{hosp} \\ &= C_{eco} \times u_{conf}^2 + C_{vacc} \times \frac{v(t)}{v_{\max}} + C_{vie}\Delta D + C_{hosp} \times I_t \end{aligned} \tag{7}$$

Le terme $\mathcal{L}_{eco} = C_{eco} \times u_{conf}^2$ correspond au coût économique du confinement. Nous considérons C_{co} les pertes économiques pour un niveau maximal de confinement. Nous multiplions ce facteur par le carré du niveau de confinement car nous considérons qu'à faible niveau de confinement, une augmentation n'a pas beaucoup d'impact sur l'économie tandis que à partir d'une certaine valeur, ajouter encore plus de confinement a un impact sur des systèmes critiques (livraisons, industrie, ...) ou à haute valeure ajoutée. Nous aurions pu considérer un facteur α pour u_{conf}^{α} mais une évolution quadratique nous semble déjà adaptée.

Le terme $\mathcal{L}_{log} = C_{vacc} \times \frac{v(t)}{v_{\max}}$ quantifie les coûts logistiques et de production de la vaccination. Le coût est proportionnel au nombre de vaccinations réellement administrées $v(t)$, normalisé par la capacité maximale $v_{\max}$. On ne paye que les doses effectivement injectées.

Le terme $\mathcal{L}_{deaths} = C_{vie}\Delta D$ quantifie économiquement la perte de vies humaines au vu de la société. Nous devons considérer la différence entre 2 instants car les décès n'arrivent qu'une fois et nous devons considérer le coût des pertes dans un intervalle de temps. Il est évident de noter que la fonction $D(t)$ ne peut être que croissante (pas de réssucitations).

Le terme $\mathcal{L}_{hosp} = C_{hosp} \times I_t$ dépend de la proportion de malades. C_{hosp} est une aggrégation des coûts indus par le fait qu'une personne soit malade (impact économique sur une entreprise

via les arrêts maladies, coûts des soins consommés par une personne infectée). Ces coûts sont directement proportionnels aux infectés.

2.4 Valeur numérique des paramètres

Afin de justifier nos paramètres par des données historiques nous choisissons d'étudier la pandémie de Covid-19 en France. Nous reproduisons ses paramètres caractéristiques et son impact sur l'économie.

2.4.1 Paramètres sur les états

Par notre propre expérience de la pandémie, nous estimons le temps d'incubation à 7 jours. Notre temps caractéristique étant en jours. Le temps de guérison devient 10 jours. Nous nous inspirons de [1] et [2] pour proposer les données suivantes :

— $t = 1$ jour
— $\sigma = 1/7 \approx 0.14$
— $\gamma = 1/10 = 0.1$
— $R_0 = 2.68 \approx 2.7 = \beta/\gamma \Rightarrow \beta = 0.27$
— $\mu = 1\% = 0.01$

2.4.2 Estimations économiques

Afin de faire des estimations des critères économiques pour la fonction de coûts, nous nous inspirons d'une publication de l'OFCE [3] : *Évaluation au 30 mars 2020 de l'impact économique* qui évalue l'impact du Covid-19 sur l'économie française.

Dans cette étude, la perte économique du Covid-19 était de $30 - 35\%$ du PIB annuel sur 2020. Si nous l'estimons à ≈ 2500 milliards d'euros pré-pandémie [1]. Nous avons alors un PIB journalier à ≈ 6.5 milliards d'euros.

Pour quantifier la valeur de la perte de vies humaines, afin de pouvoir proposer un compromis entre arrêt économique et quantité de décès, nous nous appuyons sur l'estimation admise par le gouvernement sur la valeur d'une vie française [2], qui s'élève à 3 millions d'euros.

Enfin, nous allons tenter d'estimer les pertes économiques impliquées par la production des vaccins et et par le coût des infectés sur l'économie.

Pour les vaccins, en nous basant sur des valeurs estimées pour les vaccins contre le Covid [3], on retient une valeur d'environ 21.5 € par dose pour les vaccins Moderna ou Pfizer. Si on rajoute un coût de transport et logistique (estimé à 10€ de manière assez arbitraire), on arrive, pour 2 doses de vaccin par personne, à une estimation de $2 \times 21.5 + 10 = 53$ € par personne pour la vaccination.

En multipliant par la population totale on obtient le coût de vaccination de la population entière C_{vacc}. On se basera sur cette valeur, qu'on multipliera par le taux de vaccination journalier u_{vacc} pour obtenir un coût journalier de vaccination.

On a $C_{vacc} = 6,7E7 \times 53 \approx 3,5$ milliards d'€

Pour le coût hospitalier, on trouve des estimations qui chiffrent un séjour complet pour une infection au Covid à 6272€ pour une durée de 10,8j en moyenne [4]. Cela revient à un coût

1. INSEE : `https://www.insee.fr/fr/statistiques/4500483`

2. Luc Baumstark, Benoît Dervaux, Nicolas Treich. Eléments pour une révision de la valeur statistique de la vie humaine. Commission présidée par E. Quinet. L'évaluation socio-économique des investissements publics, Commissariat général à la stratégie et à la prospective, 28 p., 2013. ⟨halshs-00958423⟩

3. Le Monde : `https://www.lemonde.fr/planete/article/2021/08/03/des-contrats-plus-exigeants-et-des-prix-plus-e 6090393_3244.html?`

4. S. Gallien et al. Évaluation des coûts des hospitalisations pour COVID-19 en France, en 2020 `https: //static.hevaweb.com/web/PDF/6ffdbba002a26-janssen-cohco-poster-jni2022-v1r7-web.pdf?`

journalier d'environ 580€ par patient. Ici on a $\gamma = 0.1$ donc la durée d'hospitalisation sera de 10j (ce qui ne change rien au coût journalier).

