# Utilisation de données cliniques pour la construction de modèles en oncologie

Thibaut Kritter

université de BORDEAUX

Inria INVENTEURS DU MONDE NUMÉRIQUE

# Thèse

pour l'obtention du grade de

**Docteur de l'université de Bordeaux**

Ecole doctorale de mathématiques et d'informatique

*Spécialité :*

Mathématiques appliquées et calcul scientifique

**Thibaut KRITTER**

---

## Utilisation de données cliniques pour la construction de modèles en oncologie

---

sous la direction de : **Olivier SAUT** et **Clair POIGNARD**

*soutenue le : 1er octobre 2018*

*Jury :*

| | | |
|---|---|---|
| **Hermine Biermé** | Professeur | Examinatrice (Présidente) |
| **Annabelle Collin** | Maître de conférences | Encadrante |
| **François Dufour** | Professeur | Examinateur |
| **Florence Hubert** | Professeur | Rapportrice |
| **Philippe Moireau** | Directeur de recherche | Rapporteur |
| **Elizabeth Moyal** | Professeur | Examinatrice |
| **Clair Poignard** | Directeur de Recherche | Directeur de thèse |
| **Olivier Saut** | Directeur de Recherche | Directeur de thèse |

# Remerciements

Je tiens tout d'abord à remercier mes deux directeurs de thèse, Olivier Saut et Clair Poignard. Merci à Olivier d'avoir guidé mes recherches et de m'avoir conseillé durant ces trois années. Ton calme et tes encouragements m'ont permis de travailler en toute sérénité. Merci également à Clair, avec qui les séances devant le tableau noir ont toujours été très enrichissantes. J'ai apprécié ta rigueur et ta détermination. Votre disponibilité ainsi que votre bonne humeur au quotidien ont rendu ces années très agréables pour moi, et vous êtes en grande partie responsables de la bonne ambiance qui règne dans l'équipe.

Un grand merci également à Annabelle Collin pour avoir co-encadré cette thèse. Ca a été un plaisir de travailler avec toi, ton investissement a été très motivant pour moi.

Merci à Florence Hubert et Philippe Moireau pour avoir accepté de rapporter mon manuscrit de thèse. Je remercie également Elizabeth Moyal, Hermine Biermé et François Dufour pour leur présence dans mon jury.

Je souhaite aussi remercier les autres permanents de l'équipe, Thierry Colin et Sébastien Benzekry, pour leurs remarques pertinentes sur mon travail ainsi que leur bonne humeur communicative.

Je tiens également à dire merci à Lorenzo Bello et Marco Rossi pour leur collaboration et pour m'avoir fourni les données sur lesquelles j'ai pu travailler.

Ces trois années n'auraient pas eu la même saveur sans toutes les personnes avec qui j'ai pu partager mon bureau. Un grand merci donc à Guillaume L., Boris, Vladimir (King Magnet), et Cristina pour tous ces moments passés ensemble. Une pensée particulière pour Guillaume D. pour ton aide et ton humour. Longue vie à Fencinnov !

Je souhaite également remercier les stagiaires, doctorants et post-doctorants des équipes MONC et Memphis pour la bonne ambiance qui règne dans ces couloirs et pour tous les moments passés en salle de pause et en dehors : Cynthia, Antoine, Florian, Olivier, Sergio, Costanza, Cédrick, Romain, Manon, Agathe, Jean, Claudia, Vivien, Marco, Federico, Andrea, Thomas, Etienne, Chiara, Guillaume, Sébastien, Erwann et Cécile.

Je tiens à adresser un grand merci à Nathan, Momo, Zaza et Cécilia pour toutes les bitoches que l'on a pu faire ensemble : j'ai apprecié nos soirées et votre compagnie. Je remercie aussi Louis-Marie et Luis pour les matchs d'extrême ping-pong.

Merci aux autres membres de l'IMB : en particulier merci à Jo de m'avoir accueilli à Bordeaux quand je suis arrivé, mais aussi à Nico, Elsa, Sami et Niko, toujours prêts pour aller boire un verre.

Merci aux équipes de foot de l'INRIA et du CNRS pour tous les midis passés sur le gazon.

J'en profite également pour remercier mes amis du Grand-Est : Alex, Frate, Delphine, Nico, Coco, Aurélien, Julie et Etienne : c'est toujours un grand plaisir de rentrer et de vous retrouver ! Merci également à Lolo et Jéjé qui m'ont montré la voie !

Un grand merci à mes parents Michel et Michèle qui m'ont toujours soutenu dans mes études et mes projets personnels. Merci également à ma soeur Céline et à Paco pour tous les bons moments passés à Metz et lors de vos venues à Bordeaux. Je remercie aussi Pascaline, dont les petites attentions me font toujours chaud au coeur. Je tiens à remercier toute ma famille, mes grands-parents, oncles, tantes et cousins. J'ai une pensée particulière pour Gilbert et Arlette, qui m'ont aidé à grandir et sans qui je ne serais pas là aujourd'hui.

Enfin merci à toi Cannelle pour tout ce que tu m'apportes au quotidien, pour toutes les belles aventures vécues et pour toutes celles à venir.

# Résumé

Cette thèse présente des travaux en lien avec l'utilisation de données cliniques dans la construction de modèles appliqués à l'oncologie. Les modèles actuels visant à intégrer plusieurs mécanismes biologiques liés à la croissance tumorale comportent trop de paramètres et ne sont pas calibrables sur des cas cliniques. A l'inverse, les modèles plus simples ne parviennent pas à prédire précisément l'évolution tumorale pour chaque patient. La multitude et la variété des données acquises par les médecins sont de nouvelles sources d'information qui peuvent permettre de rendre les estimations des modèles plus précises. A travers deux projets différents, nous avons intégré des données dans le processus de modélisation afin d'en tirer le maximum d'information. Dans la première partie, des données d'imagerie et de génétique de patients atteints de gliomes sont combinées à l'aide de méthodes d'apprentissage automatique. L'objectif est de différencier les patients qui rechutent rapidement au traitement de ceux qui ont une rechute plus lente. Les résultats montrent que la stratification obtenue est plus efficace que celles utilisées actuellement par les cliniciens. Cela permettrait donc d'adapter le traitement de manière plus spécifique pour chaque patient. Dans la seconde partie, l'utilisation des données est cette fois destinée à corriger un modèle simple de croissance tumorale. Même si ce modèle est efficace pour prédire le volume d'une tumeur, sa simplicité ne permet pas de rendre compte de l'évolution de forme. Or pouvoir anticiper la future forme d'une tumeur peut permettre au clinicien de mieux planifier une éventuelle chirurgie. Les techniques d'assimilation de données permettent d'adapter le modèle et de reconstruire l'environnement de la tumeur qui engendre ces changements de forme. La prédiction sur des cas de métastases cérébrales est alors plus précise.

Mots clés : Modélisation, Assimilation de données, Apprentissage automatique, Gliomes.

# Abstract

This thesis deals with the use of clinical data in the construction of models applied to oncology. Existing models which take into account many biological mechanisms of tumor growth have too many parameters and cannot be calibrated on clinical cases. On the contrary, too simple models are not able to precisely predict tumor evolution for each patient. The diversity of data acquired by clinicians is a source of information that can make model estimations more precise. Through two different projets, we integrated data in the modeling process in order to extract more information from it. In the first part, clinical imaging and biopsy data are combined with machine learning methods. Our aim is to distinguish fast recurrent patients from slow ones. Results show that the obtained stratification is more efficient than the stratification used by cliniciens. It could help physicians to adapt treatment in a patient-specific way. In the second part, data is used to correct a simple tumor growth model. Even though this model is efficient to predict the volume of a tumor, its simplicity prevents it from accounting for shape evolution. Yet, an estimation of the tumor shape enables clinician to better plan surgery. Data assimilation methods aim at adapting the model and rebuilding the tumor environment which is responsible for these shape changes. The prediction of the growth of brain metastases is then more accurate.

Key words : Modeling, Data assimilation, Machine learning , Glioma.

# Table des matières

**Remerciements** **3**

**Résumé** **5**

**Abstract** **7**

**1 Introduction** **11**

1.1 Contexte de la thèse . . . . . 12

1.2 Modélisation mathématique en oncologie . . . . . 13

1.3 Effets de traitements . . . . . 21

1.4 Données cliniques collectées . . . . . 23

1.5 Objectifs de la thèse . . . . . 27

1.6 Organisation de la thèse . . . . . 28

**Bibliographie** **35**

**2 Classification des patients ayant un gliome en fonction du temps de rechute par apprentissage** **41**

2.1 Introduction . . . . . 42

2.2 Méthodes d'apprentissage automatique appliquées à l'oncologie . . . . . 44

2.3 Données collectées par l'Humanitas Research Hospital et critère d'hétérogénéité . . . . . 56

2.4 Méthodes de pré-traitement . . . . . 63

2.5 Résultats de classification . . . . . 68
2.6 Discussion et conclusion . . . . . 79

**Bibliographie** **81**

**3 Assimilation de données pour la croissance de tumeurs** **85**
3.1 Introduction . . . . . 87
3.2 Modèle de croissance de tumeurs . . . . . 90
3.3 Estimation des paramètres par calibrage volumique . . . . . 98
3.4 Assimilation de données . . . . . 112
3.5 Résolution numérique . . . . . 133
3.6 Validation de la méthode sur données synthétiques . . . . . 139
3.7 Application aux données réelles . . . . . 162
3.8 Conclusion . . . . . 170

**Bibliographie** **173**

**Conclusion et Perspectives** **177**

**Annexes** **181**

# Chapitre 1

# Introduction

**Sommaire**


**1.1 Contexte de la thèse . . . . . 12**
**1.2 Modélisation mathématique en oncologie . . . . . 13**
1.2.1 Présentation du cancer . . . . . 14
1.2.2 Croissance tumorale . . . . . 17
**1.3 Effets de traitements . . . . . 21**
**1.4 Données cliniques collectées . . . . . 23**
1.4.1 Imagerie . . . . . 24
1.4.2 Biopsie . . . . . 26
**1.5 Objectifs de la thèse . . . . . 27**
**1.6 Organisation de la thèse . . . . . 28**
1.6.1 Partie I : Prédiction du temps de rechute de gliomes par apprentissage automatique . . . . . 28
1.6.2 Partie 2 : Assimilation de données pour la croissance tumorale . 30

## 1.1 Contexte de la thèse

Avec plus de 14 millions de nouveaux cas par an, le cancer est l'un des enjeux majeurs de santé publique. C'est la deuxième cause de mortalité dans les pays développés, après les maladies cardio-vasculaires : près d'un décès sur 6 dans le monde est dû au cancer [1]. En France métropolitaine, 400 000 nouveaux cas ont été recensés en 2017 et on compte 150 000 décès liés au cancer cette même année [2]. Chez l'homme, les cancers du poumon, du colon-rectum et de la prostate sont les plus courants, tandis que ce sont les cancers du sein et du poumon qui sont les plus fréquents chez la femme. L'apparition d'une tumeur est liée à la prolifération anormale de cellules dans un organe. Ces cellules acquièrent des mutations qui leur permettent d'échapper aux systèmes de régulation naturels [3]. Si la plupart des tumeurs sont bénignes et sont éliminées par le système immunitaire, certaines peuvent devenir malignes et invasives. Elles sont alors capables de se détacher de la tumeur primaire, d'entrer dans le système sanguin et d'atteindre de nouveaux organes distants, pour former des tumeurs secondaires, appelées métastases. Dans 90% des cas, ce sont ces métastases qui causent des infections et des défaillances d'organes, entraînant la mort du patient [4].

Les progrès en biologie permettent de mieux comprendre la maladie et le processus de formation de ces tumeurs. Des traitements ont alors été développés dans le but d'arrêter ou de ralentir la croissance tumorale, tels que la chirurgie, la radiothérapie, et la chimiothérapie. De plus, de nouvelles méthodes d'imagerie médicale permettent d'améliorer le suivi des patients et d'adapter le traitement plus spécifiquement qu'auparavant. Les effets de ces nouvelles technologies sont visibles, puisque par exemple, entre 1985 et 2005, le taux de survie à 5 ans est passé de 80 à 87% pour le cancer du sein, et de 70 à 94% pour le cancer de la prostate. Cependant, la marge d'amélioration est énorme puisque par exemple, les taux de survie à 10 ans des cancers du foie, du poumon et du pancréas sont très faibles (inférieurs à 10%) [5]. La maladie semble en effet parvenir à s'adapter et à résister aux traitements, notamment aux chimiothérapies.

La recherche mathématique appliquée à l'oncologie a pour but d'aider les cliniciens à comprendre la maladie et à anticiper son comportement. La modélisation des mécanismes biologiques liés au cancer permet de prévoir une évolution ou d'améliorer les thérapies actuelles. De plus, les données acquises par les médecins sont de plus en plus riches et précises, ce qui renforce l'utilisation de modèles pour coupler ces données. Un modèle trop simple ne permet pas de prédire efficacement les comportements atypiques. Ces nouvelles

données permettent alors de complexifier les modèles afin de les rendre plus précis.

Nous verrons dans cette introduction quelques modèles mathématiques habituellement utilisés en oncologie. Nous nous intéresserons ensuite aux données collectées par les médecins. L'idée de cette thèse est de tirer le maximum d'informations de ces données afin de corriger et calibrer les modèles. Nous détaillerons enfin plus précisément les deux projets de la thèse.

## 1.2 Modélisation mathématique en oncologie

Les mécanismes biologiques régissant la croissance tumorale sont complexes puisque de multiples phénomènes entrent en jeu. La modélisation mathématique peut permettre de mieux comprendre ces phénomènes. Le développement de nouvelles méthodes d'imagerie, d'analyses médicales, et d'expérimentations sont autant de moyens de mesurer une évolution et sont propices à l'utilisation de modèles. On distingue deux branches d'applications principales des modèles.

Au niveau biologique, ils permettent d'améliorer la compréhension des interactions entre les différentes entités. La comparaison entre la théorie et les expériences réalisées peut permettre de valider ou de rejeter une hypothèse du modèle. Dans les deux cas, l'information est intéressante puisqu'elle peut signifier soit que les mécanismes connus semblent corrects, soit qu'il en manque pour décrire l'évolution. La modélisation mathématique permet donc de formaliser une théorie biologique qui pourrait expliquer un phénomène et de la confronter aux données. [6, 7]

Au niveau clinique, la modélisation a pour but d'aider le clinicien à prendre des décisions thérapeutiques, comme la planification d'une chirurgie ou le choix d'un traitement. Les modèles peuvent en effet permettre de prédire l'évolution du volume ou de la forme tumorale, ou d'anticiper l'effet d'un traitement.

Dans cette thèse, nous nous intéresserons principalement à cette seconde application. Après avoir rappelé les mécanismes biologiques liés au cancer, nous nous pencherons sur les différents types de modèles qui existent, pour modéliser la croissance tumorale et les effets de traitements. La modélisation du processus métastatique n'est pas traitée dans cette thèse, mais elle fait l'objet d'une revue de Maini et al. [8].

### 1.2.1 Présentation du cancer

#### Les mécanismes biologiques

Afin de modéliser la croissance tumorale, il est nécessaire de comprendre les mécanismes biologiques à l'origine de cette croissance. La prolifération des cellules cancéreuses est due à des mutations génétiques qui leur permettent de se développer de manière incontrôlée [3]. A l'âge adulte, la division cellulaire ne s'effectue que de manière occasionnelle, par exemple pour remplacer des cellules mortes, ou pour cicatriser une blessure. Des points de contrôle lors du cycle cellulaire permettent de maîtriser cette prolifération, et de ne l'enclencher que lorsque cela est nécessaire. Ce contrôle est assuré par le phénomène d'apoptose, c'est-à-dire de mort cellulaire programmée. Un dérèglement de ces points de contrôles peut engendrer la formation d'une tumeur : les cellules affectées prolifèrent de manière incontrôlée. De plus, les mutations acquises se propagent. Les principales caractéristiques du cancer sont les suivantes :

— **Emission continue de signaux de prolifération**. Les cellules saines régulent leur prolifération, ainsi que celle des autres cellules, en envoyant des signaux de prolifération lorsque cela est nécessaire. Les cellules cancéreuses sont, elles, capables d'envoyer des signaux de prolifération en continu afin d'assurer une croissance permanente [3].

— **Echappement aux inhibiteurs de croissance**. Des protéines sont à l'origine des points de contrôle du cycle cellulaire et permettent d'inhiber la division cellulaire. Des mutations des gènes codant ces protéines permettent de désactiver ce rôle inhibiteur [9].

— **Résistance à la mort cellulaire**. L'apoptose est le phénomène de mort cellulaire programmée au cours de laquelle la cellule se suicide suite à la réception de signaux appelant à la destruction. Les cellules tumorales sont capables d'échapper à l'apoptose par envoi de signaux régulateurs. Elles ne meurent donc pas par apoptose mais par nécrose, c'est-à-dire par mort prématurée non programmée. Cela peut être dû à un manque d'oxygène ou à un traitement administré au patient. Cependant, cette nécrose peut entraîner une réparation du tissu environnant par les cellules inflammatoires, ce qui peut favoriser la prolifération.

— **Réplication à l'infini**. Une cellule saine ne peut effectuer qu'un nombre fini de divisions cellulaires, dû à un rétrécissement d'une extrémité non codante de son ADN. Les cellules cancéreuses sont, elles, capables de se diviser à l'infini grâce à une en-

zyme, la télomérase, qui permet de rallonger son ADN [10].

— **La reprogrammation du métabolisme énergétique**. Les cellules tumorales sont capables de s'adapter à leur environnement en modifiant leur besoin en oxygène et nutriments. En particulier, en cas de manque d'oxygène, les cellules cancéreuses compensent ce manque en métabolisant du glucose par glycolyse.

— **Angiogénèse**. La division cellulaire nécessite la consommation d'oxygène et de nutriments. Lorsque la vascularisation de l'environnement tumoral n'est plus suffisante pour alimenter toute la tumeur, les cellules tumorales sont capables d'envoyer des signaux induisant le phénomène d'angiogénèse. Ce phénomène consiste en la création de vaisseaux sanguins à partir des cellules endothéliales qui seront raccordées à la tumeur. Cette émission de signaux est une cible thérapeutique importante, puisque l'angiogénèse est essentielle pour que la tumeur se développe [11].

— **Invasion métastatique**. Suite à l'angiogénèse, la tumeur bénéficie de sa propre vascularisation. Certaines cellules sont alors capables de se détacher de la masse tumorale et d'entrer dans le système sanguin. La plupart du temps, ces cellules sont éliminées par le système immunitaire, mais il est possible qu'elles parviennent à coloniser un nouvel organe. Les organes les plus vascularisés sont donc les plus touchés par les métastases, et chaque tumeur principale colonise certains organes de manière préférentielle, en fonction de leur localisation et de l'affinité avec l'organe. Par exemple, le cancer de la prostate métastase généralement dans les os. De la même façon, le cancer du côlon tend à métastaser dans le foie.

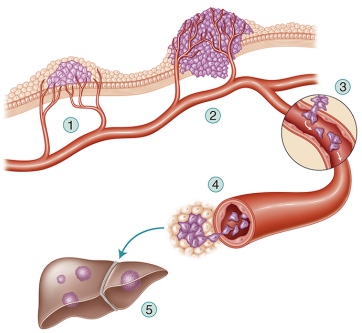



FIGURE 1.1 – Schéma du processus d'invasion métastatique [12].

A partir de ces caractéristiques du cancer, on distingue alors plusieurs étapes de développement de la maladie. Tout d'abord, la phase avasculaire correspond à la croissance

tumorale due à la consommation d'oxygène qui arrive par diffusion à partir des vaisseaux voisins. Lorsque la tumeur devient trop grosse, le centre de la tumeur n'a plus accès à ces nutriments : il se forme une zone de cellules au repos, dites quiescentes, ou une zone de nécrose [13]. La tumeur déclenche alors le phénomène d'angiogénèse qui lui permet d'être suffisamment vascularisée : c'est la phase vasculaire. La phase métastatique est ensuite le moment où des cellules cancéreuses commencent à atteindre d'autres organes.

#### Les traitements de la maladie

La connaissance de toutes ces phases permet alors de trouver des thérapies bloquant un, voire plusieurs de ces mécanismes de croissance. On distingue plusieurs types de traitements. Ils peuvent être combinés ou non, selon l'agressivité de la tumeur. Nous verrons dans la partie 1 qu'adapter le traitement de manière spécifique à chaque patient est un enjeu clinique important. On distingue les thérapies locales-régionales, qui s'attaquent à la tumeur localement et les thérapies systémiques qui cherchent à atteindre les cellules du corps tout entier par voie sanguine.

#### Thérapies locaux-régionales

— **Chirurgie.** La chirurgie consiste à extraire la tumeur lorsqu'elle est située dans des zones accessibles. Elle peut également servir pour la diagnostic grâce au prélèvement et à l'analyse d'un échantillon de tissu potentiellement cancéreux.

— **Radiofréquence et cryothérapie.** L'ablation par radiofréquence et la cryothérapie peuvent être utilisées dans le cas où la zone tumorale ne permet pas la chirurgie. Elles consistent respectivement en la destruction de la tumeur par chaleur ou par froid intense.

— **Radiothérapie.** La radiothérapie consiste en l'utilisation de radiations qui brisent les brins d'ADN et entraînent la mort cellulaire.

— **Electroporation.** L'électroporation est l'application d'un champ électrique sur une zone ciblée, à l'aide d'électrodes. Elle peut être utilisée soit à forte intensité pour détruire les cellules, soit à plus faible intensité pour perméabiliser la membrane et permettre à certaines molécules médicamenteuses d'accéder au coeur de la tumeur.

#### Thérapies systémiques

— **Chimiothérapie.** La chimiothérapie est l'administration de molécules qui s'attaquent aux cellules se divisant rapidement. Elles touchent donc les cellules tumorales, mais aussi d'autres cellules saines de l'organisme. Les effets secondaires sont donc importants.

— **Thérapies ciblées.** Les thérapies ciblées regroupent les méthodes où les molécules envoyées visent à inhiber un aspect de la croissance tumorale, comme le blocage dans le cycle de division cellulaire ou de l'angiogénèse.

La chirurgie est souvent combinée avec une des autres thérapies. Lorsque la thérapie complémentaire est administrée avant la chirurgie, on parle de traitement néo-adjuvant : il permet de réduire la taille de la tumeur afin de faciliter la chirurgie. Plus fréquemment, le traitement est utilisé après la chirurgie. Il est dit adjuvant et a pour objectif d'éviter la rechute.

### 1.2.2 Croissance tumorale

Les premiers modèles de croissance tumorale datent de 1932, quand Mayneord s'est aperçu, en confrontant son modèle à ses données, que la prolifération ne pouvait avoir lieu que sur la périphérie de la tumeur, avec une couche proliférante de plus en plus mince [14]. Cette revue des types de modèles utilisés pour la croissance tumorale n'a pas pour but d'être exhaustive, mais de montrer les différentes manières de représenter la tumeur. Des revues plus précises sont présentes dans la littérature [15, 16]. Dans un premier temps, nous décrirons des modèles EDO (équations différentielles ordinaires), qui régissent l'évolution du volume tumoral. Nous verrons ensuite les modèles spatiaux, formulés à partir d'EDP (équations aux dérivées partielles). Ils sont distingués en deux types, selon que l'on travaille au niveau cellulaire ou sur les densités de cellules. Notons que contrairement au cas de la physique (Schrödinger, Maxwell), il n'y a actuellement pas d'équation maître pour le cancer.

#### Modèles EDO

Les modèles les plus simples permettant d'estimer la croissance tumorale sont basés sur des équations différentielles ordinaires régissant le volume tumoral. La modélisation de la prolifération par un modèle exponentiel ne permet pas de retranscrire le comportement observé expérimentalement. En effet, en fin de phase avasculaire, la tumeur prolifère à un taux moins important dû au manque de nutriments. L'ajout de compétition entre les

cellules mène au modèle logistique, utilisé en dynamique des populations. Cependant, la modélisation de la prolifération et de la compétition seules ne permet pas d'obtenir une simulation de la croissance assez proche des données. Le modèle le plus couramment utilisé est le modèle Gompertz [17] :

$$\frac{dV}{dt} = aV \ln\left(\frac{K}{V}\right), \tag{1.1}$$

où $V$ est le volume tumoral, $a$ le taux de prolifération, et $K$ le volume limite que peut atteindre la tumeur. Cela permet de rendre compte de la quantité limitée de nutriments et d'oxygène dans le milieu. Ce modèle s'avère efficace pour prédire l'évolution volumique sur des données de rats [18]. Il l'est également dans le cas clinique, pour décrire la dynamique des tumeurs du sein [19].
Un autre modèle utilisé est le modèle de loi de puissance [20] :

$$\frac{dV}{dt} = aV^{\gamma}, \tag{1.2}$$

l'idée de ce modèle phénoménologique étant que seulement une proportion des cellules cancéreuses est proliférante. Par exemple, pour $\gamma = \frac{2}{3}$, cela revient à considérer que seules les cellules situées à la surface de la tumeur sont proliférantes. Ce modèle montre lui aussi de bons résultats de prédictions, chez l'animal [21] et chez l'homme [22]. Les modèles comportant plus de paramètres ne donnent pas nécessairement de meilleurs résultats, en particulier à cause du manque d'identifiabilité des paramètres.

Les modèles EDO énoncés précédemment ont l'avantage d'être simples et de modéliser l'aire ou le volume tumoral. Cependant, ils restent limités parce qu'il n'est pas possible d'y intégrer d'informations spatiales. Or il paraît clair sur les données d'imagerie clinique et expérimentale que la tumeur peut être très hétérogène. Les modèles spatiaux permettent de rendre compte de ces comportements. On distingue les modèles discrets et les modèles continus.

### Modèles discrets spatiaux

Dans les modèles discrets spatiaux, chaque cellule est suivie individuellement. Le premier type de modèle discret est le modèle sur grille. L'espace est discrétisé et chaque case de la grille peut être en interaction avec ses voisines. La facilité d'implémentation de ces méthodes est compensée par le fait que le mouvement des cellules n'est pas libre, mais fixé par la grille. Des exemples de tels modèles sont les modèles d'automates cellulaires [24, 25].

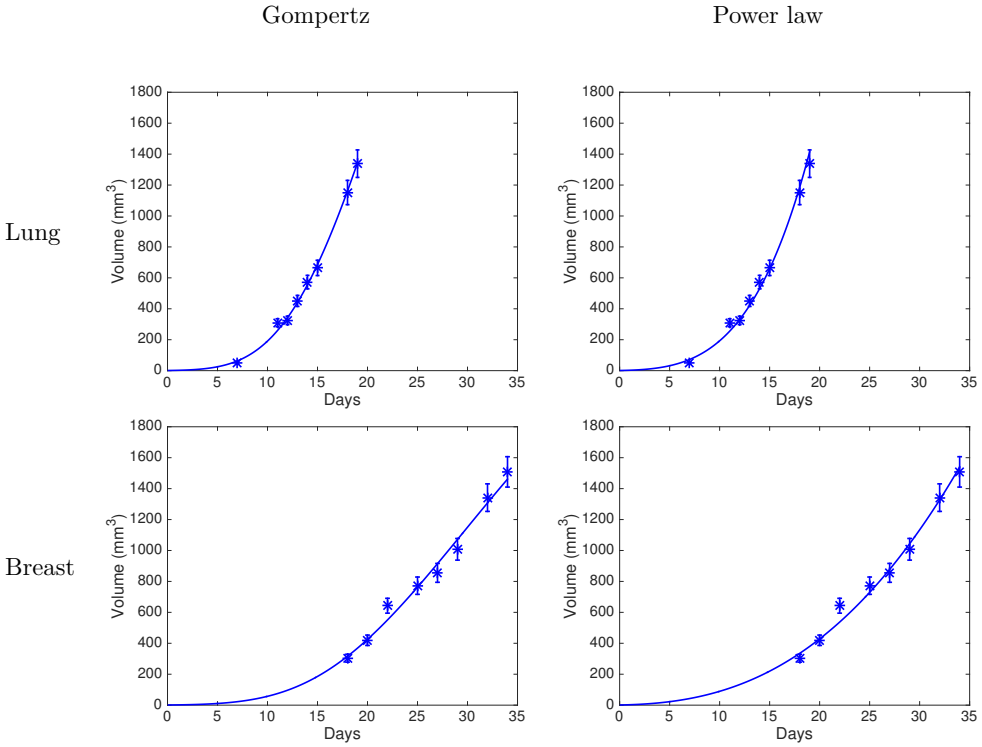



FIGURE 1.2 – Exemple de calibration volumique par les modèles de Gompertz (à gauche) et de loi puissance (à droite) sur des tumeurs au poumon et au sein [23].