Finalement, les quantités économiques que nous retenons sont :

— $C_{eco} = 0.3 \times 6 = 2$ milliards d'€

— $C_{vie} = 3$ millions d'€ / personne

— $C_{vacc} = 3,5$ milliards d'€

— $C_{hosp} = 580$ €/ personne et /jour

Ces valeurs théoriques aux échelles trop éloignées ont été par la suite, lors de nos expériences adaptées pour avoir des simulations stables et intéressantes à analyser. Nous revenons par la suite sur l'évolution de ces paramètres.

Afin de renormaliser nos données, nous prenons C_{eco} comme normalisation. Nous posons alors :

— $C_{eco} = 1.0$

— $C_{vie} = \frac{cout\ vie \times population \times \mu}{C_{eco1}} = \frac{3\times10^6\times67\times10^6\times0.01}{2\times10^9} \approx 1000.$
Avec ce calcul un peu alambiqué, nous considérons le coût de décès d'1% de la population (nous nous basons sur $\mu = 1\%$). Avec d'autres choix implémentés, les résultats n'étaient pas stables ou intéressants à analyser, donc cette normalisation nous semble pertinente par la suite.

— $C_{vacc} = \frac{v_{\max} \times population \times prix\ vaccin}{C_{eco1}} = \frac{0.001\times67\times10^6\times53}{2\times10^9} \approx 0.002.$
$v_{\max} = 0.001$ est la capacité logistique maximale de vaccination (fraction de la population par jour). En pratique nous conservons $C_{vacc} = 0.02$ pour la stabilité numérique.

— $C_{hosp} = \frac{cout\ hopital \times population}{C_{eco1}} = \frac{580\times67\times10^6}{2\times10^9} \approx 20$

Ces calculs peuvent sembler parachutés mais ils nous permettent d'utiliser des valeurs stables pour notre implémentation tout en gardant un certain ordre de grandeur dans les rapports entre les constantes (en considérant la quantité multiplicatrice avec laquelle elles sont liées dans la fonction de coût).

3 Résolution du problème par renforcement

3.1 Astuces de calcul

Nous remarquons dans notre modèle que certains états sont redondants. En effet, puisque nous avons une population fixe qui est régie par :

$$N(t) = S(t) + E(t) + I(t) + R(t) + D(t) = N_0$$

nous pouvons déterminer $D(t)$ grâce au niveau de tous les autres états. De plus, lorsque nous observons l'évolution de notre problème, nous voyons qu'aucune des équations ne dépend ni de D, ni de R. L'évolution du problème dépend exclusivement de S, E, I et des variables de contrôle. Nous pouvons déterminer par exemple l'évolution du D grâce à $\gamma\mu I$ d'après (5).

3.2 Algorithmes

Pour nos calculs nous avons utilisé plusieurs algorithmes, que nous allons ensuite comparer :

Le premier est un algorithme de type SARSA (algorithme *online*), le second est un algorithme de Q-Learning (*offline*), et le dernier est un algorithme de Programmation Dynamique (DP).

Pour effectuer des opérations nous avons besoin de discrétiser les variables. On passe donc par une méthode de "*bins*" (ou *intervalles*) : On découpe les variables I, E, S en petits intervalles.

En pratique on a pris les intervalles suivants :

— S : 12 bins — $[0, 0.2, 0.4, 0.6, 0.7, 0.8, 0.85, 0.9, 0.93, 0.95, 0.97, 0.985, 1.00]$

— E, I : 12 bins — $[0, 10^{-5}, 3\times10^{-5}, 10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}, 3\times10^{-2}, 0.1, 0.3, 0.5, 1.0]$

Ensuite, on attribue à la variable I un indice qui correspond à un intervalle de valeurs (ici par exemple, si $S = 0.5$ on prend $i_I = 3$). On remarque qu'on prend un découpage non uniforme : l'épidémie ne se propage pas de manière linéaire donc il est plus judicieux de choisir une échelle logarithmique. Également, on ne prend pas un très grand nombre d'intervalles pour ne pas surcharger les calculs. Nous avons estimé que prendre 12 intervalles était un bon compromis entre précision et vitesse de calcul.

L'ensemble des états est l'ensemble des triplets $[i_S, i_I, i_E]$, ce qui nous fait un total de $12^3 = 1728$ états accessibles. En pratique certains états ne sont pas atteignables : on doit toujours avoir $S + E + I < 1$.

Pour les actions, on agit sur u_{vacc} et u_{conf} que l'on prend tous les deux parmi [0.0, 0.25, 0.5, 0.75, 1.0]^2.

Cela donne donc 25 actions réalisables.

En pratique les algorithmes fonctionnent comme suit :

Algorithm 1 Algorithme SARSA

Données : Initialiser $Q(s, a)$ arbitrairement (ex : à 0)
Données : Paramètres : N_{episodes}, α, γ, ϵ
pour *episode* $= 1$ **à** $N_{episodes}$ **faire**
  Initialiser l'état s
  Choisir l'action a depuis s en utilisant ϵ-greedy sur Q
  tant que s *n'est pas terminal* **faire**
    Prendre l'action a, observer la récompense r et l'état suivant s'
    Choisir l'action a' depuis s' en utilisant ϵ-greedy sur Q
    $Q(s, a) \leftarrow Q(s, a) + \alpha\,[r + \gamma Q(s', a') - Q(s, a)]$
    $s \leftarrow s'$
    $a \leftarrow a'$

Dans cet algorithme, à chaque étape l'agent exécute l'action qu'il a choisie, reçoit sa récompense, puis choisit immédiatement sa prochaine action. Il regarde ensuite ce qu'il va se passer avec cette action et effectue la mise à jour.

Algorithm 2 Algorithme Q-learning

Données : Initialiser $Q(s, a)$ arbitrairement
Données : Paramètres : N_{episodes}, α, γ, ϵ
pour *episode* $= 1$ **à** $N_{episodes}$ **faire**
  Initialiser l'état s
  tant que s *n'est pas terminal* **faire**
    Choisir l'action a depuis s en utilisant ϵ-greedy sur Q
    Prendre l'action a, observer r et l'état suivant s'
    $Q(s, a) \leftarrow Q(s, a) + \alpha\,[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$
    $s \leftarrow s'$

Enfin, on applique un algorithme de Programmation dynamique. On va utiliser deux méthodes différentes et les comparer.

Itération de la politique Le principe est d'alterner entre deux phases jusqu'à obtenir une politique stable :

- Évaluation de la politique : On a une politique et on calcule la somme des récompenses futures espérées pour chaque état afin d'obtenir la valeur $V(s)$ de chaque état.

- Amélioration de la politique : Pour chaque état, l'algorithme teste toutes les actions possibles et garde celle qui maximise l'espérance de gain futur. Si une action de la politique a changé pour au moins un état, on considère que la politique est instable, et on réitère.

Itération de la valeur Cette méthode combine les 2 étapes de la méthode précédente en une seule. L'algorithme boucle sur l'ensemble des états, et pour chaque état s elle met à jour la valeur $V(s)$ avec la valeur maximale atteignable parmi toutes les actions. Lorsque les valeurs V ne changent plus ou très peu, on extrait la politique.

L'équation de Bellman correspondante pour la fonction Valeur est :

$V_t(s) = \max_a \left(\hat{\mathbb{E}}[r(s, a, S')] + \gamma \hat{\mathbb{E}}[V_{t+1}(S')] \right)$

L'espérance est estimée par K simulations Monte-Carlo depuis le centre de chaque intervalle. On parle ici de programmation dynamique *approchée* car l'espace continu est discrétisé et les transitions sont estimées empiriquement.

3.3 Implémentation

Nous proposons une implémentation [5] fondée sur la bibliothèque `gym` [6] de OpenAI. Nous utilisons aussi les environnements de `Stable Baselines 3.0` [7].

Voici, ci-dessous les résultats d'entraînement des deux algorithmes (figure 1) :

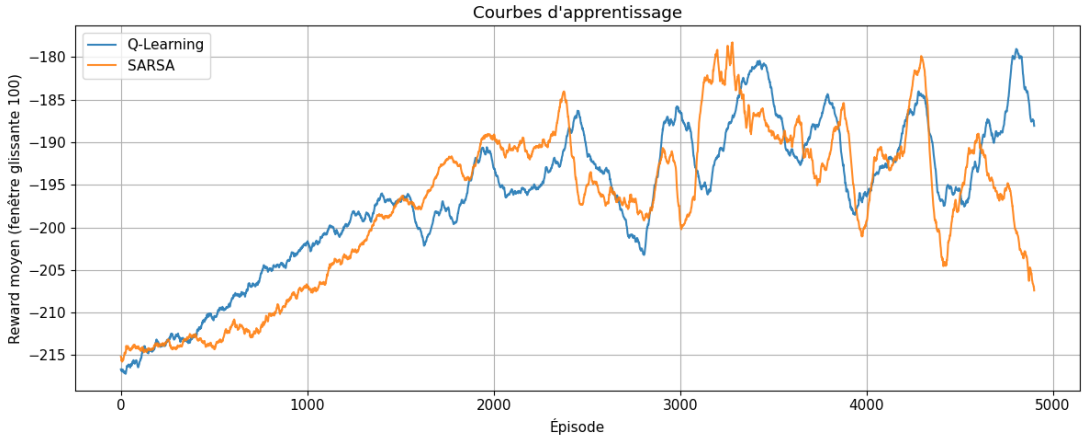

FIGURE 1 – Évolution de la récompense pour les algos SARSA et Q-Learning

Afin de juger la qualité des politiques obtenues par apprentissage par renforcement, il est indispensable de les comparer à des politiques de référence simples et interprétables. C'était notamment une des propositions de la définition du projet. Nous définissons quatre heuristiques, chacune incarnant une stratégie sanitaire archétypale. Elles couvrent un spectre allant de l'inaction totale à un pilotage adaptatif sur deux leviers simultanés.