Le second type de modèle est le modèle d'agent-centré. Les cellules sont alors représentées comme une entité à part entière, et peuvent avoir des changements de propriété ou de mouvement indépendamment les unes des autres. L'avantage de ce genre de modèles est de pouvoir prendre en compte la taille, la forme et la vitesse des cellules [26, 27].

Les modèles discrets sont très adaptés aux expériences biologiques où les cellules sont suivies individuellement, par exemple pour des cellules cancéreuses étudiées in vitro. Au niveau macroscopique, elles peuvent également être utilisées afin de modéliser un comportement qualitatif.

Cependant, ces modèles présentent deux inconvénients principaux si l'on souhaite les utiliser de manière quantitative au niveau macroscopique. Lorsque le nombre de cellules à décrire est trop important, les simulations ont un temps de calcul qui les rend inutilisables en pratique. De plus, les images cliniques ne sont pas assez précises pour distinguer le mouvement individuel des cellules. Il est donc difficile de calibrer les modèles à partir de l'imagerie. Le modèle ne peut que reproduire qualitativement un phénomène, comme c'est le cas dans l'exemple ci-dessus. Pour ces raisons, la modélisation à but clinique se fait par

des modèles spatiaux continus.

### Modèles continus spatiaux

Les modèles spatiaux continus sont basés sur des équations aux dérivées partielles, où l'on ne regarde plus le mouvement individuel des cellules, mais plutôt l'évolution de la densité de cellules. Ces modèles sont donc bien adaptés à l'échelle macroscopique et peuvent rendre compte de l'hétérogénéité tumorale. En effet, plusieurs populations de cellules peuvent être prises en considération. Là encore, on distingue deux types de modèles de croissance tumorale.

— **Equations de réaction-diffusion**
Le modèle de Swanson et al. [28] est un modèle de réaction-diffusion qui s'écrit :

$$\partial_t P - \nabla \cdot (D \nabla P) = \text{prolifération} - \text{mort}, \tag{1.3}$$

où $P$ correspond à la densité de cellules proliférantes, et $D$ est la diffusivité du milieu. L'hétérogénéité du milieu environnant est donc prise en considération dans ce terme $D$. Le terme de diffusion implique que les cellules ont un mouvement d'invasion. Ce modèle est donc particulièrement adapté aux tumeurs primitives qui sont très diffusives, comme par exemple les glioblastomes. Il a été utilisé pour confirmer les suppositions de forte invasion des gliomes dans le cerveau [29]. En particulier, il permet de montrer que la diffusion dans la matière blanche est plus forte que dans la matière grise.

— **Equations de transport**
Les modèles basés sur les équations de transports, développés par Preziosi et Ambrosi [30], s'écrivent sous la forme :

$$\partial_t P + \nabla \cdot (vP) = \text{prolifération} - \text{mort}, \tag{1.4}$$

où $P$ représente la densité de cellules proliférantes, $v$ la vitesse de déplacement des cellules. La loi de Darcy peut être choisie pour exprimer cette vitesse :

$$v = -\nabla \pi, \tag{1.5}$$

ce qui signifie que la vitesse des cellules et donc la croissance de la tumeur provient d'une trop forte pression $\pi$ entre les cellules proliférantes, comme le montre le schéma

de la figure 1.3. Ce choix de loi sur la vitesse est le plus simple, mais d'autres lois peuvent être considérées, comme une loi visco-élastique [31] ou dépendant de la tension de surface [32]. Ici, le mouvement des cellules n'est plus actif, contrairement aux modèles de réaction-diffusion, mais passif et uniquement lié à la prolifération. C'est ce type de modèle qui sera utilisé dans la partie 2 sur la modélisation de la croissance tumorale.

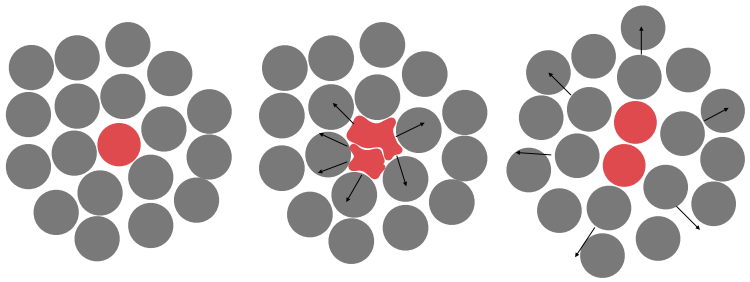


FIGURE 1.3 – Schéma de croissance tumorale : prolifération des cellules puis déplacement des cellules dû à la pression interne.

Dans les deux cas, il est possible de considérer plusieurs populations de cellules. Habituellement, on distingue les cellules proliférantes, quiescentes et invasives. Contrairement aux cellules proliférantes, les cellules quiescentes ne se divisent pas et sont dormantes, et les cellules invasives s'éloignent de la masse tumorale à la recherche de nouvelles zones de nutriments. Différencier les populations permet également de prendre en compte les différentes réactions possibles d'une cellule à un traitement.

## 1.3 Effets de traitements

La modélisation des effets de traitements sur le comportement tumoral a deux objectifs majeurs. Le premier est de prédire le volume tumoral à un temps futur, et également les différents phénotypes qui composent la tumeur. Le second est de parvenir à adapter le traitement de manière spécifique à chaque patient en déterminant la distribution de dose qui est optimale. Comme on l'a vu précédemment, plusieurs types de traitements sont possibles, et le choix de la thérapie dépend de l'agressivité et de la localisation de la tumeur, mais aussi de l'âge du patient. L'action de chaque thérapie étant différente, les modèles utilisés le sont aussi. Nous dressons ici une liste non exhaustive de modèles utilisés pour rendre compte de l'effet de différents traitements.

— **Chimiothérapie et radiothérapie.** Dans le cas de chimiothérapie et de radiothérapie, la cible visée par le traitement est directement la cellule cancéreuse. Les modèles comportent donc un terme de diminution de la densité de cellules proliférantes, dépendant de la dose administrée. On considère par exemple le modèle suivant :

$$\partial_t P + \nabla \cdot (vP) = \alpha P - \beta \exp(-\frac{t}{\tau})P, \tag{1.6}$$

où $P$ désigne la densité de cellules proliférantes. Cette équation est adaptée de l'équation de transport présentée à la section précédente, à laquelle un terme de thérapie $-\beta \exp(-\frac{t}{\tau})P$ est ajouté. Le paramètre $\beta$ correspond à la dose administrée, et $\tau$ rend compte de la durée d'efficacité du traitement. Il est également possible de différencier plusieurs phénotypes de cellules, qui ne réagissent pas de la même manière au traitement. Considérons le modèle [33] :

$$\begin{cases} \dfrac{dC}{dt} = -\alpha C, \\ \dfrac{dP}{dt} = \lambda_P P \left(1 - \dfrac{T}{K}\right) + k_{Q_P P} Q_P - k_{PQ} P - \gamma_P C P, \\ \dfrac{dQ}{dt} = k_{PQ} P - \gamma_Q C Q, \\ \dfrac{dQ_P}{dt} = \gamma_Q C Q - k_{Q_P P} Q_P - \gamma_{Q_P} Q_P, \\ T = P + Q + Q_P, \end{cases} \tag{1.7}$$

où $C$ désigne la concentration médicamenteuse, et $P, Q$ et $Q_P$ les densités de cellules proliférantes, quiescentes et quiescentes endommagées respectivement. Le traitement élimine directement les cellules proliférantes et altère les cellules quiescentes qui deviennent endommagées. Ces cellules endommagées peuvent réparer leur ADN afin de redevenir proliférantes, mais peuvent également mourir si la réparation n'est pas effectuée. Dans ces modèles, on ne considère qu'un seul cycle de traitement. Il est également possible de modéliser plusieurs cycles thérapeutiques. Un exemple de tels modèles est [34] :

$$\begin{cases} \partial_t C = D\Delta C + \rho(1 - C - C_d)C, \\ \partial_t C_d = D\Delta C_d - \dfrac{\rho}{k}(1 - C - C_d)C_d, \end{cases} \tag{1.8}$$

et pour chaque temps $t_i$ de traitement :

$$\begin{cases} C(x, t_i^+) = S(t_i) C(x, t_i^-), \\ C_d(x, t_i^+) = C_d(x, t_i^-) + [1 - S(t_i)] C(x, t_i^-), \end{cases} \tag{1.9}$$

où $C$ et $C_d$ représentent respectivement les densités de cellules cancéreuses et cancéreuses endommagées, et $S(t_i)$ correspond à la fraction de cellules capable de réparer les dommages causés par le traitement, au temps $t_i$. Tous ces modèles ont pour but d'aider à la planification et au dosage des traitements.

— **Thérapies ciblées.** Comme vu précédemment, les thérapies ciblées visent à bloquer un aspect du mécanisme de croissance. En particulier, des molécules peuvent agir pour bloquer l'angiogénèse. Le bevacizumab est un exemple de telle thérapie, et est utilisé pour traiter les cancers colorectaux, du poumon, du sein, du rein et du cerveau. Les cellules tumorales sont ainsi confrontées au phénomène d'hypoxie : il n'y a plus assez d'oxygène pour alimenter toute la tumeur. Nous ne considérons ici qu'une population de cellules cancéreuses. Une modélisation possible de cette thérapie est la suivante :

$$\begin{cases} \partial_t P + \nabla \cdot (vP) &= \alpha(M)P - dP, \\ \partial_t M &= -\eta M P, \end{cases} \tag{1.10}$$

avec :

$$\alpha(M) = \begin{cases} \alpha_0(M - M_{\text{hyp}}) & \text{si } M > M_{\text{hyp}}, \\ 0 & \text{sinon }, \end{cases} \tag{1.11}$$

où $P$ est la densité de cellules proliférantes, et $M$ est la quantité d'oxygène présent dans le milieu. La valeur $M_{\text{hyp}}$ correspond au seuil d'hypoxie : lorsque $M$ est en dessous de ce seuil, les cellules ne sont plus oxygénées et ne prolifèrent plus. Le terme $-dP$ modélise la mort cellulaire. L'équation régissant $M$ modélise la consommation d'oxygène due à la prolifération. Notons que contrairement aux modèles précédents, le traitement ne participe pas directement à la décroissance de la masse tumorale, mais ralentit la croissance.

— **Electroporation.** Les modèles d'électroporation régissent l'évolution du potentiel électrique dans les différents tissus de l'environnement tumoral. Ils se basent donc sur des lois physiques et peuvent être construits en raisonnant sur un circuit éléctrique équivalent au système considéré [35, 36].

## 1.4 Données cliniques collectées

Les progrès en imagerie médicale et en analyse génomique sont directement corrélés aux avancées dans la compréhension des mécanismes du cancer. En imagerie, l'amélioration de la qualité des images ainsi que la multiplication des modalités d'examens disponibles

permettent un meilleur dépistage de la maladie et un meilleur suivi du patient. En génomique, les moyens technologiques permettent de séquencer les génomes bien plus rapidement qu'auparavant. Il est possible aujourd'hui d'identifier un gène altéré ou une protéine à l'activité anormale, menant à la cancérisation d'une cellule saine. Ces données peuvent permettre d'identifier de nouvelles causes du cancer et de comprendre pourquoi les patients réagissent différemment à un traitement. Nous allons ici traiter tout d'abord des méthodes d'imagerie utilisées fréquemment en clinique, puis nous nous intéresserons aux données extraites des biopsies. Dans les deux cas, les techniques abordées sont celles qui peuvent être utilisées en modélisation mathématique.

### 1.4.1 Imagerie

L'imagerie médicale consiste en l'acquisition de signaux, tels que des champs magnétiques ou des rayonnements, qu'il est possible d'interpréter afin de visualiser les différentes structures de l'organe visé. On distingue deux grandes familles d'imageries : l'imagerie structurelle et l'imagerie fonctionnelle.

L'imagerie structurelle a pour but de visualiser les données anatomiques. Elle permet d'acquérir les informations de taille, de forme, et les différences de tissus d'un organe. Les méthodes d'imagerie structurelle les plus fréquemment utilisées sont les suivantes :

— **Echographie.** L'échographie permet une cartographie en 2D des tissus en évaluant la propagation d'ultrasons dans ces tissus. En oncologie, elle est souvent accompagnée d'un produit de contraste et permet de distinguer une tumeur bénigne d'une maligne.

— **Radiographie.** La radiographie utilise des rayons X pour visualiser les os (en blanc) et les autres organes du corps humains (en gris), en 2D. Elle est utilisée pour la détection de cancer du poumon.

— **Scanner.** Le scanner utilise également des rayons X, mais cette fois il est possible de reconstituer le signal en 3D grâce à plusieurs coupes 2D réalisées. Il est alors possible de distinguer les tissus qui n'ont pas la même densité.

— **Imagerie à résonnance magnétique (IRM).** L'IRM repose sur le principe de la résonnance magnétique. Un champ magnétique puissant, créé par un aimant supraconducteur, permet une magnétisation des tissus : les atomes d'hydrogène s'orientent

alors dans la même direction. Ils sont ensuite mis en résonnance par application d'ondes radio. Le signal émis lors de la relaxation des atomes est récupéré et interprété afin de reconstituer une image 3D. Le temps entre deux excitations ainsi que le temps entre l'excitation et la mesure du signal reçu peuvent être modifiés afin d'obtenir différentes modalités d'examens (T1, T2, T1c, Flair). La qualité des images permet leur utilisation dans un but de modélisation, comme pour les scanners, ce qui n'est pas le cas de l'échographie et de la radiographie. L'IRM de diffusion est une variante qui permet de distinguer, pour chaque voxel, si le milieu est isotrope ou s'il y a des directions préférentielles des fibres. Cette donnée peut donc permettre d'aider à la modélisation de la croissance tumorale.

L'imagerie fonctionnelle a, elle, pour but de visualiser le fonctionnement métabolique des tissus, afin de réveler leur activité physiologique. Deux types d'imagerie fonctionnelle sont principalement utilisés :

— **Topographie par Emission de Photons (TEP scan).** Le TEP scan permet de visualiser en 3D l'activité métabolique d'un organe. Un traceur, marqué par un atome radioactif (carbone, fluor, azote,...), est injecté dans l'organe cible. Le traceur se dépose alors dans l'organe, dans les zones qui le consomment le plus. Il émet ensuite des positons dont l'annihilation par le milieu produit deux photons, qui s'émettent dans la même direction en sens opposé. La détection de ces photons par les capteurs de la machine permet de localiser les zones où le traceur était le plus concentré. Habituellement, le traceur choisi est le glucose radioactif, puisqu'il va se fixer sur les tissus qui consomment une grande quantité de sucres, comme les tissus cancéreux. Le signal PET reconstitué est plus pertinent que le signal IRM pour détecter le tissu malin. Le traceur choisi peut également être la méthionine. On parle alors de MET PET scan. Le MET PET scan reflète les besoins cellulaires en précurseurs de la synthèse des protéines. Elle est corrélée à la prolifération cellulaire, nettement plus élevée dans la tumeur que dans le tissu sain. L'intéret principal de cette méthode est de permettre de différencier les cellules cancéreuses proliférantes, donc actives, des cellules cancéreuses quiescentes.

— **IRM fonctionnelle.** L'IRM fonctionnelle est surtout utilisée en imagerie cérébrale. Elle permet de visualiser l'activité cérébrale, en enregistrant les variations locales des propriétés du flux sanguin lorsque ces zones sont stimulées.

— **Spectroscopie.** La spectroscopie par résonnance magnétique fonctionne sur le même

appareil que l'IRM. Mais au lieu de recueillir le signal des molécules d'eau, le signal de l'eau est supprimé et c'est le signal des autres molécules dissoutes qui est analysé. Le principe est d'utiliser les différentes fréquences de résonnance des protons des différentes molécules. Cela permet alors de connaître la composition des tissus [37]. Par exemple, cela peut permettre de distinguer une cellule cancéreuse souche d'une cellulle différenciée. L'utilisation de cet examen en modélisation est compliquée puisque la composition des tissus n'est pas connue pour chaque voxel.

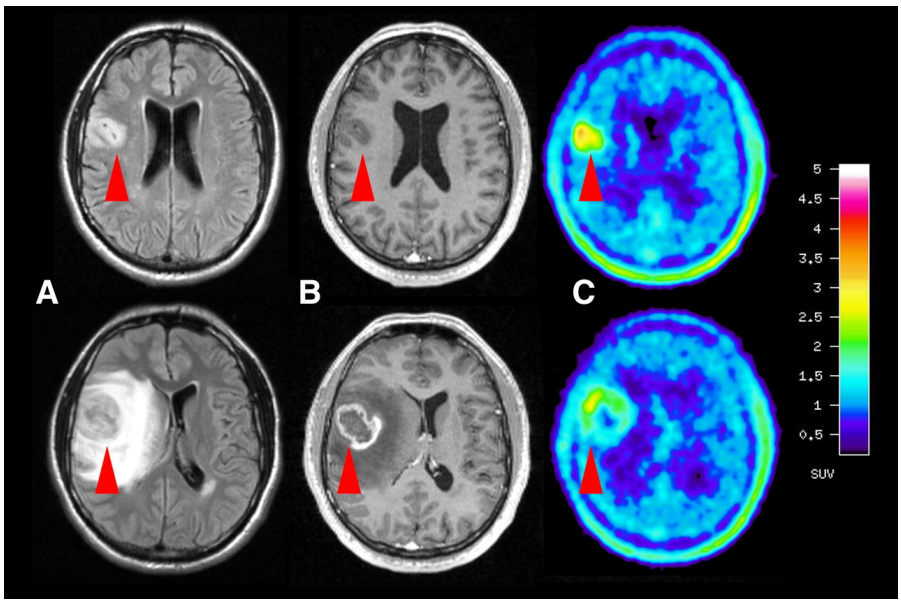



FIGURE 1.4 – Examens d'un glioblastome (gliome de grade IV) au diagnostic (première ligne) et deux mois plus tard (deuxième ligne) : (A) IRM pondérée T2, (B) IRM pondérée T1 avec agent de contraste, et (C) PET scan [38].

La combinaison d'imagerie structurelle et fonctionelle, avec différentes modalités d'examens, permet de recouper les informations afin d'estimer au mieux les caractéristiques tumorales, comme sur la figure 1.4.

### 1.4.2 Biopsie

La biopsie est une intervention lors de laquelle un échantillon de tissu ou de tumeur est prélevé afin d'être analysé. Ce prélèvement peut se faire de différentes manières, selon le tissu concerné : par prélèvement chirurgical (cerveau), par ponction (sein, prostate), par endoscopie (système digestif, vessie) ou par frottis (col de l'utérus). La biopsie sert à diagnostiquer un cancer, à déterminer à quel point le cancer s'est propagé, ou a déterminer le

grade d'un cancer. Le grade de la tumeur donne une indication sur son agressivité et est souvent utilisé pour choisir le traitement adapté. De plus, dans le cas où le tissu extrait est cancéreux, il permet d'analyser l'expression de certains gènes dans la tumeur. La recherche actuelle permet ainsi de détecter de nouveaux gènes corrélés au devenir du patient.

Dans le cas des gliomes, les tumeurs au cerveau les plus répandues, le gène IDH1 est utilisé depuis 2016 pour distinguer les patients ayant une bonne chance de répondre positivement au traitement, des autres patients. Il a même remplacé le grade dans la classification des gliomes [39]. L'analyse génomique permet également de quantifier l'activité de la tumeur. Par exemple, l'anticorps monoclonal MIB-1 permet de mesurer un index de prolifération au moment de la biopsie [40]. Enfin, dans le cas de neuroblastome, l'existence de délétions chromosomiques 1p 11q est très souvent associée à un mauvais pronostic clinique [41]. Tous ces résultats ont un impact direct sur la prise en charge de ces pathologies en permettant de mettre en place une thérapie spécifique en fonction de leur profil génomique. Dans le projet de la partie 1, de telles données seront utilisées afin d'améliorer la prédiction du temps de survie sans progression de patients atteints de gliomes de bas grade.

## 1.5 Objectifs de la thèse

La modélisation mathématique des mécanismes biologiques liés au cancer a pour but d'apporter de l'information, au niveau biologique et clinique. Les modèles permettent en effet de représenter formellement une théorie biologique afin de comparer le comportement théorique au comportement expérimental. Cela vise à une meilleure compréhension de la maladie. Nous allons ici plutôt nous intéresser à l'apport des modèles en clinique. La recherche médicale avance, les connaissances sur le cancer progressent et de nouveaux traitements voient le jour. Cependant, toutes ces avancées sont autant de questions auxquelles sont confrontés les cliniciens : Quelle est la meilleure manière d'administrer le traitement en question ? Quelle évolution de la maladie peut-on prévoir ? Comment combiner les traitements et la chirurgie ? Quand planifier la chirurgie ? Souvent, seule l'expérience du clinicien acquise sur de précédents cas permet d'obtenir un début de réponse.

Les modèles mis en place peuvent alors entrer en jeu et aider le clinicien dans les choix à effectuer. La prédiction par un modèle de l'évolution d'une tumeur, en croissance seule ou sous l'effet de traitement est une information qui peut conforter le clinicien dans sa prise de décision, ou à l'inverse l'alerter sur un comportement atypique. Afin d'être appliqués directement en clinique, ces modèles se doivent d'être précis et testés pour montrer

leur efficacité. Or, parvenir à construire un modèle pertinent biologiquement et calibrable mathématiquement est tout l'enjeu de la recherche en modélisation. Ils dépendent évidemment de la question posée par le clinicien, mais aussi des données disponibles qui permettent d'enrichir ou de valider le modèle. Habituellement, un modèle est construit en formalisant les principes biologiques que l'on souhaite intégrer. Les données collectées servent alors à confronter le modèle aux observations, et à valider ou non le modèle.

Dans cette thèse, l'idée est d'intégrer au maximum les données dans la calibration des modèles, afin d'estimer une évolution de la maladie de manière spécifique pour chaque patient. La principale difficulté pour établir un modèle est de respecter à la fois une structure assez simple pour permettre de calibrer le modèle, mais aussi assez complexe pour pouvoir simuler des mécanismes eux même complexes. Ici, nous allons voir deux moyens d'améliorer la prédiction du comportement tumoral, en utilisant au mieux l'information issue des données. Dans le cas où beaucoup d'informations sont collectées, mais à un seul instant, il est impossible d'utiliser un modèle mécanistique puisqu'il n'y a pas de données longitudinales. Nous verrons donc que les techniques d'apprentissage automatique permettent de tirer le maximum d'informations de ces données, afin de prédire l'évolution de la maladie. Dans le second cas où des données sont collectées durant plusieurs examens séparés dans le temps, il est possible de modéliser la croissance tumorale. Comme l'environnement de la tumeur n'est pas pris en compte car il n'est pas exploitable à partir de l'imagerie, nous verrons comment le maximum d'informations peut être tiré des données afin de corriger le modèle et d'améliorer la prédiction. Ces deux projets sont résumés dans la section suivante.

## 1.6 Organisation de la thèse

### 1.6.1 Partie I : Prédiction du temps de rechute de gliomes par apprentissage automatique

Les gliomes représentent 80% des tumeurs primaires du cerveau. Afin de les distinguer selon leur agressitivité, les grades I à IV ont été définis par l'Organisation Mondiale de la Santé en 2007 [42]. Depuis 2016, la mutation du gène IDH1 a remplacé le grade dans la classification des gliomes : lorsque le gène est muté, le pronostic clinique du patient est plus mauvais que lorsque le gène est sauvage [39]. Dans la plupart des cas, les patients atteints de gliomes subissent une chirurgie. Cela peut être suivi d'un traitement adjuvant, de type radiothérapie ou chimiothérapie, dû au caractère invasif de ces tumeurs. Le choix du traitement est fixé par l'Organisation Mondiale de la Santé, selon les caractéristiques

de la maladie et du patient (age, grade, IDH1, etc...). Malheureusement, on observe dans la plupart des cas une rechute de la maladie. Le temps entre la chirurgie et la rechute s'appelle le PFS (progression-free survival). Il varie de quelques mois à plusieurs années, même pour des patients traités de la même manière. **La question posée par les cliniciens est de savoir s'il est possible de distinguer, avant même que le traitement soit administré, quels patients vont probablement rechuter rapidement et sur quels patients le traitement sera efficace**. Cette information permettrait au médecin d'adapter le traitement de manière spécifique pour chaque patient. Si le traitement n'est pas efficace sur un patient, il peut être préférable de tenter une autre thérapie. A l'inverse, lorsque le traitement semble efficace, adapter la dose peut permettre d'améliorer le temps de survie.

Cette stratification des patients se fait actuellement à partir du grade et du statut IDH1 de la tumeur. L'idée ici est d'incorporer d'autres données afin d'améliorer cette classification. Nous avons travaillé sur une base de données issue de Humanitas Research Hospital à Milan. Cette base contient les données de 90 patients souffrant de gliomes de grade II et III. Les données collectées sont principalement issues de l'imagerie et de la biopsie. En imagerie, le volume tumoral et la position de la tumeur sont déterminés à partir de l'IRM. De plus, plusieurs caractéristiques de l'activité métabolique des tumeurs sont collectées à partir du MET-PET scan. Cela comprend l'intensité maximale du signal PET, mais aussi le volume métabolique et d'autres informations du signal. **Afin de rendre compte de l'hétérogénéité spatiale de la tumeur, nous avons également développé un indicateur d'hétérogénéité**. L'idée principale est de compter le nombre de composantes connexes du signal PET. Un nombre élevé de spots de cellules actives peut en effet être un bon indicateur de l'agressivité de la tumeur. De plus, la rechute semble avoir plus de chance de se produire si la tumeur posséde avant la chirurgie plusieurs sites actifs. Puisque le nombre de composantes connexes du signal dépend fortement du seuil choisi pour segmenter le signal PET, nous avons évalué l'évolution du nombre de composantes connexes lorsque le seuil parcourt toute la gamme d'intensités. L'intégrale de la courbe obtenue est l'indicateur que nous avons retenu.

Afin d'appliquer des méthodes d'apprentissage automatique à ce tableau de données de taille 90x20, il est nécessaire de réduire l'espace des caractéristiques. Dans un premier temps, cet espace est réduit en ne sélectionnant que les caractéristiques les plus utilisées lors d'une classification par forêts aléatoires. **Cela permet de constater que les données issues du MET PET sont les plus influentes**. En particulier, l'indicateur d'hé-

térogéneité précédemment défini figure parmi les caractéristiques les plus significatives. Une seconde réduction est effectuée en utilisant le principe de l'analyse en composante principale. L'algorithme est adapté ici puisque les données sont à la fois qualitatives et quantitatives. La base de donnée est alors projetée sur les deux directions principales, obtenues par combinaison linéaire des caractéristiques significatives.

Plusieurs méthodes d'apprentissage automatique sont appliquées à la base obtenue. La méthode de validation croisée à 10 ensembles est utilisée pour compenser la petite taille de la base de donnée. **Cela permet de montrer que dans un peu moins de 80% des cas, la classification des patients obtenue est correcte. La comparaison de la stratification de notre méthode avec les stratifications obtenues par le grade et le statut IDH1 montre que l'on améliore largement la prédiction.** En particulier, on remarque que les patients mieux stratifiés par notre méthode sont les patients de grade II IDH1 sauvage et les patients de grade III IDH1 muté.