Aucun contrôle (laissez-faire). Cette politique correspond au scénario dans lequel aucune mesure n'est prise :

$$u_{\text{conf}}(t) = 0, \qquad u_{\text{vacc}}(t) = 0 \quad \forall\, t.$$

L'épidémie se propage sans entrave et la totalité du coût provient des décès et de la charge hospitalière. Ce scénario constitue la borne supérieure du coût : toute politique raisonnable doit faire significativement mieux. C'est notre base de comparaison de nos résultats. De façon

5. `https://github.com/Charsutty/RL-pandemic-modeling`
6. https ://gymnasium.farama.org/
7. https ://stable-baselines3.readthedocs.io/en/master/

historique, c'est plus ou moins la politique prise par le Brésil, qui a voulu limiter les politiques prises par le gouvernement.

Vaccination maximale (sans confinement). On mobilise l'intégralité de la capacité logistique de vaccination sans imposer de restrictions à la population :

$$u_{\text{conf}}(t) = 0, \qquad u_{\text{vacc}}(t) = 1 \quad \forall\, t.$$

Cette heuristique isole l'effet de la vaccination seule. Elle permet de quantifier dans quelle mesure un programme vaccinal intensif peut, à lui seul, compenser l'absence de confinement. Le coût économique du confinement est nul, mais le coût logistique de vaccination est maximal. Nous limitons ici la perte économique aux coûts logistiques ; cependant, la capacité logistique limitée de vaccination restreint l'efficacité d'une telle stratégie.

Confinement par seuil. On introduit ici une politique réactive fondée sur la proportion d'infectés. Dès que celle-ci franchit un seuil critique, un confinement total est déclenché, accompagné d'un effort vaccinal modéré :

$$u_{\text{conf}}(t) = \begin{cases} 1 & \text{si } I(t) > 0.02, \\ 0 & \text{sinon,} \end{cases} \qquad u_{\text{vacc}}(t) = 0.5 \quad \forall\, t.$$

Ce type de politique « tout ou rien » s'apparente aux mesures prises lors des premiers confinements du Covid-19 en France. Cela peut être justifié par des considérations d'engorgement maximal des hôpitaux. Nous aurions pu traiter une capacité maximale des hôpitaux mais en première instance, notre modèle semble suffisant.

Double seuil adaptatif. La dernière heuristique module les deux leviers de contrôle en fonction de l'état épidémique. Le confinement est gradué selon le niveau d'infection :

$$u_{\text{conf}}(t) = \begin{cases} 0.75 & \text{si } I(t) > 0.02, \\ 0.25 & \text{si } 0.005 < I(t) \leq 0.02, \\ 0 & \text{sinon,} \end{cases}$$

tandis que l'effort vaccinal est adapté à la taille de la population encore susceptible :

$$u_{\text{vacc}}(t) = \begin{cases} 1 & \text{si } S(t) > 0.3, \\ 0.25 & \text{sinon.} \end{cases}$$

L'idée sous-jacente est double : d'une part, un confinement progressif évite les coûts économiques extrêmes d'un verrouillage total ; d'autre part, la vaccination est concentrée sur la phase où le réservoir de susceptibles est encore large, puis réduite lorsque l'essentiel de la population a déjà été immunisé ou infecté. Cette stratégie s'inspire des réactions proposées par les décideurs politiques lors de la crise du Covid-19 et propose des réactions à la fois adaptées et moins susceptibles aux aléas puisque déterministe comme nous pouvons le voir dans l'analyse des différents algorithmes.

4 Evaluations

À la suite de l'entraînement de nos différentes heuristiques et de nos modèles. Nous proposons les évaluations graphiques suivantes [8] :

8. L'esthétique de ces graphiques a été assistée par IA.

4.1 Politiques des heuristiques

Ici, les politiques prises par les heuristiques sont déterministes donc de nombreux graphiques semblent évidents. On voit notamment que la politique de confinement par seuil propose des confinements complets presque tous les 10 jours, ce qui provoque des oscillations sur le nombre représenté dans la courbe épidémique en figure 5. De manière similaire, le double seuil propose des confinements de manière régulière afin de protéger la population tout en menant une politique de vaccination *agressive* puis relâche l'effort en fin de vaccination. Nous voyons des oscillations plus souples autour du seuil dans la figure 5.

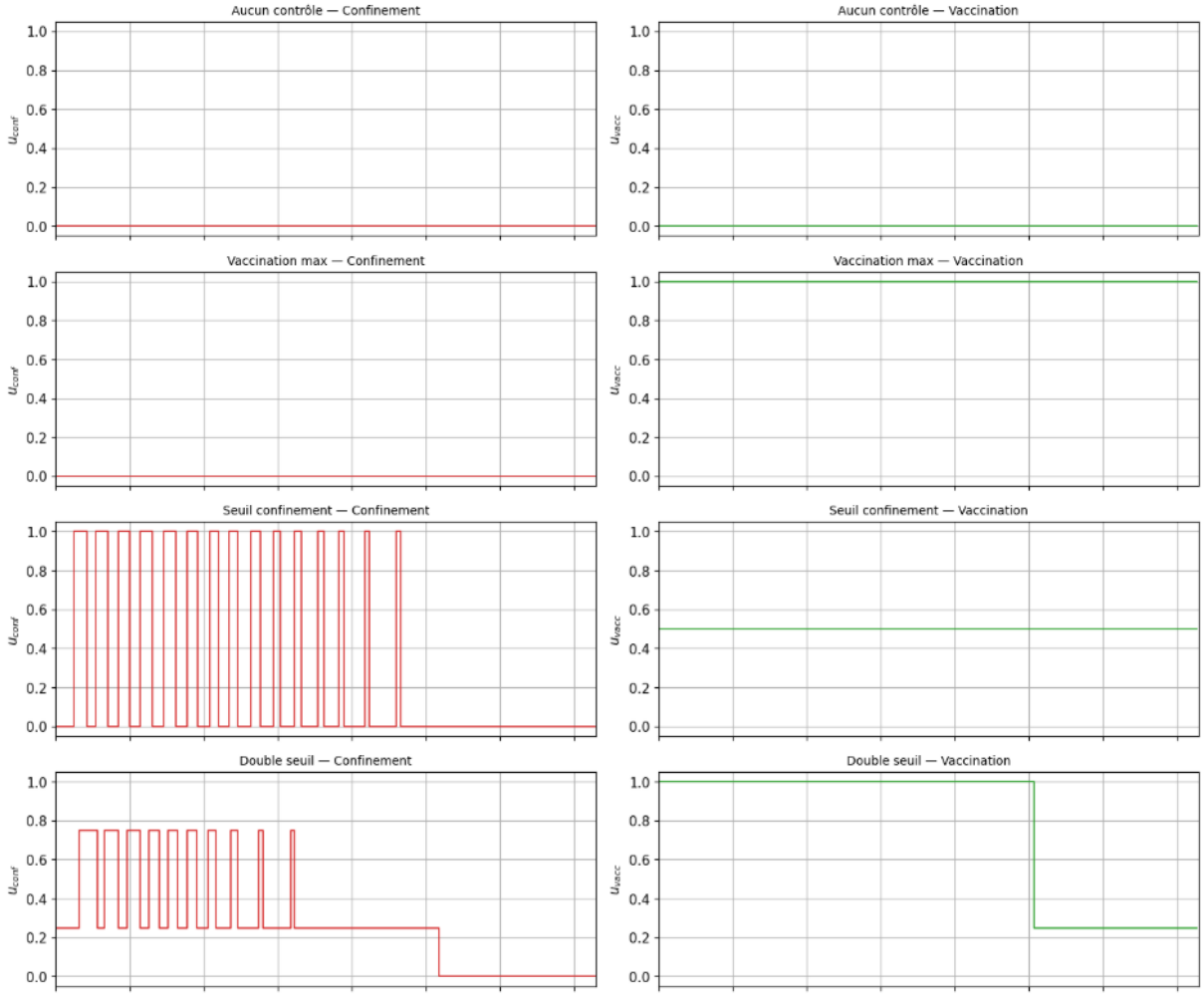

FIGURE 2 – Politiques des heuristiques

Nous observons aussi en figure 3 que le nombre d'infectés croit beaucoup plus rapidement dans les deux premières mesures tandis qu'il est bien plus faible sur les deux dernières, mais que la maladie met bien plus de temps à s'éliminer. Dans le même temps, la quantité de personnes susceptibles (S) est rallongée ; la priorité n'est pas à la vaccination de la population.