Cette étude est une preuve de concept qui donne des résultats prometteurs. Ils doivent évidemment être validés sur une plus grande base de donnée, mais la combinaison des données d'imagerie et des données génomiques semble être une source d'informations intéressante.

### 1.6.2 Partie 2 : Assimilation de données pour la croissance tumorale

Lorsque des données longitudinales sont disponibles, il est cette fois possible d'établir un modèle qui régit l'évolution de la densité de cellules cancéreuses. Parvenir à prédire l'évolution d'une tumeur est d'une importance capitale pour les cliniciens. En effet, si l'on sait prédire le volume et la forme d'une tumeur à un temps futur, il est alors possible de planifier plus précisement une chirurgie, ou d'adapter le traitement de manière spécifique au patient. **L'enjeu de cette section est de construire un modèle qui reproduit au mieux la croissance d'une tumeur, à partir des premiers examens du patient.** Pour cela, le modèle suivant, inspiré d'autres travaux [30, 43] est considéré :

$$\begin{cases} \partial_t P + \nabla \cdot (\mathbf{v}\, P) &= MP, \quad \mathcal{B}, \\ \partial_t S + \nabla \cdot (\mathbf{v}\, S) &= 0, \qquad \mathcal{B}, \end{cases} \tag{1.12}$$

où $P$ représente la densité de cellules proliférantes, et $S$ la densité de cellules saines. On considère également l'équation sur la vascularisation $M$ :

$$\partial_t M = -\alpha MP, \qquad \mathcal{B}, \tag{1.13}$$

avec les conditions initiales suivantes :

$$\begin{cases} P(0,x) &= P_0(x), \\ S(0,x) &= 1 - P_0(x), \\ M(0,x) &= M_0(x). \end{cases} \tag{1.14}$$

On suppose également que le milieu est saturé, c'est-à-dire que $P+S=1$, et que la vitesse $\mathbf{v}$ suit une loi de Darcy, à savoir :

$$\mathbf{v} = -\nabla\pi. \tag{1.15}$$

Cela signifie que la vitesse dérive d'une pression $\pi$ qui représente la compétition spatiale entre les cellules. On obtient alors l'équation suivante qui permet de fermer le système et de calculer $\mathbf{v}$ :

$$\nabla \cdot \mathbf{v} = -\Delta\pi = MP, \tag{1.16}$$

avec des conditions de Dirichlet homogènes pour la pression $\pi$, modélisant le fait que l'influence de la tumeur sur le tissu sain est localisée. Ce modèle présente donc une prolifération des cellules proliférantes $P$ ainsi qu'une consommation de l'oxygène et des nutriments $M$ due à cette prolifération. Il a l'avantage de ne contenir que deux paramètres $\alpha$ et $M_0$, puisque $P_0$ est fixé à partir du premier examen du patient. De plus, le volume tumoral $V(t)$ est calculable explicitement dans le cas où $M_0$ est constant spatialement et égal à $m_0$ :

$$V(t) = V_0 \left(1 + \frac{m_0}{m_0 - \alpha} \left(\exp(m_0 - \alpha)t - 1\right)\right). \tag{1.17}$$

Cela permet donc de calibrer le modèle et de trouver $(\alpha, m_0)$ à partir des mesures de volume des premiers examens. La calibration volumique est testée sur des cas synthétiques et réels et permet d'estimer le volume tumoral avec une erreur relative de moins de 12% sur les exemples considérés [44]. La simulation 3D lancée avec les paramètres estimés permet alors de comparer la forme tumorale simulée à la forme tumorale observée. **Lorsque la tumeur croît de manière plutôt isotrope, l'estimation est bonne. Par contre, lorsque la forme tumorale évolue, le modèle n'est pas capable de l'estimer.** Cela s'explique par le fait que la calibration volumique considère $M_0$ constant spatialement. Or en pratique, le milieu environnant influe grandement sur la forme tumorale. Les vaisseaux et fibres présents autour de la tumeur peuvent entraîner une croissance plus forte dans les directions concernées. L'information des examens qui est utilisée jusqu'ici se limite aux volumes tumoraux. L'idée de cette section est de corriger le modèle en utilisant les formes tumorales aux temps intermédiaires des examens.

L'assimilation de données permet cette correction de la dynamique du modèle en utilisant

les observations. Les simplifications biologiques dues à la modélisation ainsi que les erreurs possibles de segmentation des observations seront compensées par une correction de l'état, c'est-à-dire de $P$, ainsi qu'une correction des paramètres du modèle. En particulier, on autorise le fait que $M_0$ puisse prendre des valeurs différentes dans plusieurs zones. Ces deux types de corrections nécessitent que l'on puisse mesurer la différence entre le front tumoral simulé et le front observé. Notons $B$ le domaine d'étude, et $\Omega_P^{\text{in}} = \{P > 0.5\}$. Notons également $z(t, x)$ les observations. La mesure de similarité choisie est issue de la fonctionnelle de Chan Vese [45] :

$$D(z, P) = \int_{\Omega_P^{\text{in}}} (z - C_{\max}(z, P))^2 dx + \int_{\mathcal{B}\setminus\overline{\Omega_P^{\text{in}}}} (z - C_{\min}(z, P))^2 dx, \tag{1.18}$$

où $C_{\max} = \max(C_1, C_2)$, $C_{\min} = \min(C_1, C_2)$ avec :

$$C_1(z, P) = \frac{\displaystyle\int_{\Omega_P^{\text{in}}} z \, d\mathcal{B}}{\displaystyle\int_{\Omega_P^{\text{in}}} d\mathcal{B}}, \quad C_2(z, P) = \frac{\displaystyle\int_{\mathcal{B}\setminus\overline{\Omega_P^{\text{in}}}} z \, d\mathcal{B}}{\displaystyle\int_{\mathcal{B}\setminus\overline{\Omega_P^{\text{in}}}} d\mathcal{B}}. \tag{1.19}$$

Le premier terme de la fonctionnelle mesure à quel point le front simulé est en avance par rapport au front observé, alors que le second terme mesure le retard par rapport aux observations. Une fois cette mesure établie, on définit les termes de correction d'état et de paramètres.

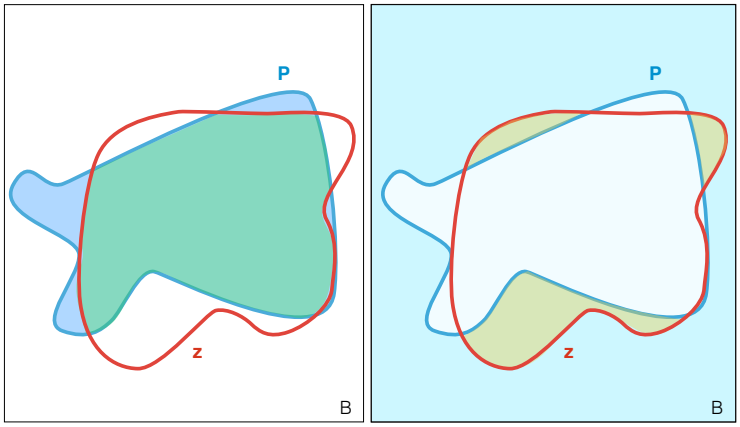



FIGURE 1.5 – Schéma pour le calcul des termes $C_1$ et $C_2$ dans la fonctionnelle de Chan Vese [45]. $C_1$ est égal à la fraction entre l'aire de la zone verte et celle de la zone bleue. $C_2$ est la fraction entre l'aire de la zone jaune et celle de la zone bleu clair.

La correction d'état se fait par filtre de Luenberger, inspiré par [46]. Notons $\Gamma = \{P = 0.5\}$ le contour tumoral et notons avec un chapeau les variables du système corrigé. On considère donc le modèle suivant :

$$\left\{\begin{array}{rcll} \partial_t \hat{P} + \mathrm{div}(\hat{v}\,\hat{P}) & = & \hat{M}\hat{P} + \lambda\,\delta_{\hat{\Gamma}}\left(-\big(z - C_{\max}(z,P)\big)^2 + \big(z - C_{\min}(z,P)\big)^2\right), & \mathcal{B} \\ \partial_t \hat{S} + \mathrm{div}(\hat{v}\,\hat{S}) & = & \lambda\,\delta_{\hat{\Gamma}}\left(-\big(1 - z - C_{\max}(1-z,S)\big)^2 + \big(1 - z - C_{\min}(1-z,S)\big)^2\right), & \mathcal{B} \\ \partial_t \hat{M} & = & -\alpha \hat{M}\hat{P}, & \mathcal{B} \\ \hat{v} & = & -\nabla\hat{\pi}, & \mathcal{B} \end{array}\right. \tag{1.20}$$

où $\lambda$ est une constante positive, appelée le gain, qui détermine à quel point l'on souhaite donner de l'importance à la correction par rapport au modèle. Le terme de correction est calculé à partir du gradient en $P$ de la mesure de similarité. **On montre théoriquement dans la thèse que la simulation corrigée converge bien vers la simulation cible lorsque la condition initiale de la cible est suffisamment régulière**.

A cette correction d'état, on ajoute une correction de paramètres par filtre de Kalman réduit, comme initié dans [47]. **Le schéma de correction jointe état-paramètres permet alors de tirer de l'information sur la vascularisation autour de la tumeur pendant la phase où les examens sont disponibles, afin d'améliorer la prédiction**

**lorsque l'on ne possède plus de données.** Cette méthode d'estimation jointe sur notre modèle est tout d'abord **validée sur des données synthétiques**. On parvient alors à corriger l'état même lorsque la condition initiale est décalée et à corriger les paramètres lorsqu'ils sont mal estimés initialement. La méthode est ensuite appliquée à des données réelles en 3D d'une métastase cérébrale. **La comparaison au temps du dernier examen entre la simulation et la tumeur réelle montre que la méthode améliore grandement la prédiction, et qu'elle parvient également à estimer grossièrement la vascularisation tumorale.**

# Bibliographie

## Bibliographie


[1] Lindsey A. Torre, Rebecca L. Siegel, Elizabeth M. Ward, and Ahmedin Jemal. Global Cancer Incidence and Mortality Rates and Trends—An Update. *Cancer Epidemiol Biomarkers Prev*, 25(1) :16–27, January 2016.

[2] Santé publique France - Le cancer en France métropolitaine : projections d'incidence et de mortalité par cancer en 2017.

[3] Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer : the next generation. *Cell*, 144(5) :646–674, March 2011.

[4] Gaorav P. Gupta and Joan Massagué. Cancer metastasis : building a framework. *Cell*, 127(4) :679–695, November 2006.

[5] Le cancer en chiffres | Fondation ARC pour la recherche sur le cancer.

[6] J. J. Casciari, S. V. Sotirchos, and R. M. Sutherland. Mathematical modelling of microenvironment and growth in EMT6/Ro multicellular tumour spheroids. *Cell Prolif.*, 25(1) :1–22, January 1992.

[7] Dirk Drasdo and Stefan Höhme. A single-cell-based model of tumor growth in vitro : monolayers and spheroids. *Phys Biol*, 2(3) :133–147, July 2005.

[8] Jacob G. Scott, Philip Gerlee, David Basanta, Alexander G. Fletcher, Philip K. Maini, and Alexander RA Anderson. Mathematical modeling of the metastatic process. *arXiv :1305.4622 [q-bio]*, May 2013. arXiv : 1305.4622.

[9] Bert Vogelstein, Surojit Sur, and Carol Prives. p53 - The most frequently altered gene in human cancers. *Nature Education at Scitable*, 3(9) :6, 2010.

[10] L. Hayflick. Mortality and immortality at the cellular level. A review. *Biochemistry Mosc.*, 62(11) :1180–1190, November 1997.

[11] Judah Folkman and Mark Hochberg. SELF-REGULATION OF GROWTH IN THREE DIMENSIONS. *J Exp Med*, 138(4) :745–753, October 1973.

[12] Cancer du côlon : Deux altérations génétiques à l'origine des métastases - Cancer colorectal - Patientsworld.

[13] J. Folkman. What is the evidence that tumors are angiogenesis dependent ? *J. Natl. Cancer Inst.*, 82(1) :4–6, January 1990.

[14] W. V. Mayneord. On a Law of Growth of Jensen's Rat Sarcoma. *The American Journal of Cancer*, 16(4) :841–846, July 1932.

[15] R. P. Araujo and D. L. S. McElwain. A history of the study of solid tumour growth : the contribution of mathematical modelling. *Bull. Math. Biol.*, 66(5) :1039–1091, September 2004.

[16] Helen M. Byrne. Dissecting cancer through mathematics : from the cell to the animal model. *Nat. Rev. Cancer*, 10(3) :221–230, March 2010.

[17] Thomas B. L. Kirkwood. Deciphering death : a commentary on Gompertz (1825) 'On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies'. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 370(1666), April 2015.

[18] A. K. Laird. DYNAMICS OF TUMOR GROWTH. *Br. J. Cancer*, 13 :490–502, September 1964.

[19] L. Norton. A Gompertzian model of human breast cancer growth. *Cancer Res.*, 48(24 Pt 1) :7067–7071, December 1988.

[20] L. Von Bertalanffy. Quantitative laws in metabolism and growth. *Q Rev Biol*, 32(3) :217–231, September 1957.

[21] L. A. Dethlefsen, J. M. Prewitt, and M. L. Mendelsohn. Analysis of tumor growth curves. *J. Natl. Cancer Inst.*, 40(2) :389–405, February 1968.

[22] D. Hart, E. Shochat, and Z. Agur. The growth law of primary breast cancer as inferred from mammography screening trials data. *Br. J. Cancer*, 78(3) :382–387, August 1998.

[23] Sébastien Benzekry, Clare Lamont, Afshin Beheshti, Amanda Tracz, John M. L. Ebos, Lynn Hlatky, and Philip Hahnfeldt. Classical Mathematical Models for Description and Prediction of Experimental Tumor Growth. *PLoS Comput Biol*, 10(8), August 2014.

[24] Andreas Deutsch and Sabine Dormann. *Cellular Automaton Modeling of Biological Pattern Formation : Characterization, Applications, and Analysis*. Modeling and Simulation in Science, Engineering and Technology. Birkhäuser Basel, 2005.

[25] Sabine Dormann and Andreas Deutsch. Modeling of self-organized avascular tumor growth with a hybrid cellular automaton. *In Silico Biol. (Gedrukt)*, 2(3) :393–406, 2002.

[26] Jörg Galle, Markus Loeffler, and Dirk Drasdo. Modeling the effect of deregulated proliferation and apoptosis on the growth dynamics of epithelial cell populations in vitro. *Biophys. J.*, 88(1) :62–75, January 2005.

[27] Stefan Hoehme and Dirk Drasdo. A cell-based simulation software for multi-cellular systems. *Bioinformatics*, 26(20) :2641–2642, October 2010.

[28] K. R. Swanson, E. C. Alvord, and J. D. Murray. A quantitative model for differential motility of gliomas in grey and white matter. *Cell Prolif.*, 33(5) :317–329, October 2000.

[29] Kristin R. Swanson, Carly Bridge, J. D. Murray, and Ellsworth C. Alvord. Virtual and real brain tumors : using mathematical modeling to quantify glioma growth and invasion. *J. Neurol. Sci.*, 216(1) :1–10, December 2003.

[30] D. Ambrosi and L. Preziosi. On the closure of mass balance models for tumor growth. *Math. Models Methods Appl. Sci.*, 12(05) :737–754, May 2002.

[31] Didier Bresch, Thierry Colin, Emmanuel Grenier, Benjamin Ribba, and Olivier Saut. A viscoelastic model for avascular tumor growth. report, 2009.

[32] Thierry Colin, Guillaume Dechriste, Jérôme Fehrenbach, Ludivine Guillaume, Valérie Lobjois, and Clair Poignard. Experimental estimation of stored stress within spherical microtissues. *Journal of Mathematical Biology*, 2018.

[33] Benjamin Ribba, Gentian Kaloshi, Mathieu Peyre, Damien Ricard, Vincent Calvez, Michel Tod, Branka Cajavec-Bernard, Ahmed Idbaih, Dimitri Psimaras, Linda Dainese, Johan Pallud, Stéphanie Cartalat-Carel, Jean-Yves Delattre, Jérôme Honnorat, Emmanuel Grenier, and François Ducray. A tumor growth inhibition model for low-grade glioma treated with chemotherapy or radiotherapy. *Clin. Cancer Res.*, 18(18) :5071–5080, September 2012.

[34] Víctor M. Pérez-García, Magdalena Bogdanska, Alicia Martínez-González, Juan Belmonte-Beitia, Philippe Schucht, and Luis A. Pérez-Romasanta. Delay effects in

the response of low-grade gliomas to radiotherapy : a mathematical model and its therapeutical implications. *Math Med Biol*, 32(3) :307–329, September 2015.

[35] Clair Poignard, A. Silve, and L. Wegner. Different Approaches used in Modeling of Cell Membrane Electroporation. report, Inria Bordeaux Sud-Ouest, July 2016.

[36] Damien Voyer, Aude Silve, Lluis M. Mir, Riccardo Scorretti, and Clair Poignard. Dynamical modeling of tissue electroporation. *Bioelectrochemistry*, 119 :98–110, February 2018.

[37] Kate E. R. Hollinshead, Debbie S. Williams, Daniel A. Tennant, and Christian Ludwig. Probing Cancer Cell Metabolism Using NMR Spectroscopy. *Adv. Exp. Med. Biol.*, 899 :89–111, 2016.

[38] Frank Willi Floeth, Michael Sabel, Gabriele Stoffels, Dirk Pauleit, Kurt Hamacher, Hans-Jakob Steiger, and Karl-Josef Langen. Prognostic Value of 18f-Fluoroethyl-l-Tyrosine PET and MRI in Small Nonspecific Incidental Brain Lesions.

[39] David N. Louis, Arie Perry, Guido Reifenberger, Andreas von Deimling, Dominique Figarella-Branger, Webster K. Cavenee, Hiroko Ohgaki, Otmar D. Wiestler, Paul Kleihues, and David W. Ellison. The 2016 World Health Organization Classification of Tumors of the Central Nervous System : a summary. *Acta Neuropathol.*, 131(6) :803–820, June 2016.

[40] Lian Tao Li, Guan Jiang, Qian Chen, and Jun Nian Zheng. Ki67 is a promising molecular target in the diagnosis of cancer (Review). *Molecular Medicine Reports*, 11(3) :1566–1572, March 2015.

[41] Recep Basaran, Serap Uslu, Berrin Gucluer, Mustafa Onoz, Nejat Isik, Mehmet Tiryaki, Cengiz Yakicier, Aydin Sav, and Ilhan Elmaci. Impact of 1p/19q codeletion on the diagnosis and prognosis of different grades of meningioma. *Br J Neurosurg*, 30(5) :571–576, October 2016.

[42] David N. Louis, Hiroko Ohgaki, Otmar D. Wiestler, Webster K. Cavenee, Peter C. Burger, Anne Jouvet, Bernd W. Scheithauer, and Paul Kleihues. The 2007 WHO Classification of Tumours of the Central Nervous System. *Acta Neuropathol*, 114(2) :97–109, August 2007.

[43] Thierry Colin, François Cornelis, Julien Jouganous, Jean Palussière, and Olivier Saut. Patient specific simulation of tumor growth, response to the treatment and relapse of a lung metastasis : a clinical case. *Journal of Computational Surgery*, page 18, 2015.

[44] Julien Jouganous. *Modélisation et simulation de la croissance de métastases pulmonaires*. Bordeaux, September 2015.

[45] T. F. Chan and L. A. Vese. Active contours without edges. *IEEE Transactions on Image Processing*, 10(2) :266–277, February 2001.

[46] Annabelle Collin, Dominique Chapelle, and Philippe Moireau. A Luenberger observer for reaction–diffusion models with front position data. *Journal of Computational Physics*, 300(Supplement C) :288–307, November 2015.

[47] Philippe Moireau, Dominique Chapelle, and Patrick Le Tallec. Joint state and parameter estimation for distributed mechanical systems. *Computer Methods in Applied Mechanics and Engineering*, 197(6-8) :659–677, 2008.

# Chapitre 2

# Classification des patients ayant un gliome en fonction du temps de rechute par apprentissage

**Sommaire**


**2.1 Introduction . . . . . 42**
**2.2 Méthodes d'apprentissage automatique appliquées à l'oncologie 44**
2.2.1 Principe général de l'apprentissage automatique . . . . . 44
2.2.2 Contraintes liées à l'application en oncologie . . . . . 49
2.2.3 Méthodes classiques d'apprentissage automatique . . . . . 51
**2.3 Données collectées par l'Humanitas Research Hospital et critère d'hétérogénéité . . . . . 56**
2.3.1 Base de données . . . . . 56
2.3.2 Critère d'hétérogénéité . . . . . 60
**2.4 Méthodes de pré-traitement . . . . . 63**
2.4.1 Sélection des caractéristiques . . . . . 63
2.4.2 Réduction de l'espace des caractéristiques : méthode d'effets mixtes 64
2.4.3 Comparaison des méthodes par validation croisée . . . . . 67
**2.5 Résultats de classification . . . . . 68**
2.5.1 Comparaison des algorithmes . . . . . 68
2.5.2 Courbe ROC . . . . . 68
2.5.3 Courbes de Kaplan-Meier . . . . . 71
**2.6 Discussion et conclusion . . . . . 79**

## 2.1 Introduction

Les gliomes représentent 80% des tumeurs malignes du cerveau [1, 2]. Ils proviennent des cellules gliales et sont classifiés selon la classification de 2016 de l'Organisation Mondiale de la Santé (OMS) selon la présence de la mutation IDH1 en deux catégories : les tumeurs avec gène IDH1 de type sauvage (wild-type), principalement représentées par les glioblastomes de grade IV avec un comportement très agressif et un mauvais pronostic, et les tumeurs avec gène IDH1 muté, classées en grade II et III. Les récentes publications ont en effet tendance à réduire le rôle du grade en tant que facteur lié au pronostic clinique et à favoriser le statut IDH1.

L'introduction de cette nouvelle classification aide les cliniciens à mieux stratifier les patients et à choisir le meilleur traitement selon la biologie de la tumeur. Malgré ces avancées, le pronostic clinique demeure incertain, en particulier pour les tumeurs ayant un gène IDH1 muté.

Le temps de survie sans progression (PFS en anglais) est le temps entre le traitement chirurgical et la rechute clinique de la maladie. La chirurgie est en effet souvent utilisée quel que soit le grade, lorsque la localisation de la tumeur le permet. Cependant, la nature infiltrante des gliomes empêche la résection complète de la tumeur [3]. Des traitements additionnels sont donc utilisés après la chirurgie pour contrôler l'évolution de la tumeur, comme la radiothérapie et la chimiothérapie. Cependant, chaque traitement comprend des risques d'effets indésirables et peut s'avérer lourd pour le patient. Choisir les traitements et les dates d'administration adéquats en fonction de l'agressivité de la tumeur et des caractéristiques de chaque patient permet d'améliorer le pronostic clinique et la qualité de vie des patients.

L'Association Européenne de Neuro-oncology a récemment publié de nouvelles lignes directrices pour le traitement des gliomes en fonction de la nouvelle classification, basées sur plusieurs essais cliniques randomisés et sur l'expérience clinique [4]. Les caractéristiques principales conduisant au choix du traitement sont : la biologie de la tumeur, l'état général du malade (déterminé par l'indice de Karnofsky [5]), le statut neurologique et l'étendue de la résection lors de la chirurgie. L'introduction de ces nouvelles directives suggère de traiter les gens appartenant à une même catégorie de manière homogène, alors que la maladie est, elle, très hétérogène. Le comportement clinique suite à ces traitements reste imprévisible, ce qui laisse à penser que les connaissances moléculaires et cliniques de cette pathologie

ne sont pas encore suffisamment puissantes pour classifier les patients. Un des élements qui est occulté lors de la classification actuelle est l'hétérogéneité tumorale, en particulier parce que les analyses histologiques et moléculaires sont exercées sur de petits échantillons de tumeur. Par contre, cette hétérogénité est visible sur l'imagerie médicale. Plusieurs publications expliquent que l'imagerie métabolique, comme le PET scan, peut aider à prédire l'agressivité de la tumeur, indépendamment des données biologiques de la lésion [6, 7]. Ces résultats sont prometteurs mais encore loin d'être optimaux.

**Afin de mieux sélectionner les patients nécessitant des thérapies additionnelles et pour trouver la bonne balance entre le traitement et la qualité de vie du patient, une meilleure classification de l'agressivité tumorale est requise**. Le développement des analyses dite de "radiomics" (combiner les données d'imagerie clinique et les données issues de biopsies) vont en ce sens et se développent en particulier pour les tumeurs du cerveau [8, 9]. Pour l'instant, la plupart de ces études n'utilisent que le grade et le statut IDH1 [10]. De plus l'analyse statistique est souvent univariée, en essayant de déterminer l'effet de chaque marqueur indépendamment [11, 12]. Cependant, l'incorporation d'autres données omiques et d'imagerie peuvent apporter d'autres informations permettant une meilleure analyse statistique. **Afin de combiner toutes ces informations, l'idée est d'appliquer des méthodes d'apprentissage automatique**. Elles sont très utilisés en oncologie, pour le diagnostic, la prédiction de rechute et la prédiction de survie [13, 14]. Notre but est de classifier les patients selon leur temps de survie sans progression (PFS), en combinant toutes les données génétiques et d'imagerie par apprentissage. **La classification ainsi obtenue est plus efficace que celles obtenues par le grade ou le statut IDH1, actuellement utilisés en clinique**.

## 2.2 Méthodes d'apprentissage automatique appliquées à l'oncologie

L'apprentissage automatique est un procédé de plus en plus utilisé dans des domaines aussi variés que la reconnaissance d'objets, les moteurs de recherche, l'analyse financière, ainsi que dans le domaine médical. Cette polyvalence s'explique par le fait que ces algorithmes déterminent un comportement en tirant de l'information à partir d'une base de données et sans a priori sur les différents comportements possibles. Cet apprentissage permet alors de structurer la base de données en recherchant des motifs récurrents. Des procédés statistiques permettent alors de prédire le comportement de nouveaux cas, en les comparant aux cas déjà étudiés. En oncologie, l'apprentissage automatique permet d'aider le médecin au diagnostic ou au pronostic clinique. Les bases de données peuvent contenir un grand nombre d'informations concernant le patient, issues de l'imagerie ou de la génomique. La figure 2.1 montre que de plus en plus d'articles d'oncologie utilisent des méthodes d'apprentissage automatique.