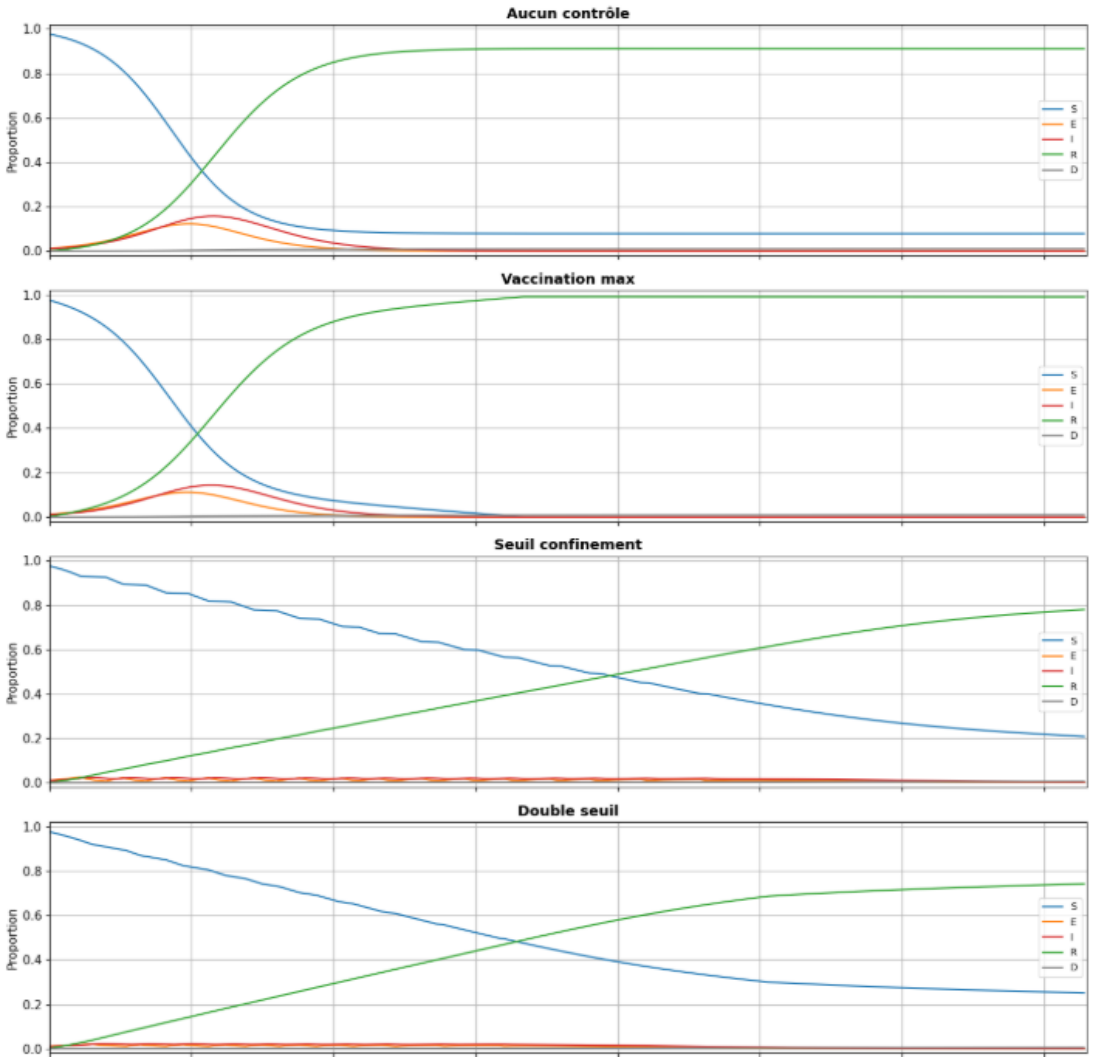

FIGURE 3 – Evolution des constantes SEIRD dans les heuristiques

4.2 Politiques des algorithmes

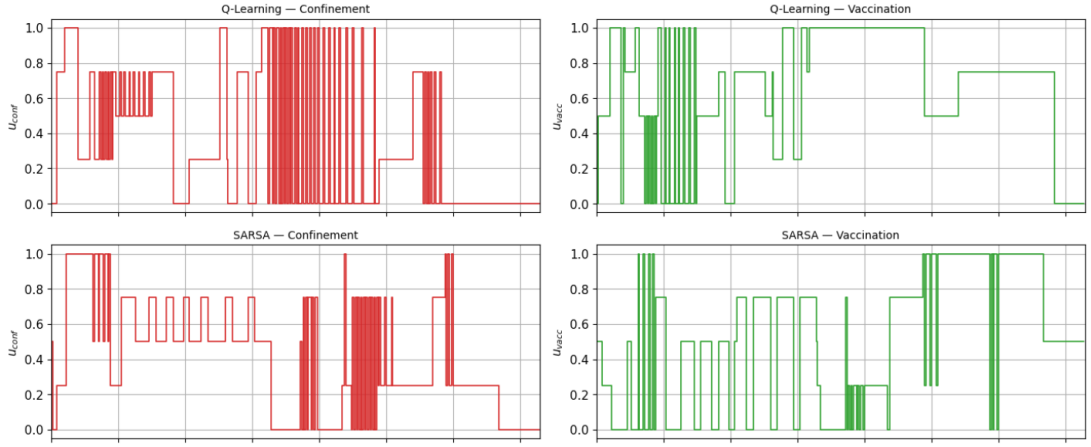

FIGURE 4 – Politiques obtenues par les algorithmes SARSA et Q-Learning

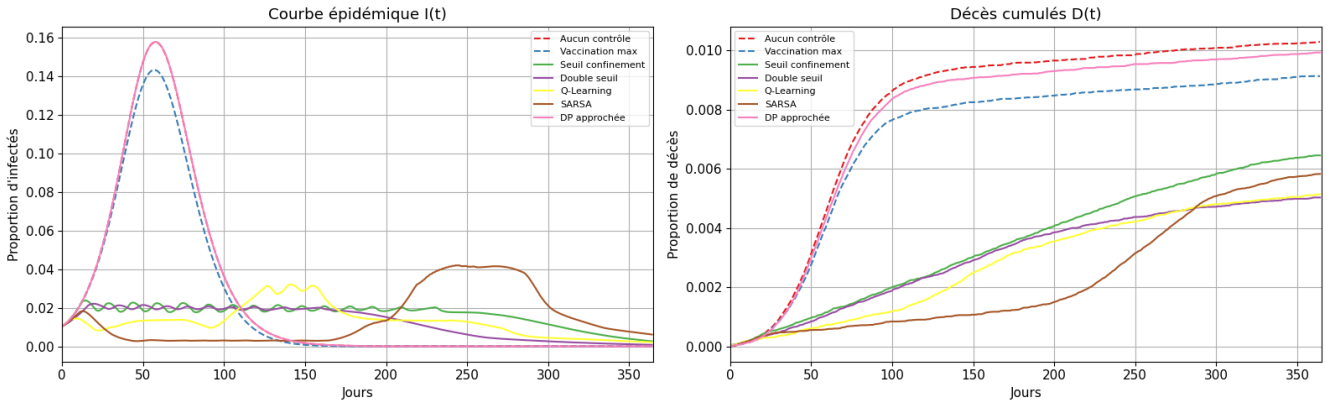

FIGURE 5 – Évolution de l'épidémie dans les différents scénarios

Les actions prises par nos algorithmes de Q-learning et SARSA sont beaucoup plus agressives. Les politiques correspondantes (Fig. 4) semblent totalement irréalistes : on alterne de manière erratique entre confinement total et inexistant, entre vaccination maximale et minimale. Cela est certainement dû à un manque de contrainte données aux algorithmes : aucune restriction n'empêche ces changements brusques. Nous pourrions dans une extension de notre modèle, introduire comme contrainte, une capacité à limiter les changements brutaux de taux de confinement et les capacités des entreprises et des citoyens à les accepter.

Au niveau du nombre de décès en figure 3 nous observons que SARSA et Q-Learning sont les plus prudents. D'autant plus en ce qui concerne SARSA qui propose de confiner intensivement au début (figure 4) puis en deuxième instance de vacciner intensément. D'abord, cette politique limite la propagation, tout en espérant l'augmentation de R. Différemment, Q-learning propose de maximiser et le confinement, (ce qui limite aussi les infectés contrairement aux heuristiques, et choisit aussi de beaucoup vacciner plus tardivement.

4.3 Evolution de l'épidémie

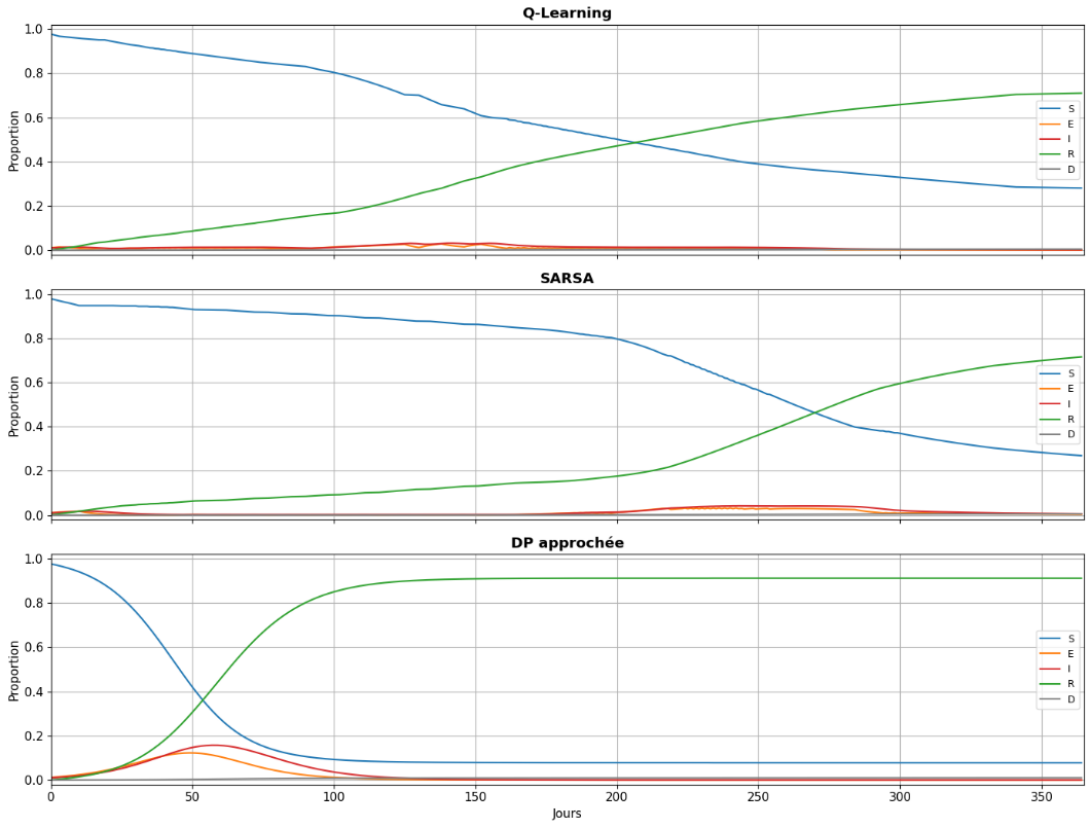

FIGURE 6 – Evolution des constantes SEIRD selon les algorithmes

Nous observons que les politiques choisies ont des impacts bien différents. Le non-contrôle a l'impact le plus désastreux que ce soit en terme d'infectés, ou de décès (la mithridatisation est une chimère). Dû au temps d'incubation et l'augmentation des infectés, vacciner seulement n'est pas non plus très efficace. Les résultats de la résolution par programmation dynamique approchée semblent aussi surprenants car ils sont proches du laisser faire. Cela pourrait venir d'une mauvaise implémentation ou d'un problème trop complexe.