Ces méthodes concurrencent les modèles mécanistiques. En effet, la construction d'un modèle nécessite de comprendre les mécanismes biologiques afin de les exprimer sous forme d'équations régissant les densités de cellules tumorales. Pour limiter la complexité de ces modèles, beaucoup de simplifications doivent être effectuées, en occultant certains de ces mécanismes. L'avantage principal de l'apprentissage automatique par rapport à ces modèles est de pouvoir modéliser des comportements plus complexes, et d'intégrer tout type d'informations. La contre partie évidente est que l'apprentissage est plus obscur et ne permet pas de mieux comprendre la biologie du cancer. De plus, il nécessite de l'information sur toute une population, alors qu'un modèle permet de ne travailler qu'avec un individu. L'apprentissage automatique apporte donc quelque chose de nouveau en oncologie par rapport à la modélisation mécanistique, ce qui explique l'intérêt qui lui est porté. **Dans ce chapitre, la notion de modèle portera sur des modèles statistiques d'apprentissage, alors que dans le chapitre suivant, nous travaillerons avec des modèles mécanistiques d'équations aux dérivées partielles.**

### 2.2.1 Principe général de l'apprentissage automatique

L'apprentissage automatique (machine learning) consiste à interpréter une base de données de la façon suivante : à partir de caractéristiques $X = (X_1, \dots X_p)$, nous souhaitons inférer une fonction $Y = f(X)$. Or la fonction $Y$ est bruitée, ce qui signifie que la prédiction ne

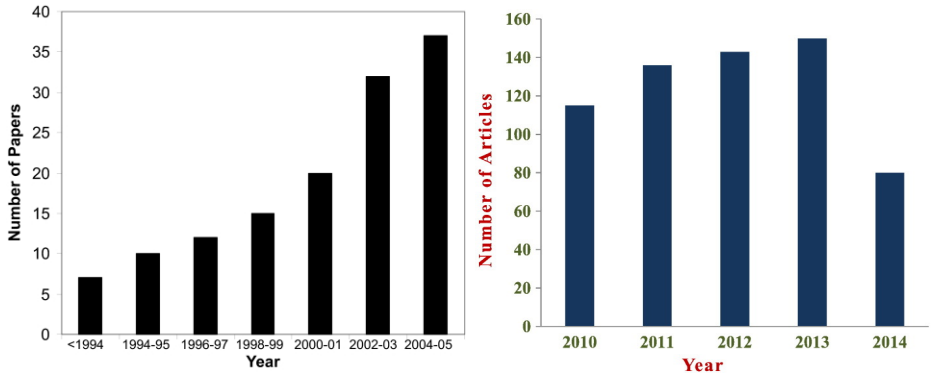



FIGURE 2.1 – Evolution du nombre d'articles d'apprentissage automatique appliqués à l'oncologie. Etudes réalisées en 2007 [13], et 2014 [15].

sera jamais parfaite, mais le but est de minimiser l'erreur entre l'estimation $\hat{f}(X)$ et la vraie valeur $f(X)$. Le but est donc d'apprendre cette fonction $\hat{f}$ pour ensuite l'appliquer à une nouvelle base de données dont on ne connaît que les valeurs $X$, et dont on souhaite estimer $Y$. Contrairement à la modélisation, il n'y a pas d'*a priori* sur la fonction $f$ à déterminer.

Sur la figure 2.2, un exemple simple de problème à deux caractérisitiques est représenté. L'âge du patient est représenté par $X_1$, et le volume tumoral est représenté par $X_2$. Le but est de déterminer le pronostic clinique de ces patients à partir de ces deux caractéristiques, en utilisant les données représentées sur la figure : on cherche à différencier les patients qui rechutent (carrés bleus) des patients qui ne rechutent pas (ronds rouges). Dans cet exemple, le motif que l'on observe est qu'un patient âgé ayant une grosse tumeur a plus tendance à rechuter qu'un jeune patient ayant une petite tumeur. L'idée de l'apprentissage automatique est de formaliser cela en séparant explicitement les deux comportements, par exemple par la ligne en pointillé. Certes, une erreur est commise dans l'apprentissage, puisqu'un patient bleu se retrouve du côté rouge. L'idée est de trouver un bon compromis entre la minimisation de l'erreur et la simplicité de l'algorithme d'apprentissage choisi.

*Apprentissage supervisé et non supervisé*

On distingue deux types de problèmes d'apprentissage automatique : l'apprentissage supervisé et non supervisé. En apprentissage supervisé, la variable $Y$ est connue pour la base de données d'apprentissage, et doit être inférée pour la base de données de prédiction.

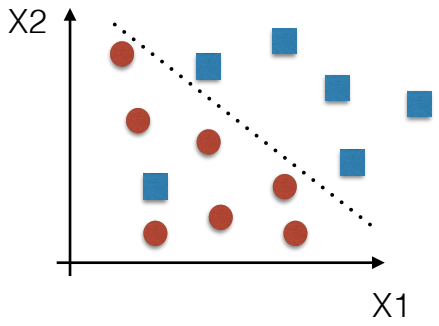



FIGURE 2.2 – Exemple simple de situation à deux caractéristiques (âge et volume tumoral). Les carrés bleus représentent les patients qui rechutent au traitement, et les ronds rouges sont ceux qui ne rechutent pas. La ligne en pointillé correspond à l'apprentissage effectué sur le jeu de données.

C'est le cas le plus courant en oncologie. En apprentissage non supervisé, on ne connaît pas la variable $Y$ : le but est alors de trouver une structure sur les points $X$, de déterminer les variables les plus importantes, et aussi de combiner ces variables afin d'en tirer un maximum d'information. Un exemple courant d'apprentissage non automatisé est de diviser un ensemble de personnes en groupes d'individus qui se ressemblent entre eux. Ce genre de méthodes sert souvent en amont de l'apprentissage, en pré-traitement des données.

*Classification et régression*

Dans le cas d'apprentissage supervisé, on distingue également les problèmes de classification et les problèmes de régression. Dans le premier cas, la variable $Y$ prend des valeurs discrètes dans l'ensemble $\{0, ..., n\}$, où $n$ est le nombre total de classes. C'est le cas de l'exemple ci dessus, où la variable $Y$ prend la valeur binaire : 0 si le patient ne rechute pas, et 1 s'il rechute. Lorsque la variable $Y$ est une variable continue, on parle de régression. Le but n'est alors plus d'estimer l'appartenance à une classe donnée, mais de calculer un score pour chaque patient. Notons qu'un problème de régression peut être ramené à un problème de classification en définissant des seuils pour chaque classe.

*Exemples d'applications en oncologie*

En oncologie, l'apprentissage automatique peut être appliqué à plusieurs problèmes liés au suivi médical. La détection de tumeur et la rechute d'un patient sont des exemples de

classification supervisée. De plus, l'estimation du temps de rechute ou du temps de survie sont des problèmes de régression supervisée classiques. L'apprentissage automatique peut également être utilisé en imagerie médicale, pour la détection et le contourage d'une tumeur. En imagerie, chaque pixel ou voxel correspond à une caractéristique. Les méthodes d'apprentissage automatique mises en place pour l'imagerie médicale sont différentes de celles utilisées pour le suivi médical. En effet, le nombre de caractéristiques est bien plus important, et le nombre de données également. En suivi médical, le nombre de patients est limité et les caractéristiques sont obtenues à partir de mesures sur l'imagerie ou de biopsie. Dans les parties suivantes, nous nous limiterons aux problématiques issues du suivi médical, et plus particulièrement au cas le plus fréquent de classification supervisée. En effet, la régression nécessite d'avoir une grande base de données, ce qui est rare en oncologie. De plus, les algorithmes non supervisés seront utilisés en pré-traitement, mais les problématiques en oncologie sont essentiellement supervisées.

*Ensemble d'apprentissage et ensemble test*

Comme évoqué précédemment, une méthode d'apprentissage automatique passe par deux phases : une phase d'apprentissage et une phase de test. Ces deux phases se font sur deux ensembles de patients distincts : un ensemble d'apprentissage et un ensemble de test. Cette indépendance entre les deux ensembles est essentielle afin de pouvoir évaluer correctement l'efficacité de l'algorithme, sans biais. Sur la figure 2.8, un problème classique de l'apprentissage est représenté. Un algorithme trop simple appliqué à la base de données peut résulter en un problème de sous-apprentissage, c'est-à-dire qu'aucun apprentissage n'est réellement effectué sur la base de donnée. A l'inverse, un modèle trop complexe peut mener à un problème de sur-apprentissage, c'est à dire que le modèle colle parfaitement aux données, mais la dynamique globale des données n'est pas apprise par l'algorithme. Dans ce cas, l'erreur commise sur l'ensemble d'apprentissage est minime, mais l'erreur sur l'ensemble test risque d'être importante. Au centre de la figure, un bon apprentissage est représenté, dans le sens où le comportement général de la population est appris.

Ces trois situations montrent donc que l'efficacité d'un algorithme doit se calculer à partir de l'ensemble test (qui est justement défini pour cela), et pas à partir de l'ensemble d'apprentissage, pour éviter le problème de sur-apprentissage.
Dans le cas des algorithmes possédant des paramètres à ajuster, l'ensemble d'apprentissage est lui même séparé en deux sous-ensembles : un premier servant à l'apprentissage habituel, et un second servant d'ensemble de validation, sur lequel les différents paramètres

sont testés. Le modèle, muni des paramètres donnant les meilleurs résultats sur l'ensemble de validation, est ensuite testé sur l'ensemble test.

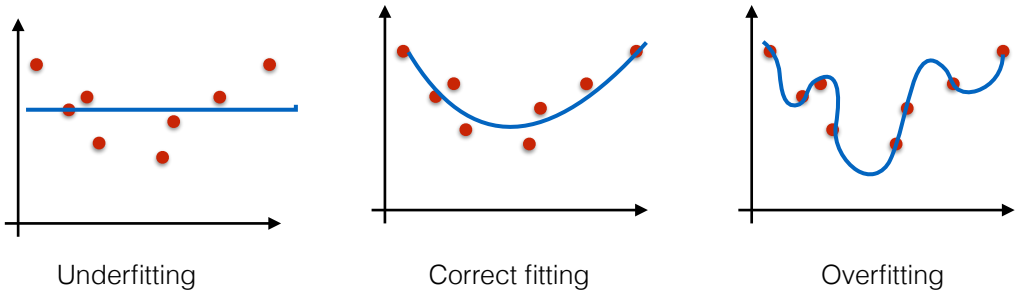



FIGURE 2.3 – Description de trois différents apprentissages : le sous-apprentissage (underfitting), le bon apprentissage (correct fitting) qui décrit bien le comportement global de la population, et le sur-apprentissage (overfitting).

*Choix du partitionnement apprentissage et test*

Différents choix sont possibles pour le partitionnement de la base de données en ensemble d'apprentissage et de test. Le choix le plus simple est de séparer la base de données en deux ensembles disjoints. Habituellement, l'ensemble d'apprentissage contient 80% de la base de données, et l'ensemble test contient les 20% restants. Afin de limiter la variabilité des résultats, il est également possible de répéter cette méthode plusieurs fois, avec des ensembles choisis aléatoirement.

Dans le cas où la base de données est de petite taille, il est cependant préférable d'utiliser une méthode de validation croisée en $k$ sous-ensembles ($k$ fold cross-validation). La base de données est divisée aléatoirement en $k$ sous-ensembles de taille égale, et chacun de ces ensembles est tour à tour l'ensemble test, les autres faisant partie de la base d'apprentissage. Cette méthode a l'avantage de permettre de classer tous les patients et est très adaptée aux petits jeux de données. Le taux d'erreur d'un algorithme est donc égal à la moyenne des pourcentages de mauvaises classifications dans chaque ensemble. Les paramètres optimaux de chaque méthode sont estimés sur la base d'apprentissage, la base de test ne servant qu'à la prédiction. De même que pour la première méthode, on peut répéter cette séparation aléatoire en $k$ sous-ensembles un grand nombre de fois pour diminuer la variabilité des résultats.

| **Réel** / **Estimé** | Pas de rechute | Rechute |
|---|---|---|
| Pas de rechute | a | b |
| Rechute | c | d |

TABLE 2.1 – Comparaison entre le résultat d'une classification par apprentissage automatisé et la classe réelle dans la base de données.

*Evaluation de l'efficacité d'un algorithme*

Après la phase d'apprentissage, le modèle est appliqué à l'ensemble test. Afin d'évaluer l'algorithme d'apprentissage automatique, on compare alors la prédiction à la valeur réelle $Y$. Le score de l'algorithme peut être calculé de plusieurs manières différentes. La plus simple est de regarder le ratio global de bonnes prédictions. Dans le cas du tableau 2.1, ce ratio vaut $\frac{a+d}{a+b+c+d}$. Dans certains cas, la sensibilité (ratio de vrais positifs) ou la spécificité (ratio de vrais négatifs) peuvent être choisies pour évaluer l'efficacité. Dans le cas du tableau 2.1, la sensibilité vaut $\frac{d}{b+d}$ alors que la sensibilité vaut $\frac{a}{a+c}$. Lorsque l'on estime si un patient va rechuter ou non, il peut être intéressant de minimiser le nombre de patients estimés sans rechute mais qui rechutent, c'est-à-dire de minimiser le nombre de faux négatifs ($\frac{b}{b+d}$ dans notre exemple), et donc de maximiser la sensibilité. Enfin, l'aire sous la courbe ROC est un autre indicateur qui permet d'évaluer et de comparer des algorithmes d'apprentissage. La courbe ROC est le tracé de la sensibilité en fonction de la valeur $(1-\text{spécificité})$, lorsque le seuil de décision d'appartenance à une classe ou à une autre varie. Le calcul de cet indicateur est détaillé dans notre exemple d'étude ci-dessous.

### 2.2.2 Contraintes liées à l'application en oncologie

L'application de méthodes d'apprentissage automatique à des problèmes d'oncologie soulève quelques contraintes, listées ci-dessous.

*Petite base de données*

La base de données est constituée de patients souffrant d'un type de tumeur donnée. Habituellement, ces bases de données contiennent les informations extraites d'un ou deux hôpitaux, ce qui limite la taille de cette base. Une autre difficulté dans l'acquisition des données provient du fait que les caractéristiques issues de l'imagerie doivent être calculées

à partir de machines semblables, afin de ne pas biaiser la comparaison. Les méthodes d'apprentissage automatique doivent donc être assez simples pour éviter le sur-apprentissage. Une autre possibilité est d'agrandir la base de données en ajoutant des patients artificiels, obtenus par combinaison des caractéristiques des autres patients [16].

*Informations manquantes*

Puisque les examens effectués sur un patient peuvent varier d'un individu à l'autre, le tableau de données collectées peut contenir des informations manquantes. Si certains algorithmes d'apprentissage automatique peuvent gérer ce problème, la plupart d'entre eux requièrent une base de données complète. Un pré-traitement doit donc être effectué. Dans le cas d'une caractéristique manquante pour une grande partie des patients, il est préférable de la retirer, puisqu'elle n'apportera pas d'information. Lorsqu'elle manque pour un petit nombre de cas, trois solutions sont envisageables. La première est de remplacer la valeur absente par une constante. Cela revient à considérer que le manque d'information est une information en soit. La seconde est de la remplacer par la valeur moyenne calculée sur les autres patients. Cela revient à limiter son influence pour les patients dont l'information est manquante. La dernière méthode est d'utiliser des méthodes d'apprentissage automatique où la caractéristique concernée est la variable à prédire.

*Classes déséquilibrées*

Lors de la classification de patients, il est possible en oncologie que la plupart des patients soient dans la même classe. Il y a alors un déséquilibre dans les classes, et traiter classiquement le problème risque de conduire à une prédiction uniforme. Afin de gérer au mieux ce problème, il faut s'assurer que lors du partitionnement en ensemble d'apprentissage et en ensemble test, la proportion globale des classes globale est respectée. En effet, si seule la classe majoritaire est contenue dans l'ensemble d'apprentissage, la classe minoritaire ne sera jamais estimée. Il est également possible de pénaliser le modèle, en ajoutant un coût plus élevé pour les erreurs de classification de la classe minoritaire.

*Réduction de l'espace des caractéristiques*

La dimension de l'espace des caractéristiques doit toujours être largement inférieure au nombre de patients de la base de données. En effet, l'apprentissage sera meilleur si tous les

comportements sont présents dans la base de données. Puisque les bases de données sont de petite taille, il est nécessaire de réduire l'espace des caractéristiques.

La première façon de le faire est de sélectionner les caractéristiques les plus influentes. Cela peut se faire grâce à une analyse univariée de chaque caractéristique, qui mesure la corrélation avec la variable à estimer. Cela permet d'enlever les caractéristiques les moins influentes. Il est aussi possible d'utiliser une méthode d'apprentissage automatique qui garde en mémoire quelles caractéristiques ont été le plus souvent utilisées pour la classification (les forêts aléatoires par exemple). L'avantage de la seconde méthode est qu'elle permet d'enlever les caractéristiques redondantes en plus des caractéristiques moins influentes, ce qui n'est pas le cas de l'analyse univariée. Cette première réduction a pour but d'enlever grossièrement les caractéristiques qui n'apportent aucune information supplémentaire.

La seconde manière de réduire l'espace des caractéristiques est d'appliquer une méthode non supervisée telle que l'analyse en composante principale. Cela permet en effet de trouver les combinaisons de caractéristiques qui représentent au mieux les données, et de projeter la base de données sur ces composantes. Puisque cette étape est non supervisée, la valeur de la variable $Y$ n'influe pas sur la projection à effectuer. Un autre intérêt de cette méthode est que les nouvelles caractéristiques projetées sont indépendantes entre elles, ce qui est préférable pour certains algorithmes (le classifieur bayésien par exemple).

*Données qualitatives et quantitatives*

Les bases de données issues du domaine médical contiennent des caractéristiques qualitatives (à valeurs discrètes) et quantitatives (à valeurs continues). Cela altère le fonctionnement de certains algorithmes comme la méthode des plus proches voisins, puique la notion de distance est ambigüe lorsque ces deux types de données sont mélangés. Dans l'exemple d'étude ci-dessous, nous verrons une méthode d'analyse mixte qui permet un pré-traitement des données.

### 2.2.3 Méthodes classiques d'apprentissage automatique

Dans cette section, les algorithmes de classification utilisés par la suite sont détaillés.

### Méthode des plus proches voisins

La méthode des plus proches voisins (KNN, K Nearest Neighbor) [17] est très intuitive : l'idée est que les patients dont les caractéristiques sont proches ont de fortes chances de se comporter de la même manière. L'apprentissage par méthode des $k$ plus proches voisins consiste donc à chercher les $k$ points aux coordonnées les plus proches du point que l'on cherche à prédire, dans l'espace des caractéristiques. La distance entre les points est habituellement une distance euclidienne, mais d'autres distances peuvent être utilisées [18]. Sur la figure 2.4, un exemple avec $k = 4$ est représenté. La classe du point à prédire est ensuite estimée en effectuant une moyenne des classes de ses plus proches voisins pondérée par les distances respectives avec le point à prédire. Le résultat est comparé à un seuil (habituellement, ce seuil vaut 0.5) afin de classer le patient dans la classe 0 ou 1. A noter que cet algorithme permet également d'effectuer une régression. L'intérêt majeur de cette méthode est qu'elle est intuitive et qu'elle permet de modéliser un comportement non linéaire. De plus, les caractéristiques non pertinentes sont naturellement inutilisées par cette méthode. Cependant, l'inconvénient principal est le fait que chaque caractéristique a la même importance au moment de la recherche des voisins proches.

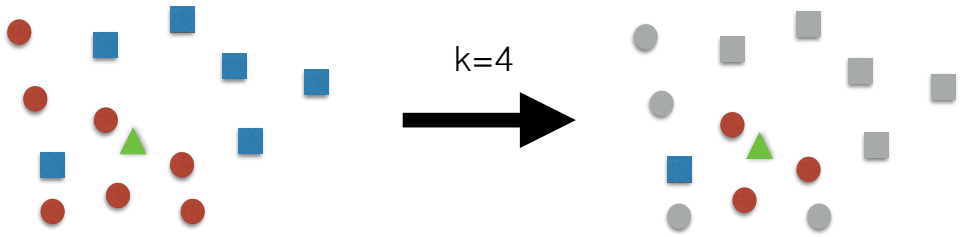



FIGURE 2.4 – Schéma de la méthode des plus proches voisins, pour $k = 4$. Le patient représenté par le triangle vert sera classifié ici dans la classe des ronds rouges.

### Arbre binaire et forêts aléatoires

Un arbre binaire [17] est un graphe connexe orienté où chaque noeud a au plus deux fils, et tous les noeuds sauf la racine ont exactement un père. En classification, chaque noeud correspond à un sous-ensemble de l'ensemble d'apprentissage. Si un noeud admet deux fils, les deux sous-ensembles associés à ces fils forment une partition du sous-ensemble associé au noeud père. Cette partition se fait à partir du choix d'une caractéristique et d'une valeur seuil : un noeud fils contient les individus du sous-ensemble du noeud père

dont la caractéristique est supérieure à la valeur seuil, tandis que l'autre fils contient les autres individus. La construction de l'arbre se fait donc par choix successifs de caractéristiques et de seuils associés. La manière la plus classique d'effectuer ces choix se base sur le critère de Gini qui évalue l'homogénéité de chacun des sous-ensemble des noeuds fils $j$ : $G^j = \sum_{i=1}^{n} p_i^j(1 - p_i^j)$ où $p_i^j$ est la proportion du sous-ensemble associé au noeud $j$ qui appartient à la classe $i$. L'indicateur retenu est alors $G = \max_j G^j$ et on choisit alors à chaque étape la caractéristique et le seuil minimisant cet indicateur, c'est-à-dire celui qui assure une bonne homogénéité de classe à chaque noeud. Un taux d'homogénéisation minimal ou une profondeur de l'arbre maximale permettent d'arrêter la construction de l'arbre et éviter le surapprentissage. La classification par arbre binaire revient ainsi à couper le domaine en zones rectangulaires suffisamment homogènes.

La figure 2.5 représente un cas simple de construction d'arbre dans le cas d'un espace de caractéristiques de dimension 2. La prédiction obtenue est donc la classe prédominante dans la case auquelle le patient appartient. L'idée des forêts aléatoires est de calculer ces arbres binaires un grand nombre de fois, en sélectionnant aléatoirement certaines caractéristiques et certains patients à chaque fois. La prédiction finale est alors la moyenne des prédictions de tous les différents arbres binaires. En plus d'être un principe intuitif, cette méthode est robuste vis-à-vis du surapprentissage et peut traiter des données à la fois continues et discrètes. Cependant, l'apprentissage peut être long et les erreurs dans la base d'apprentissage peuvent entraîner une complexification des arbres.

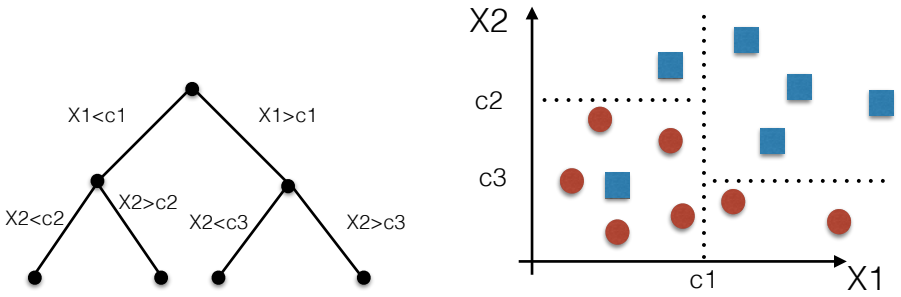



FIGURE 2.5 – Schéma de la méthode des arbres de décision. Gauche : construction de l'arbre. Droite : Cases formées par l'arbre de décision dans l'espace des caractéristiques.

### Classifieur bayésien naïf

Le classifieur bayésien naïf [17] repose sur le théorème de Bayes qui donne :

$$P(Y|X_1, X_2, ..., X_n) = \frac{P(Y)P(X_1, X_2, ..., X_n|Y)}{P(X_1, X_2, ..., X_n)}. \tag{2.1}$$

On en déduit une manière d'approcher la probabilité d'être dans une classe sachant la valeur des caractéristiques, en faisait l'hypothèse que les caractéristiques sont deux à deux indépendantes :

$$f(x_1, x_2, ..., x_n) = \underset{y}{\text{argmax}}\ P(Y = y) \prod_{i=1}^{n} P(X_i = x_i|Y = y). \tag{2.2}$$

Le dénominateur du théorème de Bayes n'étant pas dépendant de $y$, il n'apparait pas dans la fonction à maximiser. Si cette simple méthode permet un apprentissage efficace, sa principale limite réside dans l'hypothèse d'indépendance des caractéristiques, qui est rarement vérifiée en pratique. De plus, le terme en $P(Y = y)$ montre que le résultat dépend fortement de la fréquence d'apparition de chaque classe, ce qui peut poser problème dans le cas de données déséquilibrées.

### Séparateurs à vaste marge (SVM)

L'idée des séparateurs à vaste marge (Support Vector Machines) [17] est d'appliquer une transformation à l'espace de départ afin de se retrouver dans un espace dans lequel les classes sont linéairement séparables, comme le montre le schéma de la figure 2.6. Une fois la transformation faite, la droite choisie pour séparer est celle qui maximise la marge, c'est-à-dire l'écart au plus proche point de chaque classe. L'avantage est de pouvoir modéliser des comportements non linéaires et de réduire la complexité à un problème d'optimisation quadratique. Cependant, les paramètres peuvent être difficiles à estimer lorsque les données ne sont pas linéairement séparables.

### Régression Logistique

La méthode de régression logistique [19] consiste à supposer que la probabilité d'appartenir à une classe sachant la valeur des caractéristiques suit une loi logistique, à savoir :

$$\ln\left(\frac{P(Y = 1|X_1, X_2, ..., X_n)}{1 - P(Y = 1|X_1, X_2, ..., X_n)}\right) = a_0 + \sum_{i=1}^{n} a_i x_i, \tag{2.3}$$

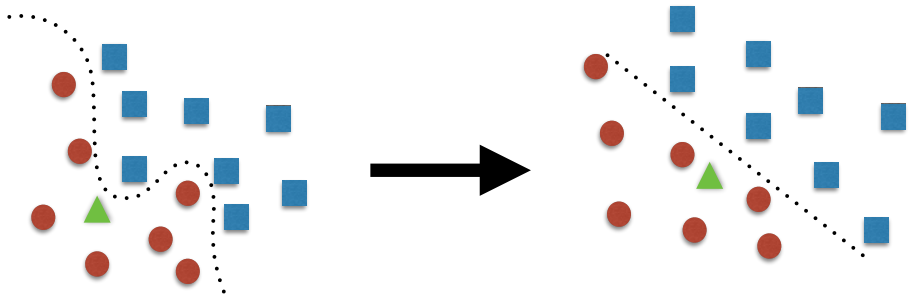


FIGURE 2.6 – Schéma de la méthode de séparateur à vaste marge. La transformation appliquée à l'espace des caractéristiques permet de se ramener à une séparation linéaire des données.

ce qui donne :

$$P(Y = 1|X_1, X_2, ..., X_n) = \frac{\exp(a_0 + \sum_{i=1}^{n} a_i x_i)}{1 + \exp(a_0 + \sum_{i=1}^{n} a_i x_i)}. \tag{2.4}$$

Les paramètres $(a_i)$ sont ensuite estimés par la méthode du maximum de vraisemblance. L'obtention de résultats stables avec cette méthode recquiert une grosse base de données, ce qui n'est pas toujours le cas en oncologie. L'avantage de cette méthode est qu'elle permet d'inclure des interactions explicites entre les caractéristiques et qu'aucun pré-traitement n'est nécessaire.

#### Réseaux de neurones artificiels

Les réseaux de neurones artificiels [17] sont l'exemple le plus simple des méthodes d'apprentissage profond (deep learning). Sur la figure 2.7, un exemple de petit réseau est représenté. Les entrées (inputs) sont les caractéristiques, qui sont envoyées sur une succession de noeuds. La valeur à ces noeuds est calculée en faisant la combinaison linéaire des valeurs entrantes, pondérée par les poids des arêtes. Le nombre de paramètres à estimer correspond donc au nombre d'arêtes du réseau, ainsi qu'aux valeurs seuils choisies afin de déterminer la classe sortante (output). La robustesse de cette méthode vis-à-vis du bruit, ainsi que la possibilité d'ajouter plusieurs sorties font de cette méthode l'une des plus utilisées en apprentissage automatique. Cependant, la structure est complexe et beaucoup de données sont nécessaires pour estimer tous les paramètres. Le surapprentissage est fréquent lorsque la base de données est plus petite, comme c'est le cas ici.

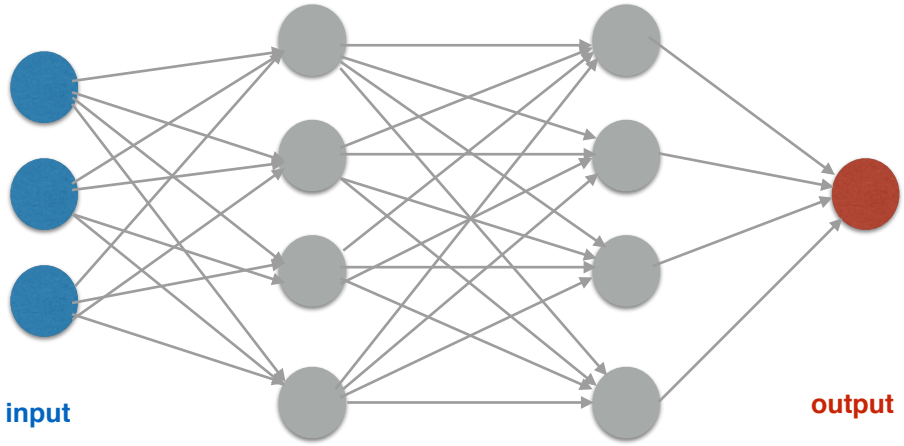



FIGURE 2.7 – Schéma des réseaux de neurones artificiels. L'entrée (input) et la sortie (output) du réseau correpondent respectivement aux caractéristiques et à la classe prédite.

## 2.3 Données collectées par l'Humanitas Research Hospital et critère d'hétérogénéité

### 2.3.1 Base de données

Cette étude rétrospective a été faite en collaboration avec l'Humanitas Research Hospital de Milan, et plus particulièrement avec le Pr Bello et le Dr Rossi. Au total, 90 patients atteints de gliomes de bas grade (II ou III) participent à cette étude. Ils ont tous subi une chirurgie partielle ou totale, suivie de traitements adjuvants, déterminés selon les lignes directrices de l'Association Européenne de Neuro-Oncology. Les données utilisées ici sont uniquement des données acquises juste avant (données d'imagerie ou démographiques) ou juste après (données génétiques) la chirurgie. Aucune information longitudinale n'a été utilisée, en particulier parce que ces informations n'étaient pas disponibles pour tous les patients. Le tableau de la figure 2.9 contient toutes les données acquises pour chaque patient.

— **IRM.** A partir de l'IRM pré-opératoire le volume tumoral ainsi que la position de la tumeur sont déterminés.

— **MET-PET scan.** Les indicateurs de l'activité métabolique dans la tumeur sont calculés sur un MET-PET scan (PET scan utilisant de la 11C-methionine comme traceur) [20, 21]. Ils comprennent des informations sur l'intensité du signal (SUVmax

**Table 1a**
Publications relevant to ML methods used for cancer susceptibility prediction.

| Publication | Method | Cancer type | No of patients | Type of data | Accuracy | Validation method | Important features |
|---|---|---|---|---|---|---|---|
| Ayer T et al. [19] | ANN | Breast cancer | 62,219 | Mammographic, demographic | AUC = 0.965 | 10-fold cross validation | Age, mammography findings |
| Waddell M et al. [44] | SVM | Multiple myeloma | 80 | SNPs | 71% | Leave-one-out cross validation | snp739514, snp521522, snp994532 |
| Listgarten J et al. [45] | SVM | Breast cancer | 174 | SNPs | 69% | 20-fold cross validation | snpCY11B2 (+) 4536 T/C snpCYP1B1 (+) 4328 C/G |
| Stajadinovic et al. [46] | BN | Colon carcinomatosis | 53 | Clinical, pathologic | AUC = 0.71 | Cross-validation | Primary tumor histology, nodal staging, extent of peritoneal cancer |

**Table 1b**
Publications relevant to ML methods used for cancer recurrence prediction.

| Publication | ML method | Cancer type | No of patients | Type of data | Accuracy | Validation method | Important features |
|---|---|---|---|---|---|---|---|
| Exarchos K et al. [24] | BN | Oral cancer | 86 | Clinical, imaging tissue genomic, blood genomic | 100% | 10-fold cross validation | Smoker, p53 stain, extra-tumor spreading, TCAM, SOD2 |
| Kim W et al. [47] | SVM | Breast cancer | 679 | Clinical, pathologic, epidemiologic | 89% | Hold-out | Local invasion of tumor |
| Park C et al. [48] | Graph-based SSL algorithm | Colon cancer, breast cancer | 437<br>374 | Gene expression, PPIs | 76.7%<br>80.7% | 10-fold cross validation | BRCA1, CCND1, STAT1, CCNB1 |
| Tseng C-J et al. [49] | SVM | Cervical cancer | 168 | Clinical, pathologic | 68% | Hold-out | pathologic_S, pathologic_T, cell type RT target summary |
| Eshlaghy A et al. [34] | SVM | Breast cancer | 547 | Clinical, population | 95% | 10-fold cross validation | Age at diagnosis, age at menarche |

**Table 1c**
Publications relevant to ML methods used for cancer survival prediction.

| Publication | ML method | Cancer type | No of patients | Type of data | Accuracy | Validation method | Important features |
|---|---|---|---|---|---|---|---|
| Chen Y-C et al. [50] | ANN | Lung cancer | 440 | Clinical, gene expression | 83.5% | Cross validation | Sex, age, T_stage, N_stage LCK and ERBB2 genes |
| Park K et al. [26] | Graph-based SSL algorithm | Breast cancer | 162,500 | SEER | 71% | 5-fold cross validation | Tumor size, age at diagnosis, number of nodes |
| Chang S-W et al. [32] | SVM | Oral cancer | 31 | Clinical, genomic | 75% | Cross validation | Drink, invasion, p63 gene |
| Xu X et al. [51] | SVM | Breast cancer | 295 | Genomic | 97% | Leave-one-out cross validation | 50-gene signature |
| Gevaert O et al. [52] | BN | Breast cancer | 97 | Clinical, microarray | AUC = 0.851 | Hold-Out | Age, angioinvasion, grade MMP9, HRASLA and RAB27B genes |
| Rosado P et al. [53] | SVM | Oral cancer | 69 | Clinical, molecular | 98% | Cross validation | TNM_stage, number of recurrences |
| Delen D et al. [54] | DT | Breast cancer | 200,000 | SEER | 93% | Cross validation | Age at diagnosis, tumor size, number of nodes, histology |
| Kim J et al. [36] | SSL Co-training algorithm | Breast cancer | 162,500 | SEER | 76% | 5-fold cross validation | Age at diagnosis, tumor size, number of nodes, extension of tumor |

FIGURE 2.8 – Exemples des problèmes et des méthodes d'apprentissage automatique utilisées en oncologie [15].

SUVratio, etc...) ainsi que sur le volume tumoral métabolique (MTV,MTB) qui correspond au volume des zones actives dans la tumeur.

— **Génomique** Les données génomiques comprennent les expressions de plusieurs gènes corrélés à la prolifération tumorale ou à la réponse de la tumeur aux traitements (IDH1, 1p, 11q, MIB1, etc...).

L'objectif de cette étude est d'utiliser ces données afin de classifier les patients selon leur temps de survie sans progression, noté PFS (Progression Free Survival). Le schéma de la figure 2.10 montre comment ce PFS est calculé : c'est le temps entre la chirurgie et une éventuelle rechute du patient après le traitement. Sur la figure 2.11, l'histogramme des PFS de ces 90 patients est représenté, en fonction du grade de la tumeur. Estimer le PFS de patients de manière précise est difficile avec les données acquises. L'idée ici est plutôt de distinguer deux catégories de patients : ceux pour qui la rechute est rapide et ceux pour

qui la rechute est lente. Le seuil qui a été choisi par les cliniciens pour différencier ces deux comportements est de 30 mois. Ce seuil a également l'avantage d'être proche de la médiane des PFS des patients de la base de données, ce qui rend les algorithmes d'apprentissage automatique plus efficace. En effet, on dispose alors de 43 patients ayant une rechute rapide (moins de 30 mois) et de 47 patients ayant une rechute lente (plus de 30 mois), et donc de quasiment autant de cas de chaque classe. Notons que parmi les 90 patients, 37 d'entre eux n'ont pas rechuté au moment de l'étude. Cependant, le PFS de ces patients étant déjà au dessus de 30 mois, leur classe est connue. Le graphique de la figure 2.11 ne distingue pas les patients ayant rechuté ou non, et la valeur du PFS affichée est donc celle au moment de l'acquisition de la base de donnée. On remarque que les patients de grade II ont en moyenne un PFS plus long que ceux de grade III, ce qui est cohérent. Cependant, on compte un nombre non négligeable de patients de grade II rechutant rapidement (11 cas sur 43) et de patients de grade III qui rechutent plus tardivement que prévu (16 cas sur 47). **Cela suggère donc que la classification par grade n'est pas optimale et peut être améliorée**.

| Demographics | Age | 44 years (17-81) |
|---|---|---|
| Clinical | Lobe | 31 frontal, 14 temporal, 10 parietal, 3 insula, 32 multiple |
| | Cerebral hemisphere | 44 left - 46 right |
| | First intervention | 45 yes - 45 no |
| MRI | GD enhancing | 25 yes- 65 no |
| | Tumor volume | 41.7 +/- 44.3 $cm^3$ |
| MET-PET | SUVmax | 3.5 +/- 1.8 |
| | SUVnorm | 1.5 +/- 0.3 |
| | SUVmean | 2.3 +/- 1.1 |
| | SUVratio (SUVmax/SUVnorm) | 2.1 +/- 0.7 |
| | MTV | 13.3 +/- 17.0 $cm^3$ |
| | MTB (SUVmean x MTV) | 32.0 +/- 46.4 $cm^3$ |
| Histomolecular diagnosis | Grade | 43 II - 47 III |
| | MIB1 | 8.2 +/- 9.2 % |
| | IDH1 status | 56 mutated - 26 wild-type - 8 unknown |
| | 1p expression | 68 mutated - 6 wild-type - 16 unknown |
| | 1q expression | 49 mutated - 25 wild-type - 16 unknown |
| | Codeletion | 48 mutated - 26 wild-type - 16 unknown |
| | MGMT | 61 yes - 13 no - 16 unknown |
| Added data | Heterogeneity indicator | 1.4 +/- 0.3 |
| Follow-up | Extent of resection | 54 total - 32 subtotal - 2 partial - 2 biopsy |
| | Recurrence | 53 yes - 37 no |
| | | Grade II : 20 recurrent patients |
| | | Grade III : 33 recurrent patients |
| | Overall survival | 32.8 +/- 12.9 months |
| | PFS | 29.0 +/- 14.2 months |
| | | Grade II : 35 +/- 12 months |
| | | Grade III : 23 +/- 13 months |
| | Adjuvant Rx | 60 yes - 30 no |
| | Lost in follow up | 25 yes - 65 no |

FIGURE 2.9 – Caractéristiques de la base de données de 90 patients. La première partie du tableau contient les caractéristiques utilisées pour la classification et la seconde partie concerne le suivi médical. Les cases grisées sont les données quantitatives, les autres étant qualitatives. La donnée *GD Enhancing* est un indicateur de la présence de réhaussement ou non sur la modalité T1 avec produit de contraste (gadolinium).

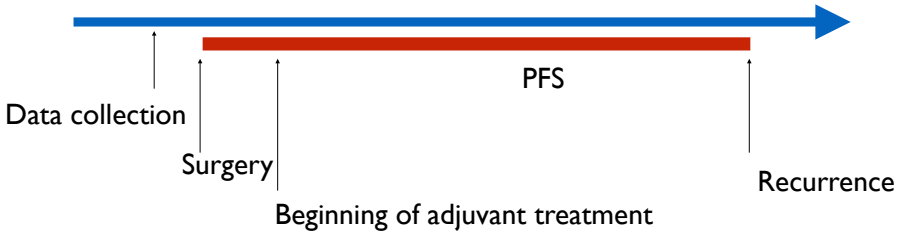



FIGURE 2.10 – Schéma du calcul du PFS : c'est le temps entre la chirurgie et la rechute éventuelle du patient. Lorsque les patients n'ont pas rechuté, on ne connaît pas leur PFS exact (ils peuvent rechuter plus tard dans le suivi clinique), mais on sait que leur PFS sera plus grand que le temps de suivi actuel.

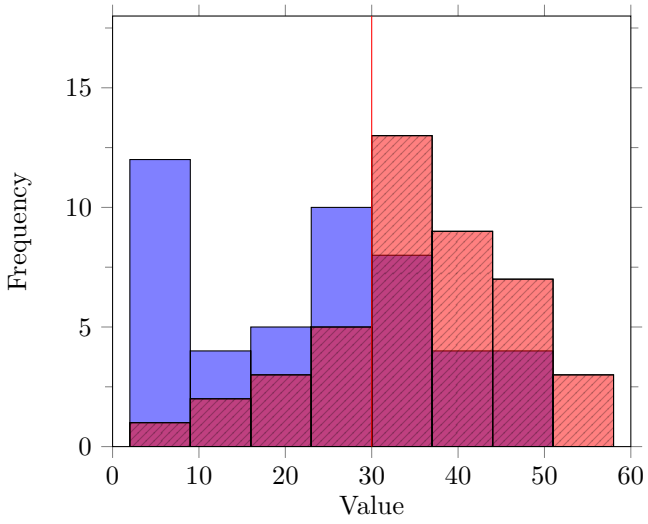



FIGURE 2.11 – Histogramme des PFS des 90 patients (en mois). Les barres bleues et pleines représentent les patients de grade III, et les barres rouges striées représentent les patients de grade II. Les zones violettes correspondent aux zones où les deux barres se superposent, donc ici les patients de grade II rechutant rapidement et les patients de grade III rechutant lentement. La ligne verticale représente le seuil de 30 mois, qui sépare les rechutes longues (plus de 30 moins) et courtes (moins de 30 mois).

### 2.3.2 Critère d'hétérogénéité

Parmi les données collectées par les cliniciens, aucune ne prend en compte l'hétérogénéité spatiale de la tumeur. En effet, les données liées au MET-PET scan sont des valeurs numériques calculées sur toute la tumeur (SUVmax, SUVratio, MTV). Cependant, le nombre de zones actives différentes aperçues sur le PET scan semble être une information qui peut être corrélée à l'agressivité de la tumeur. En effet, plusieurs zones actives peuvent signifier

un plus gros risque de rechute suite à la résection d'une partie de la tumeur. **Afin de prendre en compte cette hétérogéneité spatiale, un indicateur est mis en place**.

La figure 2.12 montre comment cet indicateur est calculé. Dans un premier temps, un atlas de cerveau [22] est recalé (recalage rigide) avec le signal PET du patient. Ce recalage nécessite une rotation et une translation de l'atlas du cerveau, effectuée grâce à la librairie PapriK, developpée dans l'équipe MONC. Ce recalage n'a pas besoin d'être extrêmement précis dans la mesure où il sert simplement à ne garder que le signal PET à l'intérieur du cerveau, et éliminer les signaux parasites des différents organes.

La segmentation du signal PET se fait par seuillage [23]. Afin de calculer le nombre de zones actives, l'idée était de trouver un seuillage puis de calculer le nombre de composantes connexes du signal résultant. Cependant, fixer un seuillage uniformément pour toute la base de données n'est pas pertinent dans la mesure où les intensités du signal sont très variées parmi les patients. L'idée ici est donc de compter le nombre de composantes connexes du signal résultant, lorsque la valeur du seuil balaye toutes les intensités de 0 à l'intensité maximale SUVmax. L'espace des valeurs de seuillage $[0, \text{SUVmax}]$ est donc discrétisé en 50 valeurs, et pour chaque seuil, le nombre de composantes connexes du signal résultant est calculé. Ce calcul du nombre de composantes connexes se fait par la fonction *scipy.ndimage.label* de Python. Notons qu'afin d'éviter le bruit du signal PET, seules les composantes connexes d'au moins 9 voxels sont prises en compte.

La courbe donnant le nombre de composantes connexes pour chaque seuil permet de comparer deux tumeurs, comme on peut le voir sur la figure 2.12. A gauche, la tumeur est hétérogène et son nombre de composantes connexes varie de 1 à 3. Au contraire, à droite, le signal est plus homogène puisque un seul spot principal est présent, qui se divise en deux petites parties pour des seuils élevés. Le critère qui est retenu afin de rendre compte de l'hétérogéneité du PET scan est l'intégrale de la courbe obtenue. L'avantage de ce choix est qu'il prend à la fois en compte le nombre maximal de zones actives, mais aussi l'éloignement entre deux zones actives. En effet, comme on le remarque dans le cas de droite, deux petites zones actives très proches ne vont pas beaucoup influer sur la valeur de l'intégrale. Au contraire, deux zones clairement distinctes et éloignées entraînent une intégrale plus grande. D'autres critères tels que le maximum de la courbe ou l'écart-type des pics ont été testés, mais l'intégrale est celui qui donne les meilleurs résultats de classification. Cet indicateur a été calculé pour les 90 patients, à partir des données du PET scan.

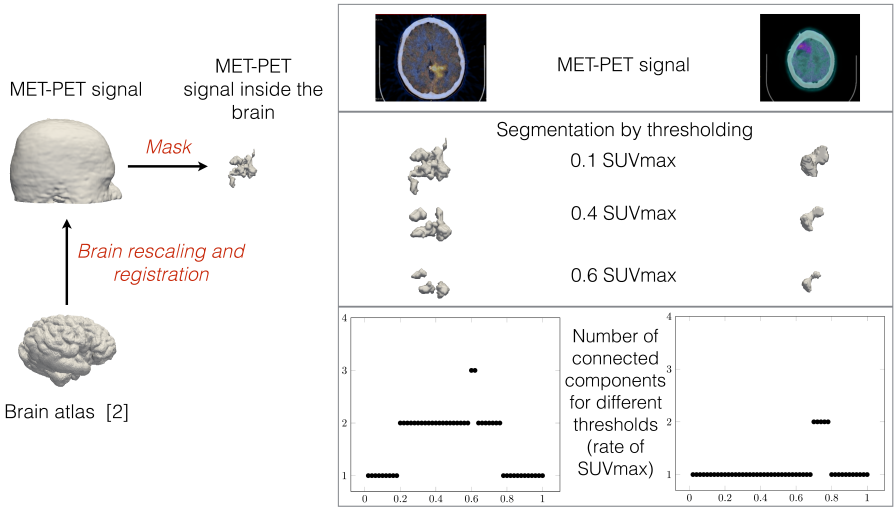



FIGURE 2.12 – Gauche : recalage d'un atlas de cerveau [24] sur le signal MET-PET. Droite : Comparaison de deux tumeurs de grade III : cas hétérogène (colonne de gauche) et homogène (colonne de droite). Le signal PET acquis (première ligne) est reconstruit en 3D (seconde ligne) par seuillage. Pour chaque seuil, le nombre de composantes connexes est calculé, ce qui donne la courbe du dessous (dernière ligne). L'intégrale de cette courbe est un indicateur de l'hétérogénéité tumorale.

## 2.4 Méthodes de pré-traitement

### 2.4.1 Sélection des caractéristiques

Afin d'appliquer à la base de données des méthodes d'apprentissage automatique, il est nécessaire de réduire dans un premier temps l'espace des caractéristiques. Parmi les 20 caractéristiques (19 acquises par les cliniciens, et l'indicateur d'hétérogénéité calculé précédemment), il faut sélectionner celles qui sont le plus fortement corrélées au PFS. Un trop grand nombre de caractéristiques produit du bruit dans les données et dégrade la classification. Un des avantages de la méthode des forêts aléatoires (random forest) [17] est qu'elle permet de calculer l'importance d'une caractéristique en recensant le nombre de fois où cette caractérisitique a été utilisée pour séparer les deux classes. Cette méthode a été présentée plus en détails dans la section concernant les méthodes d'apprentissage automatique.

En l'appliquant à notre jeu de données, les 11 caractéristiques qui se trouvent être les plus pertinentes dans l'étude du PFS sont représentées sur la figure 2.13. L'âge, le volume et le grade de la tumeur apparaissent fortement corrélées comme attendu. **Les données extraites du MET-PET scan sont également parmi les plus influentes, ce qui confirme l'intérêt porté à cette méthode d'imagerie fonctionnelle. En particulier, c'est le cas de l'indicateur d'hétérogénéité, ce qui confirme l'importance de l'hétérogénéité spatiale dans le comportement tumoral**. Les données génomiques telles que le statut IDH1 et la codélétion sont également influents, mais à un degré moindre. Ce statut IDH1 étant utilisé dans la nouvelle classification de l'agressivité des tumeurs, cette faible corrélation est plus surprenante. Cela peut s'expliquer par plusieurs raisons. La première est qu'un faible score d'une caractéristique ne signifie pas forcément que celle-ci est peu corrélée au PFS. En effet, il est également possible que l'information détenue par cette caractéristique soit également détenue dans d'autres caractéristiques qui elles sont plus souvent utilisées par l'algorithme. Une autre des explications possibles est le fait que les données génomiques soient déterminées localement et ne prennent pas en compte l'hétérogénéité tumorale. Au contraire, les données d'imagerie concernent la globalité de la lésion. Une dernière raison peut être le manque de certaines données (8 patients sur 90) pour cette caractéristique. Ces 11 caractéristiques sont ainsi les seules à être gardées pour la suite.

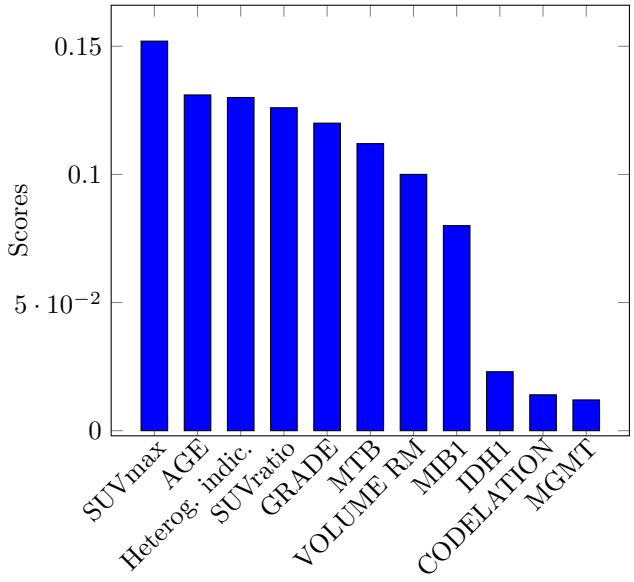



FIGURE 2.13 – Scores des plus importantes caractéristiques selon la classification des patients par leur PFS. Les caractéristiques non représentées ici ont un score plus petit que 0.01. Le score correspond au nombre de fois où la caractéristique est utilisée lors du classement par forêt aléatoire, normalisé par le nombre total d'utilisations de caractéristiques.

### 2.4.2 Réduction de l'espace des caractéristiques : méthode d'effets mixtes

L'espace des caractéristique doit encore être réduit, à cause du petit nombre de patients dans la base de données. De plus, certaines caractéristiques sont quantitatives (âge, volume tumoral, données PET scan) ou quantitatives (grade, données génomiques). Certains algorithmes d'apprentissage automatique, comme la méthode des plus proches voisins ne peuvent pas être utilisés sur un mélange de ces deux types de données. L'idée ici est de projeter les caractéristiques sur un espace de dimension plus petite, tout en gérant ce problème de mélange de type de données. L'analyse en composante principale permet de projeter les données sur un espace réduit, en ne sélectionnant que les directions principales. Cependant il n'est pas adapté aux données qualitatives. Nous allons donc utiliser une méthode d'effets mixtes [25], qui combine l'analyse en composante principale pour les données quantitatives, et l'analyse en correspondance multiple pour les données qualitatives. Notons $n$ le nombre de patients, $p_1$ le nombre de caractéristiques quantitatives, et $p_2$ le nombre de caractéristiques qualitatives. Notons $m$ le nombre total de modalités présentes dans les $p_2$ caractéristiques qualitatives, et $n_s$ le nombre de patients ayant la caractéristique $s$ ($s = 1...m$). Le tableau disjonctif complet est noté $X = (X_1|X_2)$ avec $X_1$ de taille $n$ x $p_1$ et $X_2$ de taille $n$ x $m$ : il est obtenu en remplaçant les variables

qualitatives à plusieurs modalités en autant de variables binaires, chacune correspondant à une des modalités. Les différentes étapes de la méthode d'effets mixtes sont les suivantes :

— Calcul de $Z_1$ la version centrée et réduite de $X_1$. Cela permet d'avoir les caractéristiques dans le même ordre de grandeur, et d'éviter qu'une variable à la variance élevée attire toute l'information.

— Calcul de $Z_2 = (I_n - \frac{J}{n})X_2D$, avec $J$ la matrice remplie de 1, et $D = \text{diag}(\sqrt{\frac{n}{n_s}})$, $s = 1...m$. La matrice $Z_2$ est donc une matrice indicatrice centrée (multiplication par $(I_n - \frac{J}{n})$) et avec les colonnes pondérées par la fréquence d'apparition de chaque modalité (multiplication par $D$).

— On pose $Z = \frac{1}{\sqrt{n}}(Z_1|Z_2)$.

— Décomposition en valeur singulière de $Z$ :

$$Z = U\Delta V^t, \tag{2.5}$$

avec $U^tU = V^tV = I_r$ où $r$ est le rang de $Z$, et $\Delta$ la matrice diagonale des racines carrées des valeurs propres, rangées par ordre décroissant.

— Les $k$ premières colonnes de $U$ donnent la matrice $n$ x $k$ qui correspond aux coordonnées de chaque patient dans l'espace projeté.

Ici, nous avons projeté les caractéristiques sur les 4 directions principales. La projection sur un espace de dimension plus grande est possible mais n'améliore pas la classification, en particulier parce que ces 4 directions détiennent plus de 70% de l'information. Les contributions des 11 caractéristiques pour les deux directions principales sont représentées sur la figure 2.14, et les projections de chaque patient sur l'espace à deux dimensions sont visibles sur la figure 2.15. Les algorithmes d'apprentissage automatique ont ensuite été appliqués à cette base de données de taille $n$ x 4.

Le choix dans notre cas d'effectuer l'analyse à effets mixtes après la sélection des caractéristiques s'explique par le fait que beaucoup de caractéristiques initiales étaient corrélées ou insignifiantes vis-à-vis du PFS et que ce premier élagage permet de réduire le bruit dans la méthode d'effets mixtes.

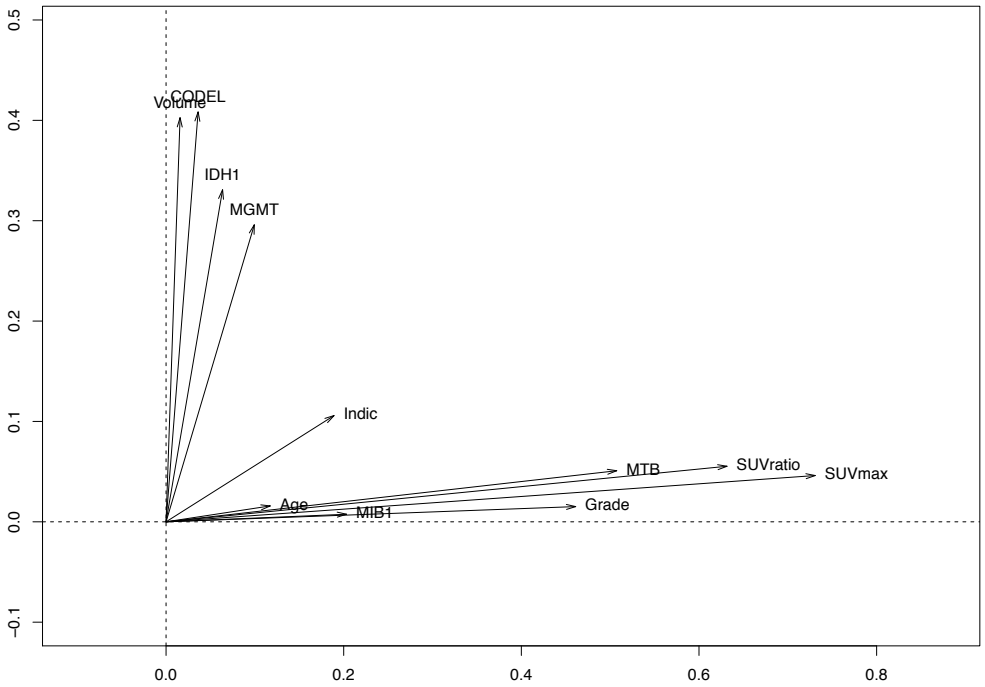



FIGURE 2.14 – Contribution de chacune des 11 caractéristiques sur les deux composantes principales : la première direction est représentée en abscisse et la seconde en ordonnée. Pour chaque caractéristique, la flèche tracée donne le pourcentage d'information de la caractéristique qui est utilisé dans chacune des deux directions principales.