Sur la figure 6, nous observons que le DP approché préfère résoudre l'économie bien plus tôt, quitte à provoquer un nombre de d'infectés bien supérieur. Nous notons notamment que la courbe des exposés semble avoir son pic avant celle des infectés ce qui concorde avec notre argument de période d'incubation. Toutefois comme présentés dans la figure 6 les pertes sont bien plus grandes.

Enfin, nous voyons que sur des critères *humains*, nos algorithmes étudiés sont les plus pertinents, ceci mettent beaucoup plus en avant les impacts de pertes de vies ou hospitalières, au détriment des impacts économiques.

5 Considération de coûts économiques

Notre considération principale était de prendre en compte l'impact économique de la pandémie. Sur ce critère l'heuristique double seuil est la plus efficace, elle limite et la perte de vie, l'hospitalisation et les pertes économiques. C'est la politique la plus efficace, qu'il serait sensé de proposer à un gouvernement.

Ensuite, différentes analyses sont possibles. Nous observons contrairement à *aucun contrôle*, ou *vaccination max*, qui favorisent les pertes humaines face à l'économie, que nos algorithmes

Q-Learning et *SARSA* proposent des solutions plus balancées entre pertes humaines ou coûts hospitaliers et pertes économiques du confinement. Ces deux solutions semblent tout à fait pertinentes en fonction de choix éthiques qui seraient pris par les décideurs politiques. De plus, le fait de ne pas prendre en compte l'engorgement des hôpitaux pourrait disqualifier instantanément en seconde instance des algorithmes présentant trop de coûts d'hospitalisation.

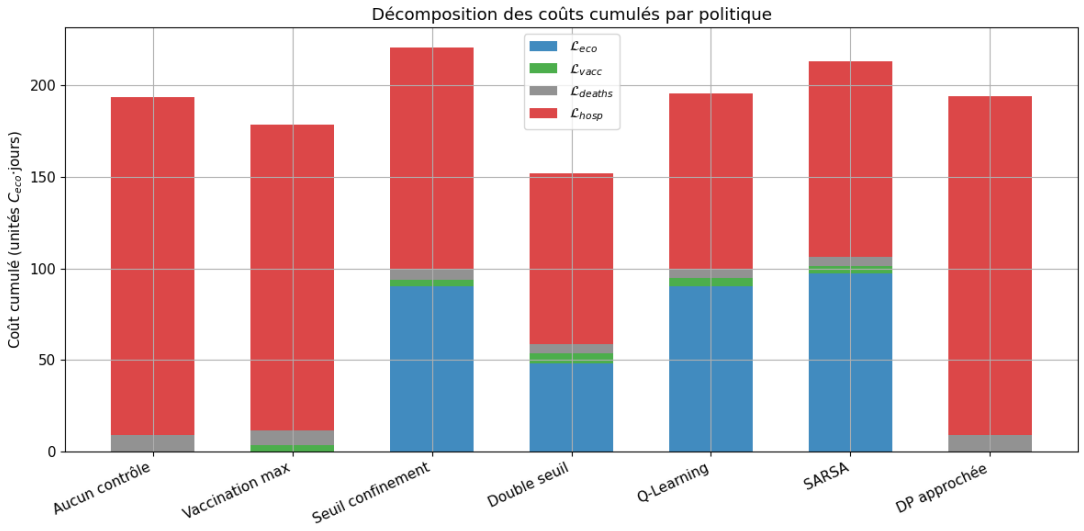

FIGURE 7 – Coûts financiers selon le scénario

5.1 Conclusion

Dans cette étude, nous avons comparé différentes manières de résoudre un problème SEIRD sur des données inspirées de nos expériences et des données historiques du Covid-19. Nous observons que si certaines solutions sont comparables, il pourrait revenir aux décideurs politiques de poser le curseur entre pertes humaines et pertes économiques. Toutefois, nous avons pu comparer nos algorithmes de RL à des cas d'usages réels et d'autres heuristiques *extrêmes* afin de proposer des politiques de vaccinations ou de confinement qui nous semblent pertinentes à l'aune de la simplicité de notre modèle.

Références

[1] Fernando E ALVAREZ, David ARGENTE et Francesco LIPPI. *A Simple Planning Problem for COVID-19 Lockdown*. Working Paper 26981. National Bureau of Economic Research, avr. 2020. DOI : 10.3386/w26981. URL : http://www.nber.org/papers/w26981.

[2] Jianhong WU, Kathy LEUNG et Gabriel LEUNG. « Nowcasting and Forecasting the Potential Domestic and International Spread of the 2019-nCoV Outbreak Originating in Wuhan, China : A Modeling Study ». In : *Obstetrical & Gynecological Survey* 75 (juill. 2020), p. 399-400. DOI : 10.1097/01.ogx.0000688032.41075.a8.

[3] Ofce OBSERVATOIRE FRANÇAIS DES CONJONCTURES ÉCONOMIQUES. « Évaluation au 30 mars 2020 de l'impact économique de la pandémie de COVID-19 et des mesures de confinement en France ». In : *OFCE Policy Brief* 65 (mars 2020), p. . URL : https://sciencespo.hal.science/hal-03610155.