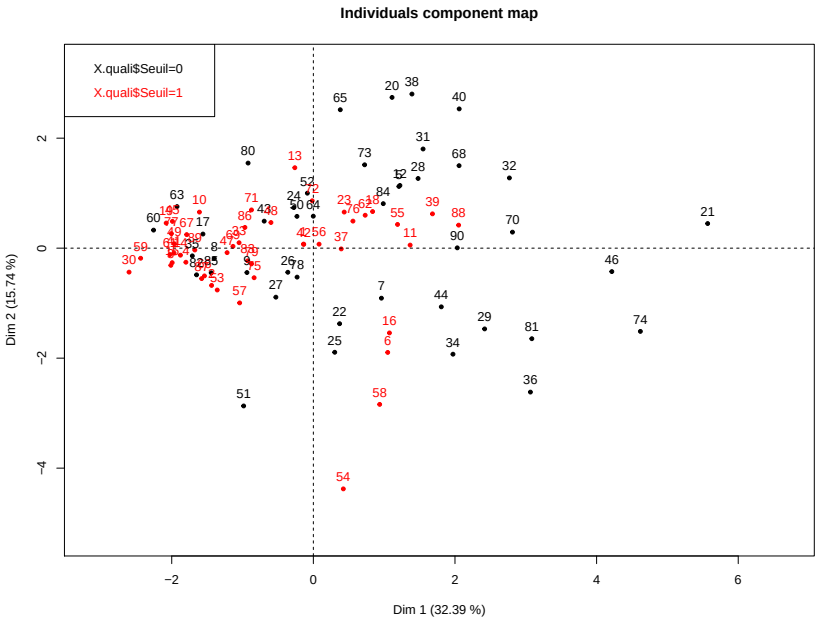



FIGURE 2.15 – Projection des caractéristiques des 90 patients sur les deux composantes principales. Les patients à rechute lente sont représentés en rouge, et ceux à rechute rapide en noir.

### 2.4.3 Comparaison des méthodes par validation croisée

La librairie Scikit Learn de Python [26] a été utilisée afin de comparer les algorithmes d'apprentissage automatisé présentés en introduction. L'efficacité d'une méthode a été calculée en utilisant une validation croisée en 10 sous-ensembles [27], comme décrit en préambule. Chaque algorithme est lancé pour 100 partitions en 10 sous-ensembles choisis aléatoirement. Les erreurs calculées sont donc les erreurs moyennes sur ces 100 prédictions, et la prédiction finale d'une méthode est la prédiction moyenne sur ces 100 prédictions.

## 2.5 Résultats de classification

### 2.5.1 Comparaison des algorithmes

Les méthodes précédemment détaillées sont appliquées au tableau de données 90 x 2 obtenu suite aux deux réductions successives de l'espace des caractéristiques. On obtient alors le tableau de la figure 2.2. **En moyenne, les algorithmes permettent une bonne classification des patients dans 77% des cas**. L'algorithme qui donne le meilleur score est la méthode des plus proches voisins avec 79% de bonne classification. Ce score est obtenu avec un nombre de voisins $k = 10$. Le fait que cet algorithme soit le plus efficace peut s'expliquer par le fait que les patients aux comportement identiques sont regroupés en plusieurs zones distinctes, ce qu'un classifieur linéaire ne peut pas capturer. Cependant, la comparaison d'un classifieur par rapport à un autre est anecdotique ici étant donné le peu de patients concernés. En effet, une différence de classification d'un seul patient entraîne un écart de 1.11%. L'obtention de résultats proches d'une méthode à l'autre suggère que le résultat global d'environ 77% est robuste. Les écarts-types de chaque méthode confirment cette stabilité. **Ce résultat est satisfaisant dans la mesure où seules les données avant chirurgie et les données génomiques ont été utilisées, sans information longitudinale**.

La méthode utilisée de validation croisée en $k$ parties permet d'obtenir une classification pour chaque patient de la base de données. Ce classement dans le cas de l'algorithme des plus proches voisins donne le tableau de la figure 2.3. On remarque que les 19 erreurs de classification sont équitablement réparties entre les rechutes lentes estimées rapides et les rechutes rapides estimées lentes. L'algorithme a ici minimisé le nombre d'erreurs totales. Nous aurions également pu tenter de minimiser l'une des deux catégories seulement. Or les cliniciens ont estimé que les deux erreurs étaient équivalentes. En effet, si une rechute rapide est estimée à tort, la dose de traitement peut être diminuée, alors qu'il aurait pu s'avérer efficace. A l'inverse, si une rechute lente est estimée à tort, le traitement est augmenté malgré sa non efficacité, et la qualité de vie du patient est dégradée sans résultats.

### 2.5.2 Courbe ROC

Le tracé de la courbe ROC [28] permet de vérifier l'efficacité de la classification. Lors de la classification d'un patient par une méthode d'apprentissage automatique, l'algorithme détermine un score $s$ entre 0 et 1, vu comme la probabilié d'être dans une classe ou dans

| Méthode | Erreur moyenne | Ecart type |
|---|---|---|
| KNN | 0.213 | 0.0084 |
| Naive Bayes | 0.237 | 0.0088 |
| Logistic Reg | 0.240 | 0.0059 |
| Random Forest | 0.267 | 0.0187 |
| ANN | 0.244 | 0.0104 |
| SVM | 0.233 | 0.0131 |

TABLE 2.2 – Six algorithmes sont comparés par validation croisée à 10 sous-ensembles. Les erreurs moyennes et les écarts-types sont calculés sur 100 partitions aléatoires.

| **Real PFS** / **PFS Prediction** | Fast recurrence | Slow recurrence |
|---|---|---|
| Fast recurrence | 33 | 9 |
| Slow recurrence | 10 | 38 |

TABLE 2.3 – Classification des patients en utilisant la méthode des plus proches voisins, soit celle qui minimise le nombre de mauvaises classifications.

l'autre. Si $s < 0.5$, le patient est classé dans la catégorie de rechute rapide, et si $s > 0.5$, le patient est classé dans la catégorie de rechute lente. L'idée de la courbe ROC est de faire varier ce seuil de 0.5 et de calculer l'évolution du taux de vrais positifs (taux des rechutes rapides estimées rapides) en fonction du taux de faux positifs (taux des rechutes lentes estimées rapides). Le taux de vrais positifs s'appelle la sensibilité, et le taux de faux positifs s'appelle l'antispécificité (1 moins la spécificité). Par exemple, lorsque le seuil vaut 0, tous les patients sont estimés à rechute lente. On a donc le tableau 2.4. La spécificité est donc égale à $\frac{0}{43} = 0$ et l'antispécificité à $\frac{0}{47} = 0$. Le point $(0, 0)$ est donc le premier point de la courbe.

A l'inverse, lorsque le seuil est de 1, tous les patients sont estimés en rechute rapide, ce qui donne le tableau de classification 2.5. Dans ce cas, la sensibilité vaut $\frac{43}{43} = 1$ et l'antispécificité vaut $\frac{47}{47} = 1$ donc le point $(1, 1)$ est le dernier point de la courbe. Dans le

| **Real PFS** / **PFS Prediction** | Fast recurrence | Slow recurrence |
|---|---|---|
| Fast recurrence | 0 | 0 |
| Slow recurrence | 43 | 47 |

TABLE 2.4 – Table de prédiction lorsque le seuil de décision est de 0.

| PFS Prediction \ Real PFS | Fast recurrence | Slow recurrence |
|---|---|---|
| Fast recurrence | 43 | 47 |
| Slow recurrence | 0 | 0 |

TABLE 2.5 – Table de prédiction lorsque le seuil de décision est de 1.

cas d'un classifieur aléatoire, la courbe ROC est une droite entre $(0, 0)$ et $(1, 1)$. Dans le cas d'un classifieur parfait, le taux de faux positifs est toujours nul, donc le seul autre point de la courbe est le point $(0, 1)$. Dans le cas de notre classifieur, le segment $[0, 1]$ est discrétisé, et pour chaque valeur du seuil appartenant à ce segment, on calcule la sensibilité et la spécificité de la classification. En particulier, grâce à la table 2.3, dans le cas où le seuil est égal à 0.5, le point correspondant est le point $(\frac{9}{47}, \frac{33}{43}) = (0.191, 0.767)$. Les courbes ROC associées aux classifications par les 6 algorithmes précédents sont représentées à la figure 2.16. La qualité d'un classifieur se mesure grâce à l'aire sous la courbe ROC (AUC). Dans le cas du classifieur aléatoire, elle vaut 0.5, et elle vaut 1 dans le cas du classifieur parfait. Notre classifieur possède une aire sous la courbe de 0.824, ce qui est au dessus de la limite de 0.8 qui sert de seuil pour distinguer les classifieurs efficaces. Plus généralement, on remarque que tous les classifieurs ont une AUC autour de 0.82, ce qui confirme la robustesse de la méthode. La classification qui est gardée pour les sections suivantes est celle obtenue par la méthode KNN, puisqu'elle maximise le nombre de bonnes classifications ainsi que l'AUC.

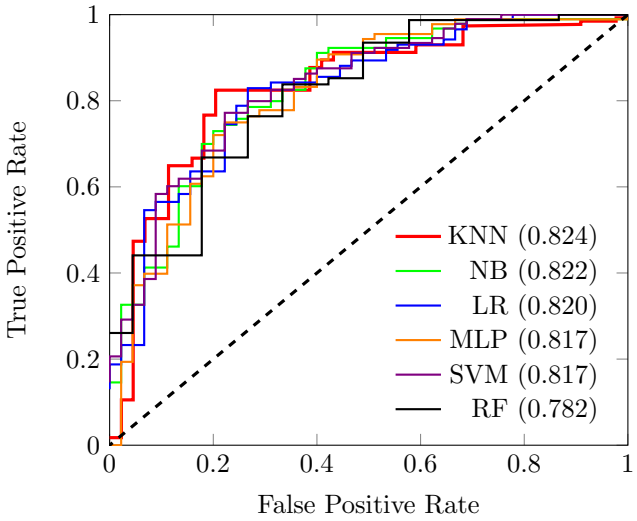



FIGURE 2.16 – Courbes ROC associées aux classifications par les 6 algorithmes. L'aire sous la courbe (AUC) est donnée entre parenthèses pour chaque méthode.

### 2.5.3 Courbes de Kaplan-Meier

#### Construction de la courbe

Afin de visualiser la classification des patients et de la comparer aux classifications par le grade ou le statut IDH1, les courbes de Kaplan-Meier [29] ont été tracées. Ces courbes de survie représentent l'évolution du taux de patients n'ayant pas encore rechuté, en fonction du temps. Notons $t_i$ ($i$ allant de 1 à N) l'ensemble des temps de rechute de la base de données, $d_i$ le nombre de rechutes à $t_i$, $c_i$ le nombre de données censurées à $t_i$ et enfin $n_i$ le nombre de patients à risque juste avant $t_i$. Notons que les patients aux données censurées (lorsque la rechute n'a pas encore eu lieu) ne font plus partie de l'ensemble des patients à risque lorsque le temps $t_i$ a dépassé le temps de censure. L'estimateur de Kaplan Meier a donc la forme suivante :

$$S(t) = \prod_{t_i<t} \left(1 - \frac{d_i}{n_i}\right), \tag{2.6}$$

avec, pour $i \in [\![0, N-1]\!]$

$$n_{i+1} = n_i - d_i - c_i. \tag{2.7}$$

Les rechutes entraînent une diminution de la courbe au temps $t_i$ et la fonction est constante entre deux dates de rechute. La date de rechute n'étant pas connue pour les données censurées, celles-ci sont représentées sur la courbe par des croix à la date du dernier examen de suivi. L'intérêt de cette courbe est de pouvoir comparer les comportements de deux populations différentes. Sur la figure 2.17, sont tracées en rouge les courbes de Kaplan-Meier des patients estimés en rechute rapide, et des patients en rechutes lentes. Ces courbes permettent ainsi de mieux visualiser la stratification des patients. En effet, une mauvaise classification d'un patient estimé à rechute lente mais qui a rechuté à 29 mois, est à différencier d'une mauvaise classification d'un patient estimé à rechute lente et qui a rechuté à 2 mois. Le tableau d'erreur ne permet pas distinguer ces deux erreurs de prédiction, mais les courbes de Kaplan-Meier le peuvent. Ces même courbes sont tracées en séparant les patients par leur grade et par leur statut IDH1. A noter que puisque l'information du statut IDH1 était manquante pour 8 des 90 patients, la courbe de Kaplan-Meier associée à la séparation par ce statut n'est faite que sur les 82 patients restants. On remarque que la classification par machine learning sépare nettement mieux les patients que les deux autres. **Cela signifie donc que l'information apportée par cette stratification est plus pertinente que celle que l'on a en regardant simplement le grade et le statut IDH1 du patient**. Afin de quantifier cet écart, deux tests ont été appliqués.

| | Groupe A | Groupe B | Total |
|---|---|---|---|
| Rechute | $r_{Ai}$ | $r_{Bi}$ | $r_i$ |
| Survie | $n_{Ai} - r_{Ai}$ | $n_{Bi} - r_{Bi}$ | $n_i - r_i$ |
| Total | $n_{Ai}$ | $n_{Bi}$ | $n_i$ |

TABLE 2.6 – Tableau des rechutes au temps $t_i$.

### Test du logrank

Notons $A$ le groupe de patients avec une rechute estimée rapide, et $B$ le groupe de patients avec une rechute estimée lente. Le but du test du logrank [30] est de montrer que le comportement des patients des groupes $A$ et $B$ est significativement différent.

Soit $H_0$ l'lhypothèse suivante :

**Hypothèse 1.** *Les temps de rechute des patients du groupe A et du groupe B suivent la même loi de probabilité.*

Supposons que cette hypothèse soit vérifiée. A chaque temps de rechute $t_i$ ($i$ allant de 0 à $N$), la proportion de rechute dans le groupe A et dans le groupe B est identique. Notons $n_i$ le nombre total de patients à risque (n'ayant pas encore rechuté) au temps $t_i$, et $n_{Ai}$ le nombre de tels patients dans le groupe A, et $n_{Bi}$ le nombre de tels patients dans le groupe B. De la même manière, soient $r_{Ai}$ et $r_{Bi}$ le nombre de rechutes au temps $t_i$ dans les groupes $A$ et $B$ respectivement, et $r_i = r_{Ai} + r_{Bi}$. On a donc le tableau 2.6 représentatif de l'état à $t_i$.

Soit $e_i$ le nombre de rechutes attendues au temps $t_i$, sous l'hypothèse $H_0$. On obtient :

$$e_i = e_{Ai} + e_{Bi} = \frac{n_{Ai} \cdot r_i}{n_i} + \frac{n_{Bi} \cdot r_i}{n_i}. \tag{2.8}$$

Soient $E_A$ et $E_B$ le nombre total de rechutes attendues, et $O_A$ et $O_B$ le nombre total de

rechutes observées, dans les groupes $A$ et $B$ respectivement, alors :

$$\begin{aligned} E_A &= \sum_{i=0}^{N} e_{Ai} = \sum_{i=0}^{N} \frac{n_{Ai} \cdot r_i}{n_i}, \\ E_B &= \sum_{i=0}^{N} e_{Bi} = \sum_{i=0}^{N} \frac{n_{Bi} \cdot r_i}{n_i}, \\ O_A &= \sum_{i=0}^{N} r_{Ai}, \\ O_B &= \sum_{i=0}^{N} r_{Bi}. \end{aligned} \tag{2.9}$$

Sous l'hypothèse $H_0$, les variables $O_A$ et $O_B$ suivent des lois normales d'espérance $E_A$ et $E_B$ et de variance $\sqrt{E_A}$ et $\sqrt{E_B}$ respectivement. La variable définie par :

$$X = \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B}, \tag{2.10}$$

suit donc une loi du $\chi^2$ à un degré de liberté $X \sim \chi^2(1)$.
Le calcul de cette valeur dans le cas de la classification par machine learning, grade et statut IDH1 donne les résultats suivants :

$$X_{\text{ML}} = 39.9,\ X_{\text{grade}} = 11.6 \text{ et } X_{\text{IDH1}} = 7.8. \tag{2.11}$$

On en déduit les probabilités que l'hypothèse $H_0$ soit vraie dans chaque cas :

$$p_{\text{ML}} < 0.00001,\ p_{\text{grade}} = 0.00066 \text{ et } p_{\text{IDH1}} = 0.0052. \tag{2.12}$$

Dans chacun des cas, la probabilité est très faible, ce qui permet de rejeter l'hypothèse $H_0$ et ainsi conclure que la différence entre les groupes $A$ et $B$ n'est pas simplement liée aux bruits des données, mais est issue d'une vraie différence de comportement. Cela permet également de quantifier ce qui était visible sur les courbes de Kaplan-Meier, à savoir que notre méthode permet une meilleure séparation des patients que le grade et le statut IDH1.

### Calcul du risque de rechute par le modèle de Cox

Un autre manière de quantifier cet écart est de calculer le risque de rechute instantanée en fonction du temps, noté $\lambda(t)$, pour un patient qui n'a pas encore rechuté juste avant le

temps $t$. Le modèle de Cox [31] consiste à exprimer cette fonction sous la forme :

$$\lambda(t, X) = \lambda_0(t) \exp(\beta X), \tag{2.13}$$

où $X$ correspond au groupe de patients considéré (0 pour les patients estimés à rechute rapide, 1 pour les patients estimés à rechute lente), et $\lambda_0(t)$ est une fonction du temps mais indépendante de $X$. L'idée ici n'est pas d'estimer $\lambda_0(t)$ mais d'estimer le ratio de risque HR ("hazard ratio"), entre les deux groupes, à savoir :

$$\mathrm{HR} = \frac{\lambda(t, 0)}{\lambda(t, 1)} = \frac{\lambda_0(t)}{\lambda_0(t) \exp(\beta)} = \exp(-\beta). \tag{2.14}$$

HR mesure donc le risque relatif de rechute, entre les patients du groupe $A$ et ceux du groupe $B$. L'hypothèse des risques proportionnels du modèle de Cox donne au ratio de risque HR une valeur constante en fonction du temps. La méthode du maximum de vraisemblance permet d'estimer $\beta$. En notant $n$ le nombre total de patients :

$$\beta = \operatorname{argmin} \prod_{i=1}^{n} \frac{\exp(\beta X_i)}{\sum_{Y_j \geqslant Y_i} \exp(\beta X_j)},$$

où $X_i$ et $Y_i$ sont respectivement le groupe et le temps de rechute du patient $i$. On obtient alors, pour chaque classifieur :

$$\mathrm{HR}_{\mathrm{ML}} = 5.59,\ \mathrm{HR}_{\mathrm{grade}} = 2.59 \text{ et } \mathrm{HR}_{\mathrm{IDH1}} = 2.04. \tag{2.15}$$

Dans le cas de notre classifieur, le risque de rechute à un instant $t$ pour les patients du groupe $A$ est donc plus de 5 fois plus grand que le risque de rechute des patients du groupe $B$. Le ratio est plus de deux fois plus faible dans le cas de la classification par grade et par statut IDH1. Cela permet une fois de plus de montrer l'intérêt de notre classifieur, et de quantifier l'avantage qu'il apporte.

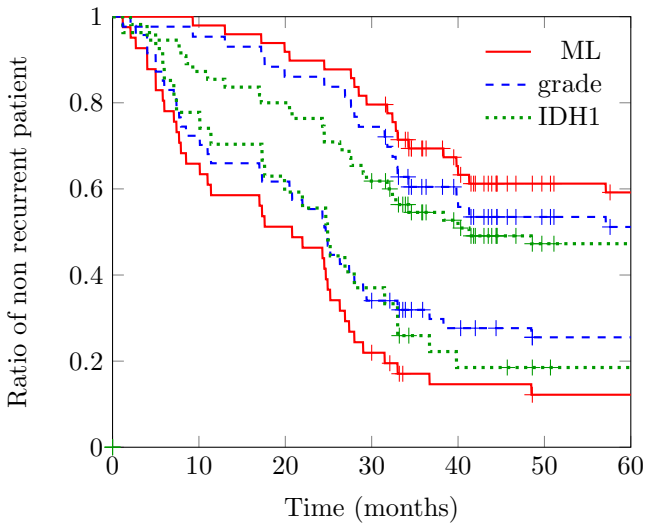



FIGURE 2.17 – Courbes de Kaplan-Meier associées aux classifications obtenues à partir de notre méthode (rouge), du grade (bleu) et du statut IDH1(vert). Les croix représentent les données censurées, c'est-à-dire les patients n'ayant pas encore rechuté au moment de l'étude.

### Comparaison avec la combinaison grade/IDH1

La section précédente a montré que le grade seul ou le statut IDH1 seul ne permettent pas de stratifier aussi bien les patients que notre classifieur. La combinaison de ces deux informations a été testée et comparée à notre classification. Sur la figure 2.18, les courbes de Kaplan-Meier associées aux quatre classes combinées (grade II muté et sauvage, grade II muté et sauvage) sont représentées. Les courbes des cas extrêmes (grade II muté et grade III sauvage) sont confondues avec la séparation par machine learning, alors que les courbes intermédiaires (grade III muté et grade II sauvage) sont mal séparées. Cela se confirme sur les graphiques de la figure 2.19. A gauche, les courbes de Kaplan-Meier ne sont tracées que pour les patients de grade II muté et de grade III sauvage. Ces courbes sont exactement les mêmes que pour notre classifieur. **Cela signifie que la classification est identique et que dans ces cas extrêmes, le grade et le statut IDH1 contiennent toutes les informations nécessaires**. La séparation est forte et cela paraît cohérent puisque l'on compare ici des tumeurs agressives qui répondent mal au traitement (grade III sauvage) à des tumeurs moins agressives qui répondent bien au traitement (grade II muté). L'imagerie IRM et PET scan n'apporte donc pas plus d'information pour ce type de patients.

Par contre, la figure de droite montre que pour les autres cohortes de patients, à savoir les

tumeurs de grade III mutées et de grade II sauvages, la méthode basée sur l'apprentissage automatique sépare beaucoup mieux les patients. **Cela signifie donc que pour ces patients, la combinaison des données d'imagerie et de génomique est utile pour diviser la cohorte en deux classes au comportement bien distinct**.

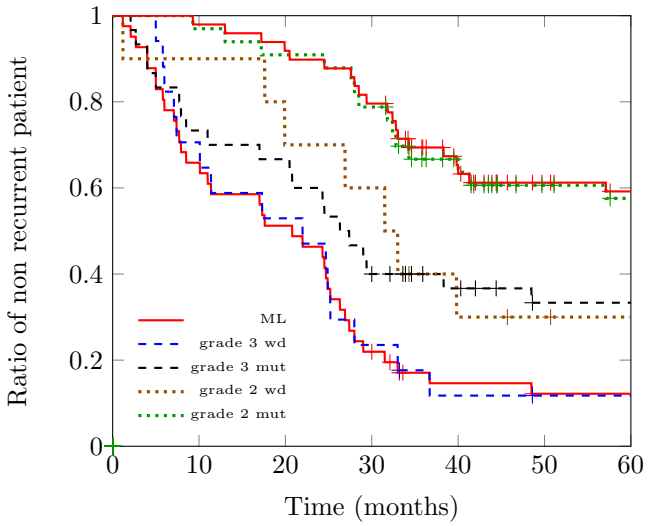



FIGURE 2.18 – Courbes de Kaplan-Meier associées aux classifications par notre méthode (rouge), et par la combinaison du grade et du statut IDH1. Les données censurées sont représentées par des croix.

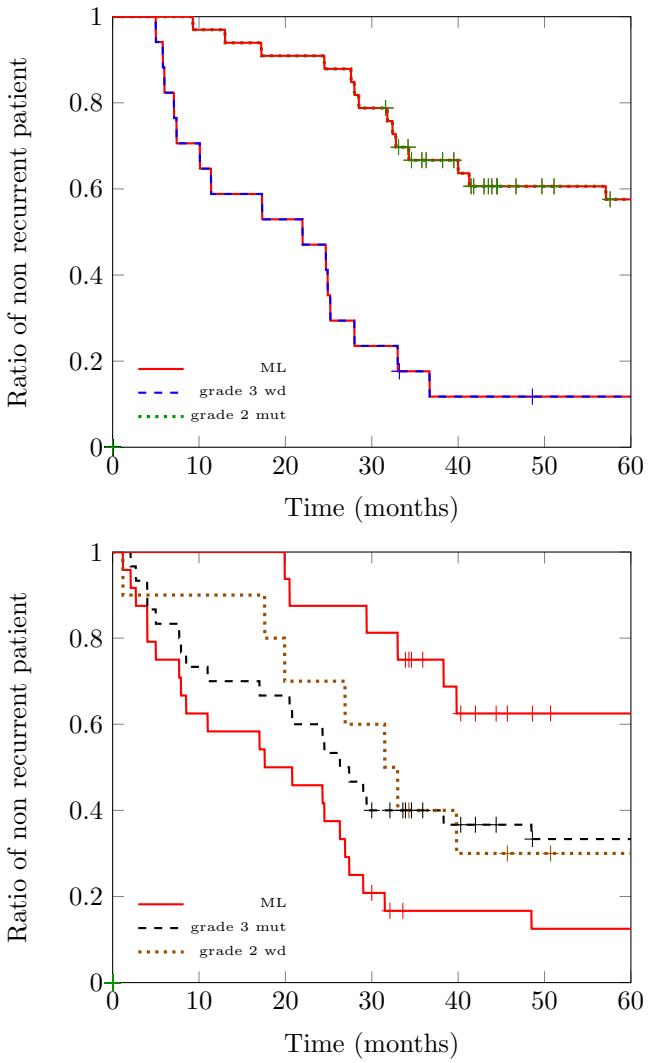



FIGURE 2.19 – Courbes de Kaplan-Meier associées aux classifications par notre méthode (rouge), et par la combinaison du grade et du statut IDH1. Première ligne : seulement pour les patients de grade II muté IDH1 et de grade III sauvage IDH1. Deuxième ligne : seulement pour les patients de grade II sauvage IDH1 et de grade III muté IDH1.

### Restriction aux données d'imagerie seulement

Dans les sections précédentes, les données utilisées étaient les données démographiques et d'imagerie, acquises avant la chirurgie, ainsi que les données de génomique acquises sur la pièce opératoire. Cela a permis de constater l'efficacité de la combinaison des informations de génomique et d'imagerie dans la prédiction du PFS. L'objectif de cette section est de savoir si la prédiction pouvait être satisfaisante en enlevant les données de génomique. Cela revient en particulier à enlever les informations de grade et de statut IDH1

qui sont actuellement utilisées pour le pronostic clinique. La même méthode a donc été appliquée à la base de données résultante, et donne 70% de bonne classification. La courbe de Kaplan-Meier correspondante est représentée à la figure 2.20. Même si le résultat de classification est moins bon qu'auparavant, on remarque que l'imagerie seule donne autant, voire même plus d'informations relatives à la durée sans rechute que le grade et le statut IDH1. **Grâce à cette stratification, le clinicien peut avoir accès à l'agressivité de la tumeur seulement à partir de l'imagerie (IRM et MET-PET), et pas à partir d'une biopsie du cerveau qui est une méthode invasive et non sans risques**. En particulier, cela peut aider à la plannification d'une éventuelle chirurgie pour les grades faibles.

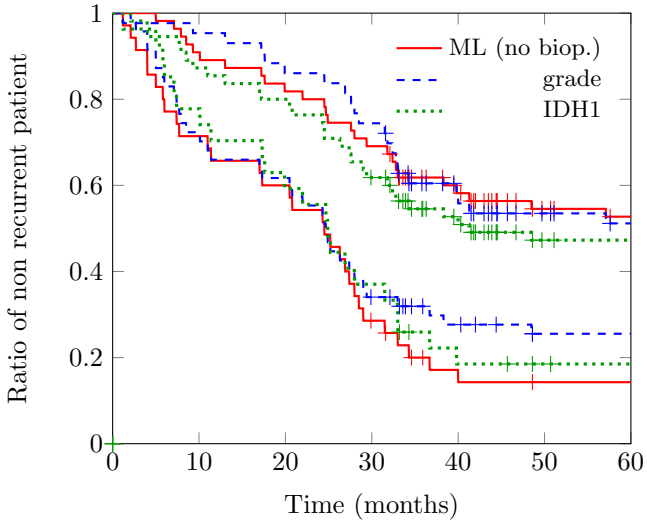



FIGURE 2.20 – Courbes de Kaplan-Meier associées aux classifications obtenues à partir de notre méthode (rouge), du grade (bleu) et du statut IDH1(vert). Pour notre méthode, seules les données acquises avant chirurgie sont utilisées ici (imagerie et démographie), et donc aucune information issue de la biopsie.

## 2.6 Discussion et conclusion

**Cette étude rétrospective montre que la combinaison des informations acquises par le neuro-chirurgien permet d'améliorer la prédiction du temps de rechute**. La méthode globale est représentée sur la figure 2.21. L'étude confirme l'intérêt des PET scans dans l'étude du pronostic clinique. En effet, les données issues de ce type d'imagerie métabolique apparaissent comme étant celles les plus fortement corrélées au temps de survie sans progression et apportent des informations en plus du grade et du statut IDH1. **De plus, l'indicateur d'hétérogénéité calculé semble avoir un réel intérêt puisqu'il ressort dans l'analyse de pré-traitement comme étant influent**.**La comparaison de la classification par notre méthode aux classifications actuellement utilisées par le médecin montre que de tels outils mathématiques peuvent permettre de mieux anticiper les comportements de certains patients atypiques**. Les patients se voient administrer un traitement dépendant de leur condition physique et de l'agressivité de la tumeur. Avoir une idée de la réaction du patient à ce traitement avant même que le traitement soit administré donne un avantage au médecin sur la maladie, lui permettant d'adapter ce traitement en fonction de la prédiction. De plus, la dernière partie de l'étude utilisant seulement les données avant chirurgie montre que le clinicien peut avoir une idée sur l'agressivité de la tumeur sans recourrir à des méthodes invasives (biopsie ou chirurgie), mais simplement à partir de l'imagerie.

**Bien sûr, les résultats ici demandent à être confirmés sur une base de données plus grande. Cependant, cette étude est une preuve de concept qui permet d'être optimiste sur l'intérêt de l'apprentissage automatique en oncologie**. De plus, le MET-PET scan est une technologie peu utilisée dans le monde, en partie en raison de son coût. On notera cependant qu'une telle étude peut très bien s'appliquer avec des données issues d'autres techniques d'imagerie (spectroscopie, imagerie de diffusion, etc...). Le travail effectué sur le PFS a également été testé sur le temps de survie global. Cependant, ce temps de survie n'est connu que pour 24 patients sur les 90 puisque la durée de suivi clinique n'est pas suffisamment longue. En conséquence, la base de données était trop petite pour pouvoir obtenir de bons résultats par les algorithmes d'apprentissage automatique.

Lorsque plus de données sont collectées, il est possible d'utiliser des méthodes d'apprentissage profond (Deep Learning). L'avantage de ce type d'algorithmes est de pouvoir combiner tout seul les caractéristiques afin d'en former de nouvelles qui sont pertinentes pour la clas-

sification. Dans notre cas d'étude, de telles méthodes ne peuvent pas s'appliquer car la base de données est trop petite. C'est donc l'ajout de connaissance qui permet de trouver de nouvelles caractéristiques, comme ça a été le cas pour l'indicateur d'hétérogénéité.

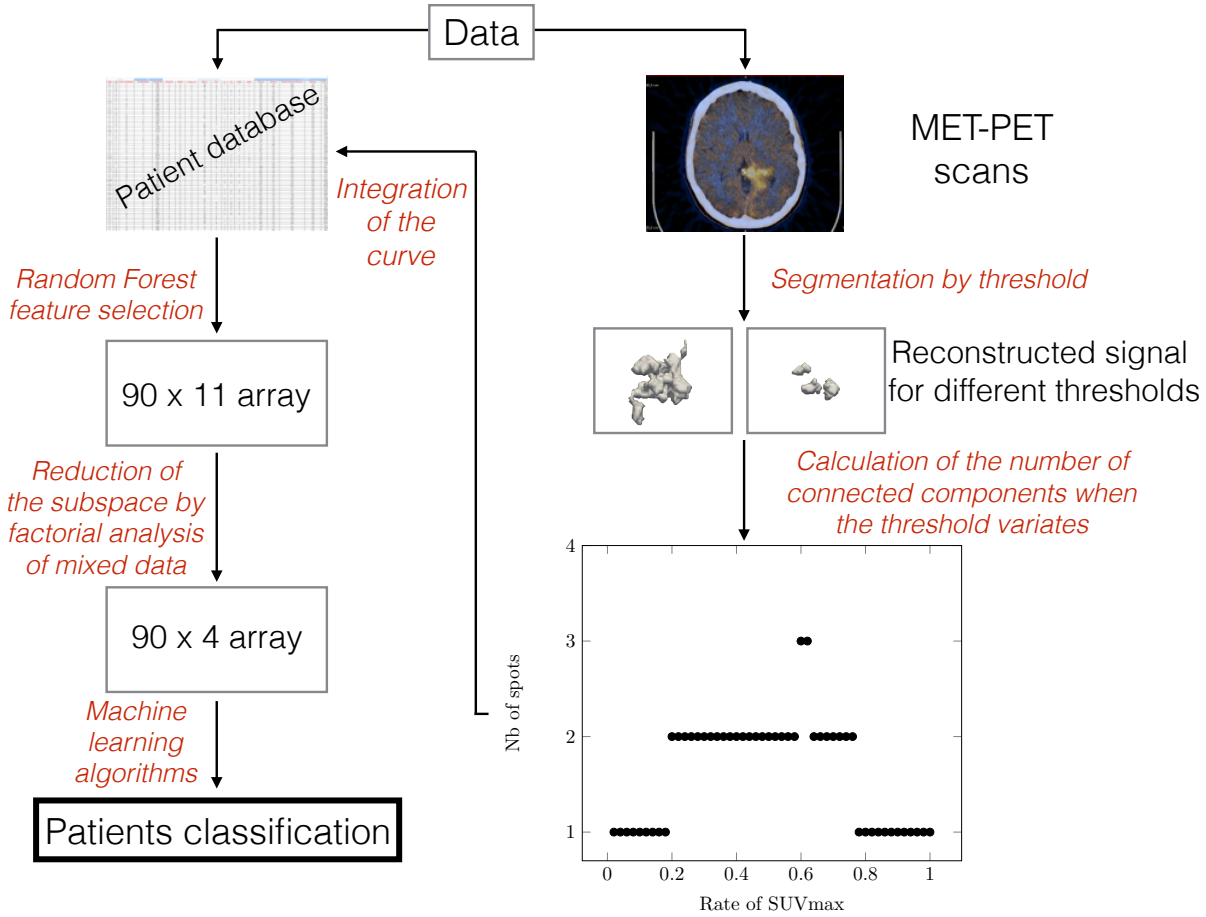



FIGURE 2.21 – Description de la méthode globale. La partie droite explique l'extraction de l'indicateur d'hétérogénéité à partir des PET scans. La partie gauche montre les différentes étapes pour passer de la base de données à la classification : sélection des caractéristiques, analyse factorielle de données mixtes et algorithmes d'apprentissage automatique.

# Bibliographie

## Bibliographie


[1] David N. Louis, Hiroko Ohgaki, Otmar D. Wiestler, Webster K. Cavenee, Peter C. Burger, Anne Jouvet, Bernd W. Scheithauer, and Paul Kleihues. The 2007 WHO Classification of Tumours of the Central Nervous System. *Acta Neuropathol*, 114(2) :97–109, August 2007.

[2] Quinn T. Ostrom, Haley Gittleman, Peter Liao, Chaturia Rouse, Yanwen Chen, Jacqueline Dowling, Yingli Wolinsky, Carol Kruchko, and Jill Barnholtz-Sloan. CBTRUS Statistical Report : Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2007–2011. *Neuro Oncol*, 16(Suppl 4) :iv1–iv63, October 2014.

[3] Asgeir S. Jakola, Kristin S. Myrmel, Roar Kloster, Sverre H. Torp, Sigurd Lindal, Geirmund Unsgård, and Ole Solheim. Comparison of a Strategy Favoring Early Surgical Resection vs a Strategy Favoring Watchful Waiting in Low-Grade Gliomas. *JAMA*, 308(18) :1881–1888, November 2012.

[4] Michael Weller, Martin van den Bent, Jörg C. Tonn, Roger Stupp, Matthias Preusser, Elizabeth Cohen-Jonathan-Moyal, Roger Henriksson, Emilie Le Rhun, Carmen Balana, Olivier Chinot, Martin Bendszus, Jaap C. Reijneveld, Frederick Dhermain, Pim French, Christine Marosi, Colin Watts, Ingela Oberg, Geoffrey Pilkington, Brigitta G. Baumert, Martin J. B. Taphoorn, Monika Hegi, Manfred Westphal, Guido Reifenberger, Riccardo Soffietti, Wolfgang Wick, and European Association for Neuro-Oncology (EANO) Task Force on Gliomas. European Association for Neuro-Oncology (EANO) guideline on the diagnosis and treatment of adult astrocytic and oligodendroglial gliomas. *Lancet Oncol.*, 18(6) :e315–e329, 2017.

[5] Catherine Terret, Gilles Albrand, Géraldine Moncenix, and Jean Pierre Droz. Karnofsky Performance Scale (KPS) or Physical Performance Test (PPT) ? That is the question. *Crit. Rev. Oncol. Hematol.*, 77(2) :142–147, February 2011.

[6] Bogdana Suchorska, Nathalie Lisa Albert, and Jörg-Christian Tonn. Usefulness of PET Imaging to Guide Treatment Options in Gliomas. *Curr Treat Options Neurol*, 18(1) :4, January 2016.

[7] Nathalie L. Albert, Michael Weller, Bogdana Suchorska, Norbert Galldiks, Riccardo Soffietti, Michelle M. Kim, Christian la Fougère, Whitney Pope, Ian Law, Javier Arbizu, Marc C. Chamberlain, Michael Vogelbaum, Ben M. Ellingson, and Joerg C. Tonn. Response Assessment in Neuro-Oncology working group and European Association for Neuro-Oncology recommendations for the clinical use of PET imaging in gliomas. *Neuro-oncology*, 18(9) :1199–1208, 2016.

[8] Zeju Li, Yuanyuan Wang, Jinhua Yu, Yi Guo, and Wei Cao. Deep learning based radiomics (dlr) and its usage in noninvasive idh1 prediction for low grade glioma. *Scientific Reports*, 7(1) :5467, 2017.

[9] C. Bourgier, J. Colinge, N. Ailléres, P. Fenoglietto, M. Brengues, A. Pélegrin, and D. Azria. Définition et applications cliniques des radiomics. *Cancer/Radiothérapie*, 19(6) :532 – 537, 2015. 26e Congrès national de la Société française de radiothérapie oncologique SFRO).

[10] David N. Louis, Arie Perry, Guido Reifenberger, Andreas von Deimling, Dominique Figarella-Branger, Webster K. Cavenee, Hiroko Ohgaki, Otmar D. Wiestler, Paul Kleihues, and David W. Ellison. The 2016 World Health Organization Classification of Tumors of the Central Nervous System : a summary. *Acta Neuropathol*, 131(6) :803–820, June 2016.

[11] Egesta Lopci, Marco Riva, Laura Olivari, Fabio Raneri, Riccardo Soffietti, Arnoldo Piccardo, Alberto Bizzi, Pierina Navarria, Anna Maria Ascolese, Roberta Rudà, Bethania Fernandes, Federico Pessina, Marco Grimaldi, Matteo Simonelli, Marco Rossi, Tommaso Alfieri, Paolo Andrea Zucali, Marta Scorsetti, Lorenzo Bello, and Arturo Chiti. Prognostic value of molecular and imaging biomarkers in patients with supratentorial glioma. *Eur. J. Nucl. Med. Mol. Imaging*, 44(7) :1155–1164, July 2017.

[12] Bogdana Suchorska, Nathalie L. Jansen, Jennifer Linn, Hans Kretzschmar, Hendrik Janssen, Sabina Eigenbrod, Matthias Simon, Gabriele Pöpperl, Friedrich W. Kreth, Christian la Fougere, Michael Weller, and Joerg C. Tonn. Biological tumor volume in 18fet-pet before radiochemotherapy correlates with survival in gbm. *Neurology*, 84(7) :710–719, 2015.

[13] Joseph A. Cruz and David S. Wishart. Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Inform*, 2 :59–77, February 2007.

[14] Laszlo Papp, Nina Poetsch, Marko Grahovac, Victor Schmidbauer, Adelheid Woehrer, Matthias Preusser, Markus Mitterhauser, Barbara Kiesel, Wolfgang Wadsak, Thomas Beyer, Marcus Hacker, and Tatjana Traub-Weidinger. Glioma survival prediction with the combined analysis of in vivo 11c-MET-PET, ex vivo and patient features by supervised machine learning. *J Nucl Med*, page jnumed.117.202267, November 2017.

[15] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13 :8–17, 2015.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE : Synthetic Minority Over-sampling Technique. *arXiv :1106.1813 [cs]*, June 2011. arXiv : 1106.1813.

[17] S. B. Kotsiantis. Supervised Machine Learning : A Review of Classification Techniques. In *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering : Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, pages 3–24, Amsterdam, The Netherlands, The Netherlands, 2007. IOS Press.

[18] Li-Yu Hu, Min-Wei Huang, Shih-Wen Ke, and Chih-Fong Tsai. The distance function effect on k-nearest neighbor classification for medical datasets. *Springerplus*, 5(1), August 2016.

[19] Sandro Sperandei. Understanding logistic regression analysis. *Biochem Med (Zagreb)*, 24(1) :12–18, February 2014.

[20] Andor W. J. M. Glaudemans, Roelien H. Enting, Mart A. A. M. Heesters, Rudi A. J. O. Dierckx, Ronald W. J. van Rheenen, Annemiek M. E. Walenkamp, and Riemer H. J. A. Slart. Value of 11c-methionine pet in imaging brain tumours and metastases. *European Journal of Nuclear Medicine and Molecular Imaging*, 40(4) :615–635, Apr 2013.

[21] Kamalakannan Palanichamy and Arnab Chakravarti. Diagnostic and Prognostic Significance of Methionine Uptake and Methionine Positron Emission Tomography Imaging in Gliomas. *Front Oncol*, 7 :257, 2017.

[22] D. L. Collins, A. P. Zijdenbos, V. Kollokian, J. G. Sled, N. J. Kabani, C. J. Holmes, and A. C. Evans. Design and construction of a realistic digital brain phantom. *IEEE Transactions on Medical Imaging*, 17(3) :463–468, June 1998.

[23] Tae Min Kim, Jin Chul Paeng, In Kook Chun, Bhumsuk Keam, Yoon Kyung Jeon, Se-Hoon Lee, Dong-Wan Kim, Dong Soo Lee, Chul Woo Kim, June-Key Chung, Il Han Kim, and Dae Seog Heo. Total lesion glycolysis in positron emission tomography is a better predictor of outcome than the International Prognostic Index for patients with diffuse large B cell lymphoma. *Cancer*, 119(6) :1195–1202, March 2013.

[24] D. L. Collins, A. P. Zijdenbos, V. Kollokian, J. G. Sled, N. J. Kabani, C. J. Holmes, and A. C. Evans. Design and construction of a realistic digital brain phantom. *IEEE Transactions on Medical Imaging*, 17(3) :463–468, June 1998.

[25] Marie Chavent, Vanessa Kuentz-Simonet, Amaury Labenne, and Jérôme Saracco. Multivariate analysis of mixed data : The PCAmixdata R package. *arXiv :1411.4911 [stat]*, November 2014. arXiv : 1411.4911.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011.

[27] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, pages 1137–1143, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[28] Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian J Intern Med*, 4(2) :627–635, 2013.

[29] Manish Kumar Goel, Pardeep Khanna, and Jugal Kishore. Understanding survival analysis : Kaplan-Meier estimate. *Int J Ayurveda Res*, 1(4) :274–278, 2010.

[30] J Martin Bland and Douglas G Altman. The logrank test. *BMJ*, 328(7447) :1073, May 2004.

[31] Spotswood L. Spruance, Julia E. Reid, Michael Grace, and Matthew Samore. Hazard Ratio in Clinical Trials. *Antimicrob Agents Chemother*, 48(8) :2787–2792, August 2004.

# Chapitre 3

# Assimilation de données pour la croissance de tumeurs

**Sommaire**


**3.1 Introduction . . . . . 87**
**3.2 Modèle de croissance de tumeurs . . . . . 90**
3.2.1 Modèle . . . . . 90
3.2.2 Résolution numérique . . . . . 91
3.2.3 Premières simulations . . . . . 94
3.2.4 Calcul du volume tumoral . . . . . 94
**3.3 Estimation des paramètres par calibrage volumique . . . . . 98**
3.3.1 Principe du calibrage volumique . . . . . 98
3.3.2 Espace de recherche des paramètres $\alpha$ et $m_0$ . . . . . 99
3.3.3 Exemple sur un cas synthétique . . . . . 100
3.3.4 Exemple sur un cas clinique 2D : métastase au poumon . . . . . 105
3.3.5 Exemple sur un cas clinique 3D : métastase cérébrale . . . . . 107
3.3.6 Insuffisance du calibrage volumique . . . . . 107
**3.4 Assimilation de données . . . . . 112**
3.4.1 Principe de l'assimilation de données . . . . . 112
3.4.2 Approche variationnelle . . . . . 113
3.4.3 Approche séquentielle . . . . . 114
3.4.4 Mesure de similarité . . . . . 115
3.4.5 Observateur de Luenberger . . . . . 116
3.4.6 Justifications mathématiques . . . . . 119
3.4.7 Filtre de Kalman réduit à l'espace des paramètres . . . . . 125
**3.5 Résolution numérique . . . . . 133**

3.5.1 Discrétisation du Dirac . . . . . 133
3.5.2 Interpolation des données . . . . . 134
3.5.3 Choix des *sigma points* . . . . . 136
3.5.4 Résolution de l'équation sur $M$ . . . . . 138
**3.6 Validation de la méthode sur données synthétiques . . . . . 139**
3.6.1 Correction d'état . . . . . 139
3.6.2 Correction jointe état-paramètres . . . . . 142
3.6.3 Correction jointe état-paramètres avec condition initiale décalée . 150
3.6.4 Importance du choix de $\lambda$ . . . . . 152
3.6.5 Estimation de $M_0$ sur une grille . . . . . 153
**3.7 Application aux données réelles . . . . . 162**
**3.8 Conclusion . . . . . 170**

## 3.1 Introduction

La croissance tumorale dépend de nombreux mécanismes biologiques, ce qui la rend difficile à prévoir. Cependant, prédire la forme tumorale à un temps futur est une information cruciale pour le clinicien, qui peut alors avoir un temps d'avance sur la maladie pour anticiper la planification d'un traitement. En particulier, pour certains organes tels que le cerveau, les zones atteintes par la tumeur peuvent être comparées à une cartographie des fonctionnalités du cerveau afin de juger de l'urgence de la situation et de la possibilité d'appliquer un traitement. La construction d'un modèle de croissance tumorale nécessite de représenter les interactions entre plusieurs entités biologiques (cellules tumorales, cellules saines, vaisseaux sanguins, système immunitaire, etc...). Il est nécessaire de comprendre les mécanismes biologiques qui sont en jeu afin de modéliser ces interactions. Cependant, il est impossible de tous les prendre en compte, et l'on doit ainsi procéder à des simplifications biologiques. Dans le cas d'un modèle de croissance tumorale, les principes pris en considération sont la division cellulaire, le phénomène d'angiogénèse ou de consommation de nutriments. Les interactions extérieures sont plus complexes à modéliser puisqu'elles sont spécifiques à l'environnement de la tumeur, et donc fortement dépendantes du patient. Elles sont donc souvent négligées dans le processus de modélisation. Ces interactions sont pourtant essentielles dans l'évolution de la forme de la tumeur. En effet, les contraintes mécaniques et biologiques sont les principales responsables du changement de morphologie de la tumeur. La segmentation de l'environnement tumoral peut permettre d'inclure ces données au modèle [1]. Cependant, la qualité d'imagerie rend difficile les segmentations d'entités biologiques fines tels les vaisseaux ou les fibres. Dans notre modèle, un champ de vascularisation rend compte de l'environnement extérieur. La stratégie ici est donc de corriger et d'estimer ce champ au temps initial sans *a priori* sur sa structure. Cette correction qui a pour but de rapprocher le modèle des observations se fait par assimilation de données.

L'assimilation de données [2] consiste à combiner un modèle et des données afin d'estimer le système réel. Elle permet entre autres de résoudre des problèmes de suivi de front, lorsque l'interface est mobile. Parmi les applications possibles, on peut citer la propagation de feux [3, 4, 5], la prédiction météorologique [6, 7], l'électrophysiologie cardiaque [8, 9, 10], les marées noires [11], et donc la croissance tumorale [12].

Le modèle utilisé ici est un modèle mécanistique à une seule population de cellules cancéreuses, qui prolifèrent et se déplacent à cause de la pression induite par cette prolifération. Ce type de modèle est bien évidemment différent des modèles d'apprentissage utilisés dans

la partie précédente : on ne considère plus une population d'individus mais chaque patient est considéré individuellement et une séquence de plusieurs examens de ce patient sont utilisées. Ces données consistent donc en une succession de contours de la tumeur du patient, collectés à plusieurs temps différents.

Le modèle est considéré comme une bonne approximation potentielle de la trajectoire réelle, mais dépend de paramètres, de condition initiale, et de conditions aux bords incertains. Les données procurent des informations sur la trajectoire réelle, mais elles sont partielles (en temps et en espace) et contiennent souvent une part non régligeable d'erreur due aux mesures. L'assimilation de donnée fonctionne en combinant ces deux sources d'informations et en tenant compte des incertitudes, afin d'estimer le système réel. Cette correction nécessite de pouvoir comparer le contour simulé au contour observé. Il est donc nécessaire de définir une mesure de similarité entre observation et simulation. Corriger le modèle signifie donc ici réduire cette mesure de discordance entre simulation et observation.

**L'objectif de cette partie est donc de construire un modèle de croissance tumorale afin de prédire le volume, mais également la forme qu'aura la tumeur à un temps donné**. Dans un premier temps, nous introduirons un modèle simple de croissance. L'avantage de ce modèle est de comporter peu de paramètres, mais aussi d'être intégrable. L'évolution du volume tumoral est alors connu explicitement en fonction de ces paramètres. Il est alors possible de calibrer les paramètres et ainsi de prédire l'évolution du volume tumoral à partir des premiers examens. La tumeur simulée a alors un volume proche de la tumeur observée, mais la forme prédite peut cependant être très éloignée de l'observation.

Dans un second temps, nous détaillerons donc la stratégie d'**assimilation de données** qui permet de corriger le modèle et la forme tumorale. Cette correction se fera conjointement de deux manières. Tout d'abord, les incertitudes liées à la segmentation de la condition initiale [13] ainsi qu'aux simplifications de modélisation sont corrigées par un filtre d'état. De plus, les paramètres du modèle sont eux aussi corrigés, afin de rendre compte de l'environnement de la tumeur. Un modèle de correction jointe état-paramètres est alors développé, comme initié dans [14]. **Nous utiliserons un filtre de Luenberger pour corriger l'état du système [8], ainsi qu'un filtre réduit de Kalman [15] pour corriger les paramètres**. **Nous prouverons également que sous des hypothèses raisonnables, le filtre d'état permet bien la convergence théorique vers la cible. Des données synthétiques sont ensuite utilisées afin de valider la méthode globale utilisée**.

Enfin, cette stratégie sera appliquée sur des **données réelles**. **La prédiction obtenue est alors bien meilleure que celle obtenue sans correction**. **Nous verrons également que cela permet de reconstituer et d'estimer la matrice extra-cellulaire qui entoure la tumeur**.

## 3.2 Modèle de croissance de tumeurs

### 3.2.1 Modèle

Nous souhaitons établir un modèle régissant la croissance tumorale. Deux phénotypes seront distingués ici : un phénotype tumoral, représenté par la densité de cellules cancéreuses $P(x,t)$, et un phénotype sain, de densité $S(x,t)$. L'hypothèse ici est que la croissance de la tumeur est liée au mouvement passif des cellules sous l'effet de la prolifération, comme représenté sur le schéma de la figure 3.1. Les nouvelles cellules qui apparaissent au sein d'un milieu saturé poussent les autres cellules vers l'extérieur.

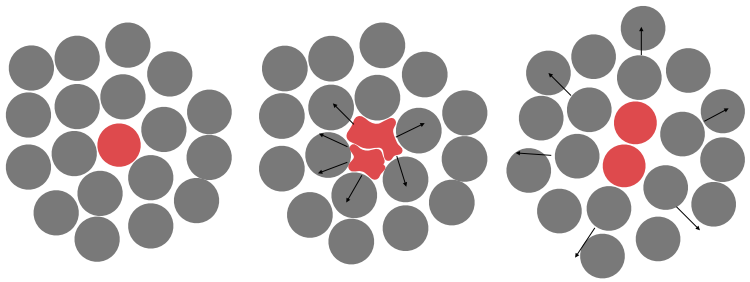


FIGURE 3.1 – Schéma de croissance tumorale : prolifération des cellules cancéreuses (en rouge) puis déplacement des cellules dû à la pression interne. Les cellules saines (gris) sont poussées vers l'extérieur.

Nous supposons également que la prolifération nécessite la consommation de nutriments et d'oxygène représentés par le champ $M$. Notons $\mathcal{B}$ le domaine d'étude, en 2D ou en 3D. Nous considérons donc le modèle suivant [17, 18] :

$$\begin{cases} \partial_t P + \nabla \cdot (\mathbf{v}\, P) &= MP, & \mathcal{B} \\ \partial_t S + \nabla \cdot (\mathbf{v}\, S) &= 0, & \mathcal{B} \\ \partial_t M &= -\alpha MP, & \mathcal{B} \end{cases} \tag{3.1}$$

avec les conditions initiales suivantes :

$$\begin{cases} P(0,x) &= P_0(x), \\ S(0,x) &= 1 - P_0(x), \\ M(0,x) &= M_0(x). \end{cases} \tag{3.2}$$

Puisque la tumeur est au centre du domaine et que la tumeur est en phase de croissance, la vitesse est sortante et il n'y a donc pas besoin de conditions aux bords sur $P$.
On suppose également que le milieu est saturé, c'est-à-dire que $P+S=1$, et que la vitesse

$\mathbf{v}$ suit une loi de Darcy, à savoir :

$$\mathbf{v} = -\nabla\pi. \tag{3.3}$$

Cela signifie que la vitesse dérive d'une pression $\pi$ qui représente la compétition spatiale entre les cellules. On obtient alors l'équation suivante qui permet de fermer le système et de calculer $\mathbf{v}$ :

$$\nabla \cdot \mathbf{v} = -\Delta\pi = MP. \tag{3.4}$$

On obtient une équation de Poisson où l'on fixe des conditions de Dirichlet homogènes au bord modélisant le fait que la pression loin de la tumeur n'est pas affectée par la croissance. Lorsqu'il existe une contrainte mécanique à la croissance de la tumeur (par exemple le crâne), cette dernière ne peut grossir qu'à l'intérieur d'un nouveau domaine $\mathcal{D}$. On fixe alors des conditions de Neumann homogènes au bord de $\mathcal{D}$. Le schéma de la figure 3.2 représente ces conditions mixtes sur un exemple simple.

Le modèle comprend donc un terme de prolifération $MP$ et un terme de transport conservatif $\nabla \cdot (\mathbf{v}P)$ qui modélise la poussée entre les cellules. La prolifération est proportionnelle à la quantité de nutriments présents dans le milieu. Les cellules saines sont transportées selon la même vitesse, subissant donc un mouvement passif lié à la croissance de la tumeur. Lorsque $\alpha > 0$, la prolifération entraîne une consommation de nutriments au taux $\alpha$. Ce modèle permet également de modéliser une croissance très forte, lorsque $\alpha < 0$. Biologiquement, cela signifie que la tumeur parvient à s'auto-vasculariser ce qui correspond au phénomène d'angiogénèse.

Les paramètres de ce modèle sont donc ce taux de consommation de nutriments $\alpha$ ainsi que les conditions initiales $P_0(x)$ et $M_0(x)$. Dans la pratique, $P_0(x)$ est issu directement des données d'imagerie par segmentation. Cela ramène donc le système à deux paramètres, dont un spatial, $M_0$.

### 3.2.2 Résolution numérique

#### Méthode générale

Ce modèle ainsi que tous les modèles suivants ont été implémentés en C++ en utilisant la librairie Cadmos développée par l'équipe MONC, qui contient des solveurs des équations aux dérivées partielles de type transport et diffusion. Ces solveurs utilisent des méthodes de type volumes finis. Le maillage choisi est un maillage cartésien, ce qui est bien adapté à l'imagerie médicale. Chaque case de la grille contient l'intensité du champ considéré. Le choix du nombre de mailles dépend à la fois de la qualité de l'image médicale utilisée, mais

aussi du temps de calcul et de la précision numérique souhaités. Les exemples synthétiques en 2D ont été réalisés avec une grille 200x200. Les exemples cliniques 3D proposés dans cette thèse sont réalisés sur la grille issue de l'imagerie. Les champs $P$, $S$, $M$ et $\pi$ sont donnés au centre des mailles et les vitesses sont calculées sur la frontière des mailles, ce qui permet d'avoir une méthode de résolution à l'ordre 2. La résolution numérique du modèle se fait alors ainsi :

**Initialisation**

— Initialisation des grandeurs $P_0$ et $S_0 = 1 - P_0$ à partir de l'imagerie médicale au premier temps.

— Initialisation de $M_0$ en prenant un champ constant spatialement de valeur $m_0$. Dans les sections suivantes, $M_0$ pourra prendre différentes valeurs dans des zones dépendantes du milieu (matière blanche et grise, vaisseau).

**Itération du modèle entre $i$ et $i+1$**

— Résolution de l'équation de Poisson (3.4) pour déterminer la pression $\pi$.

— Calcul de la vitesse $\mathbf{v}$ grâce à la loi de Darcy (3.3).

— Calcul de $P$ en résolvant l'équation (3.87) puis calcul de $S = 1 - P$.

— Mise à jour de $M$ par l'équation (1.13).

L'équation régissant $M$ est résolue par un schéma exponentiel classique. Les autres résolutions sont détaillées dans les sous-sections suivantes.

**Résolution de l'équation de transport**

L'équation régissant l'évolution des cellules tumorales $P$ contient un terme de transport ainsi qu'un terme de croissance. Ces deux parties de l'équation sont résolues par méthode de splitting de Strang [19]. Nous nous intéressons donc dans un premier temps à l'équation suivante :

$$\partial_t P + \nabla \cdot (\mathbf{v}\, P) = 0. \tag{3.5}$$

En développant le terme de divergence, on obtient :

$$\partial_t P + \mathbf{v} \cdot \nabla P = -P \nabla \cdot \mathbf{v}. \tag{3.6}$$

Or, l'équation (3.4) donne $\nabla \cdot \mathbf{v} = MP$, donc :

$$\partial_t P + \mathbf{v} \cdot \nabla P = -MP^2. \tag{3.7}$$

Le terme de droite est un terme source qui assure la conservation de la masse. Encore une fois, on résout ce genre d'équation par splitting en traitant le terme source à part. On s'intéresse alors simplement à l'équation :

$$\partial_t P + \mathbf{v} \cdot \nabla P = 0. \tag{3.8}$$

Cette équation est résolue par un schéma WENO d'ordre 5 [20, 21], de type volumes finis, avec des conditions de Dirichlet homogènes au bord. Ce schéma permet de minimiser la perte de masse ainsi que la diffusion numérique. Le splitting donne donc le schéma global suivant :

- résolution de l'équation (3.8) par WENO d'ordre 5 sur l'intervalle de temps $dt$.

- résolution de la partie réaction par un schéma exponentiel sur l'intervalle de temps $dt$.

**Résolution de l'équation de Poisson**

Lorsque la tumeur croît sans contraintes mécaniques (bord du crâne par exemple), l'équation de Poisson (3.4) est résolue en discrétisant le Laplacien par un schéma à 5 points (en 2D) ou à 7 points (en 3D). La pression étant estimée à une constante près, on fixe des conditions de Dirichlet homogènes aux bords du domaine. Lorsqu'un bord ajoute une contrainte mécanique à la tumeur, cette dernière ne peut grossir qu'à l'intérieur d'un nouveau domaine $\mathcal{D}$. On fixe alors des conditions de Neumann homogènes aux bords de $\mathcal{D}$ (voir le schéma de la figure 3.2). Notons $\pi$ la solution théorique de l'équation 3.4 avec ces conditions aux bords. On approche $\pi$ est résolvant l'équation différentielle par pénalisation :

$$-\nabla \cdot (K \nabla \pi_\epsilon) = MP_\epsilon, \qquad \text{sur } \mathcal{B} \tag{3.9}$$

où $K$ vaut 1 à l'intérieur du $\mathcal{D}$ domaine, et $\epsilon = 0.01$ à l'extérieur. Cette approximation de la solution permet de trouver $\pi_\epsilon$ tel que $\pi_\epsilon - \pi$ soit de l'ordre $\epsilon$. Le calcul de $\mathbf{v}$ se fait ensuite grâce à la loi de Darcy.

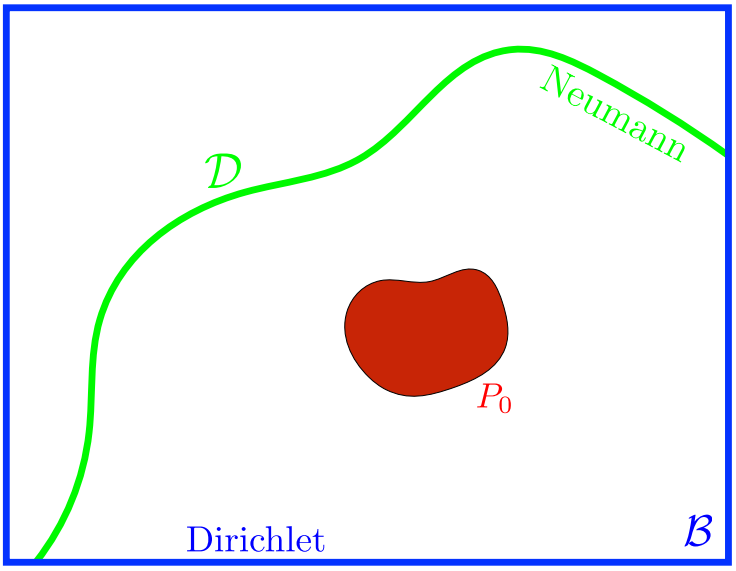



FIGURE 3.2 – Schéma des conditions aux bords considérées.

### 3.2.3 Premières simulations

Le modèle est simulé dans le cas simple où $M_0$ est un champ constant spatialement. La figure 3.3 contient l'évolution des cellules proliférantes et du champ de nutriments.

La croissance de la tumeur se fait de façon isotrope, étant donné que le champ des nutriments est choisi initialement uniforme. La consommation de nutriments sature la croissance au coeur de la tumeur.

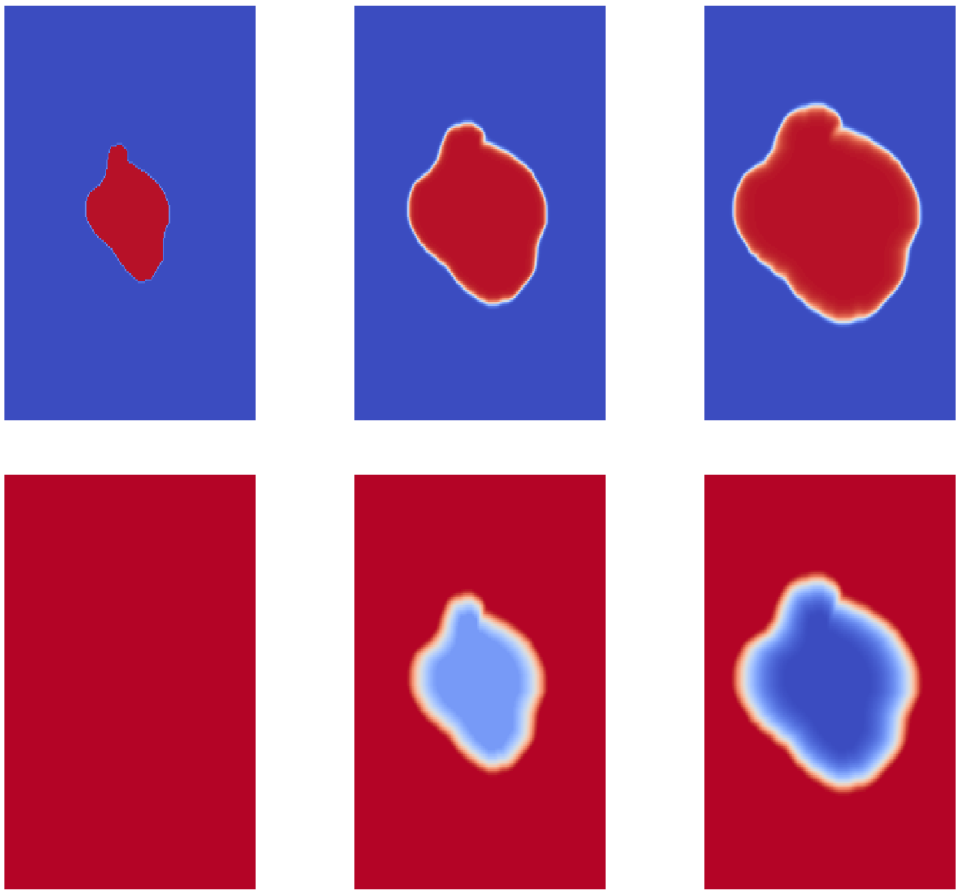
### 3.2.4 Calcul du volume tumoral

On suppose que le front de la tumeur $\Gamma_P(t)$ est défini par une certaine valeur seuil $P_{\text{th}}$, c'est-à-dire :

$$\Gamma_P(t) = \{\vec{x} \in \mathcal{B},\, P(\vec{x}, t) = P_{\text{th}}\},$$

et l'on définit également l'intérieur de la tumeur :

$$\Omega_P^{\text{in}}(t) = \{\vec{x} \in \mathcal{B},\, P(\vec{x}, t) > P_{\text{th}}\}.$$



(a) $t = 0$ (b) $t = 1.5$ (c) $t = 3$

FIGURE 3.3 – Evolution de la densité de cellules proliférantes $P$ (première ligne) et de la densité de nutriments $M$, avec $(\alpha, m_0) = (0.1, 0.02)$.

Habituellement, la tumeur prend initialement les valeurs de 0 à l'extérieur et 1 à l'intérieur, et le seuil choisi est $P_{\text{th}} = 0.5$. Puisque 0 et 1 sont des points fixes de l'équation sur $P$, on a $\forall(t, x)$, $P(t, x) \in \{0, 1\}$, c'est-à-dire que la densité $P$ prend la valeur 0 ou 1. Notons qu'en pratique, et comme on peut le voir sur la figure 3.3, on observe une couche limite autour de la tumeur qui est due à la diffusion numérique.
Notons $V(t) = \int_{\Omega_P^{\text{in}}} P d\mathcal{B}$ la masse totale de la tumeur. On suppose dans cette section que $M_0$ est un champ constant de valeur $m_0$. Comme initié dans [22], le volume tumoral est régi par l'équation suivante :

$$\frac{\partial V}{\partial t} = \frac{\partial}{\partial t} \int_{\Omega_P^{\text{in}}} P d\mathcal{B} = \int_{\Omega_P^{\text{in}}} \frac{\partial P}{\partial t} d\mathcal{B} + \int_{\Omega_P^{\text{in}}} \nabla \cdot (vP) d\mathcal{B}, \tag{3.10}$$

en utilisant le théorème de Reynolds.
Grâce à l'équation (3.87), on obtient :

$$\frac{\partial V}{\partial t} = \int_{\Omega_P^{\text{in}}} MP d\mathcal{B} - \int_{\Omega_P^{\text{in}}} \nabla \cdot (vP) d\mathcal{B} + \int_{\Omega_P^{\text{in}}} \nabla \cdot (vP) d\mathcal{B} = \int_{\Omega_P^{\text{in}}} M d\mathcal{B}, \tag{3.11}$$

puisque $P = 1$ sur $\Omega_P^{\text{in}}$.
Notons $\bar{M}(t) = \frac{1}{V(t)} \int_{\Omega_P^{\text{in}}} M d\mathcal{B}$. On obtient donc $\frac{\partial V}{\partial t} = \bar{M}(t)V(t)$ et il faut donc évaluer $\bar{M}(t)$.

$$\begin{aligned}
\bar{M}'(t) &= \frac{-V'(t)}{V^2(t)} \int_{\Omega_P^{\text{in}}} M d\mathcal{B} + \frac{1}{V(t)} \frac{\partial}{\partial t} \int_{\Omega_P^{\text{in}}} M d\mathcal{B}, \\
&= \frac{-V'(t)}{V(t)} \bar{M}(t) + \frac{1}{V(t)} \left( \int_{\Omega_P^{\text{in}}} \frac{\partial M}{\partial t} + \nabla \cdot (Mv) \right), \\
&= -\bar{M}^2(t) + \frac{1}{V(t)} \left( \int_{\Omega_P^{\text{in}}} -\alpha M + \int_{\Omega_P^{\text{in}}} \nabla \cdot (Mv) \right), \\
&= -\bar{M}^2(t) - \alpha \bar{M} + \frac{1}{V(t)} \int_{\partial\Omega_P^{\text{in}}} Mv \cdot n ds,
\end{aligned} \tag{3.12}$$

et comme $M = m_0$ sur le contour $\partial\Omega_P^{\text{in}}$ :

$$\begin{aligned}
\bar{M}'(t) &= -\bar{M}^2(t) - \alpha \bar{M} + \frac{m_0}{V(t)} \int_{\Omega_P^{\text{in}}} \nabla \cdot \mathbf{v} d\mathcal{B}, \\
&= -\bar{M}^2(t) - \alpha \bar{M} + \frac{m_0}{V(t)} \int_{\Omega_P^{\text{in}}} MP d\mathcal{B}, \\
&= -\bar{M}^2(t) - \alpha \bar{M} + m_0 \bar{M}, \\
&= -\bar{M}(\bar{M} - (m_0 - \alpha)).
\end{aligned} \tag{3.13}$$

En intégrant l'équation précédente, on obtient :

$$\frac{\bar{M}}{\bar{M} - (m_0 - \alpha)} = \frac{m_0}{\alpha} \exp(m_0 - \alpha)t, \tag{3.14}$$

ce qui donne finalement :

$$\bar{M} = \frac{-(m_0 - \alpha)\frac{m_0}{\alpha} \exp(m_0 - \alpha)t}{1 - \frac{m_0}{\alpha} \exp(m_0 - \alpha)t}. \tag{3.15}$$

Puisque

$$\frac{\partial V}{\partial t} = \bar{M} V, \tag{3.16}$$

on obtient en intégrant :

$$V(t) = V_0 \left(1 + \frac{m_0}{m_0 - \alpha} \left(\exp(m_0 - \alpha)t - 1\right)\right). \tag{3.17}$$

**Le volume tumoral est donc connu explicitement dans le cas où $M_0$ est un champ constant**.

## 3.3 Estimation des paramètres par calibrage volumique

### 3.3.1 Principe du calibrage volumique

Le but est de calibrer le modèle à partir des données d'imagerie, c'est-à-dire de trouver les paramètres $\alpha$ et $m_0$ adaptés à chaque patient, comme initié dans [22]. Les données acquises pour un patient donné sont des séquences de $k+1$ examens. Seulement les $k$ premiers examens sont utilisés pour la calibration. Le dernier point permet de mesurer l'efficacité de la prédiction en la comparant à l'examen réel. En pratique, le nombre d'examens dont on se sert pour calibrer varie entre 2 et 4. Pour chaque examen, on dispose d'une segmentation de la tumeur ainsi que de la date d'examen. Dans un premier temps, seule la donnée sur le volume tumoral sera utilisée. Chaque patient est donc caractérisé par une liste de $k$ couples $(t_i, V_i)$ où $t_i$ est le temps de l'examen et $V_i$ le volume tumoral à $t_i$. L'idée est de trouver les paramètres du modèle en utilisant l'expression du volume calculée précédemment. On souhaite minimiser l'erreur relative suivante :

$$\varepsilon(\alpha, m_0) = \max\left\{ \frac{(V_{\alpha,m_0}(t_i) - V_i)}{V_i}, \quad i = 1...k \right\}. \tag{3.18}$$

Les erreurs liées à la segmentation de la tumeur peuvent entraîner des variations importantes du volume. Afin d'obtenir une méthode robuste à ces incertitudes, on ne souhaite pas forcément trouver le jeu de paramètres qui minimise $\varepsilon(\alpha, m_0)$, mais on regarde plutôt l'ensemble des jeux de paramètres convenables. On définit pour cela une erreur maximale autorisée $\varepsilon_{max}$. Une étude comparant les erreurs de segmentation inter-opérateurs a montré que l'erreur peut atteindre 12%. Ce chiffre est obtenu en comparant les contours de plusieurs tumeurs par 20 cliniciens différents [13]. On choisit donc ici $\varepsilon_{max} = 0.12$.

On applique alors une méthode de Monte-Carlo : on effectue le tirage d'un grand nombre de jeux de paramètres, pour lesquels $\varepsilon$ est calculé. Notons alors $E$ l'ensemble :

$$E = \{(\alpha, m_0), \quad \varepsilon(\alpha, m_0) < \varepsilon_{max}\}. \tag{3.19}$$

L'ensemble $E$ contient donc tous les jeux de paramètres donnant une simulation acceptable, au sens où l'erreur relative entre les volumes estimés et observés ne dépasse pas le seuil de tolérance. Tous les jeux de paramètres de l'ensemble $E$ correspondent donc à une courbe d'évolution du volume différente. Afin de sélectionner le jeu de paramètres de $E$ le plus adapté, l'idée est de sélectionner celui qui donne un comportement volumique le plus

probable. On choisit alors le jeu $(\alpha, m_0)$ tel que :

$$V_{(\alpha,m_0)}(t_k) = \text{médiane}\left\{V_{(\alpha_i,m_{0,i})}(t_k),\ (\alpha_i, m_{0,i}) \in E\right\}. \tag{3.20}$$

Le volume choisi est donc le volume médian, soit la valeur seuil qui sépare l'ensemble des volumes acceptés en deux sous-ensembles de même taille. Cela permet alors d'obtenir un jeu de paramètres qui traduit bien l'évolution volumique, tout en prenant en compte les erreurs de segmentation.

### 3.3.2 Espace de recherche des paramètres $\alpha$ et $m_0$

La méthode de Monte Carlo nécessite de tirer aléatoirement un grand nombre de jeux de paramètres afin de maximiser le nombre d'éléments de $E$ et ainsi d'avoir un résultat précis. Cependant, un tel tirage nécessite de donner des bornes de recherches pour chacun des deux paramètres. Plutôt que d'effectuer un tirage dans une large zone, l'idée est d'utiliser le modèle pour trouver une zone de recherche plus précise dans laquelle un jeu de paramètres a des chances d'appartenir à $E$. Réduire la zone de recherche permet de réduire le nombre de tirages et diminue donc le temps de calcul. Nous supposons dans cette section que le paramètre $\alpha$ est positif, c'est-à-dire que la tumeur consomme des nutriments afin de proliférer.

Le champ $M$ de la densité de nutriments dans le milieu a une valeur initiale uniforme de $m_0$ puis diminue là où la tumeur prolifère. On a donc $\forall (t,x) \in \mathbb{R}^+ \times \mathcal{B}$ :

$$M(t,x) \leqslant m_0. \tag{3.21}$$

En reportant cette inégalité dans l'équation (3.16), et puisque $\bar{M} \leqslant m_0$, on obtient :

$$V(t) \leqslant V(0)e^{m_0 t}. \tag{3.22}$$

Rappelons que la densité $P$ vaut toujours 0 à l'extérieur de la tumeur et 1 à l'intérieur lorsque c'est le cas pour la condition initiale. Ainsi, le champ $M$ à l'intérieur de la tumeur initiale ($P_0 = 1$) est régi par l'équation :

$$\frac{\partial M}{\partial t} = -\alpha M, \tag{3.23}$$

d'où l'équation à l'intérieur de la tumeur initiale : $\forall (t,x) \in \mathbb{R}^+ \times \Omega^{\text{in}}_{P_0}(t)$

$$M(t,x) = m_0 e^{-\alpha t}. \tag{3.24}$$

Le coeur de la tumeur étant la zone où le plus de nutriments sont consommés, on en déduit donc la minoration suivante : $\forall (t,x) \in \mathbb{R}^+ \times \mathcal{B}$,

$$M(t,x) \geqslant m_0 e^{-\alpha t}, \tag{3.25}$$

ce qui donne, en intégrant et par le théorème des accroissements finis :

$$\ln\left(\frac{V(t)}{V_0}\right) \geqslant \frac{m_0}{\alpha}\left(1 - e^{-\alpha t}\right), \tag{3.26}$$

et donc :

$$V(t) \geqslant V_0 e^{\frac{m_0}{\alpha}(1-e^{-\alpha t})}. \tag{3.27}$$

Finalement, le volume tumoral est borné par :

$$V_0 e^{\frac{m_0}{\alpha}(1-e^{-\alpha t})} \leqslant V(t) \leqslant V(0) e^{m_0 t}. \tag{3.28}$$

Supposons donc que l'on possède deux examens $(t_0, V_0)$ et $(t_1, V_1)$. Notons que $t_0$ est choisi comme origine des temps : $t_0 = 0$. L'équation (3.30) donne donc :

$$V_0 e^{\frac{m_0}{\alpha}(1-e^{-\alpha t_1})} \leqslant V_1 \leqslant V_0 e^{m_0 t_1}. \tag{3.29}$$

d'où la proposition :

**Proposition 1.**

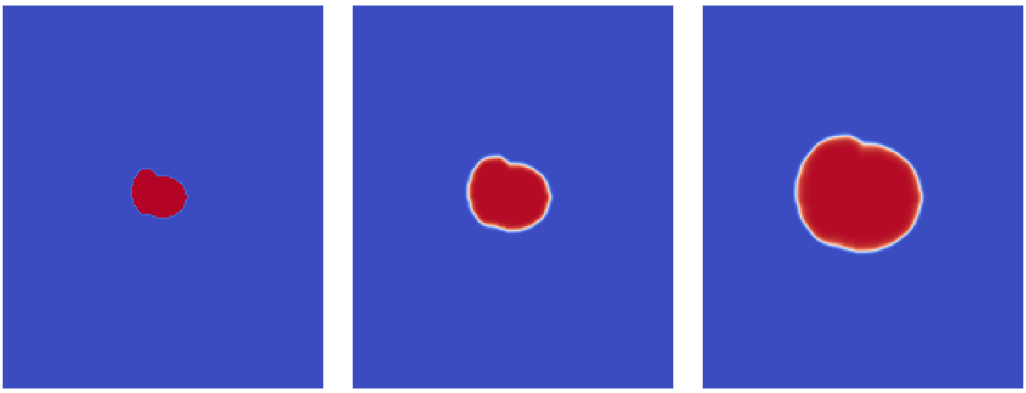
$$\frac{1}{t_1}\ln\left(\frac{V_1}{V_0}\right) \leqslant m_0 \leqslant \frac{\alpha}{1-e^{-\alpha t_1}}\ln\left(\frac{V_1}{V_0}\right). \tag{3.30}$$

Le paramètre $m_0$ est donc borné en fonction de la valeur du paramètre $\alpha$. Pour ce paramètre $\alpha$ par contre, aucun moyen simple ne permet de trouver un intervalle de recherche. On décide donc de le tirer aléatoirement dans l'intervalle $[0,1]$. Puisqu'on autorise une erreur relative de 12% sur le volume $V_1$, la recherche se fait finalement dans l'espace de paramètres suivant :

$$\begin{aligned} &\alpha \in [0,1], \\ &m_0 \in \left[\frac{1}{t_1}\ln\left(\frac{0.88V_1}{V_0}\right), \frac{\alpha}{1-e^{-\alpha t_1}}\ln\left(\frac{1.12V_1}{V_0}\right)\right]. \end{aligned} \tag{3.31}$$

### 3.3.3 Exemple sur un cas synthétique

On construit une séquence de données synthétiques en lançant le modèle avec des paramètres $(\alpha^t, m_0^t) = (0.02, 0.1)$. On obtient alors les données représentées sur la figure 3.4.



(a) $t = 0$ (b) $t = 10$ (c) $t = 20$

FIGURE 3.4 – Données synthétiques utilisées pour valider le calibrage volumique.

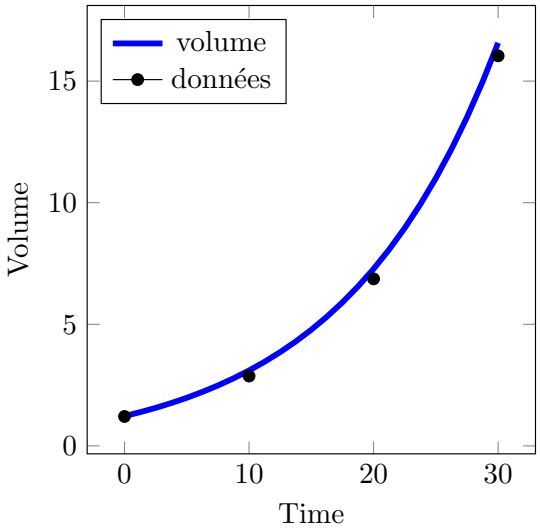



FIGURE 3.5 – Données synthétiques (points noirs) représentant 4 examens. La courbe bleue représente l'expression théorique du volume tumoral, obtenue avec les paramètres cibles $(\alpha^t, m_0^t) = (0.02, 0.1)$. Celle-ci est légèrement au-dessus des points de données à cause de la diffusion numérique.

On calcule alors le volume tumoral à chacun des trois temps, afin d'obtenir les couples $(t_i, V_i)$ pour $i = 1...3$. Ces données sont représentées sur la figure 3.5. La courbe du volume calculé à partir de l'expression théorique et avec les paramètres cibles passe bien par ces points. Le quatrième examen est gardé de côté pour comparer ensuite à la prédiction du modèle. La méthode de calibrage volumique décrite précédemment est appliquée, en limitant l'espace de recherche des paramètres à l'espace décrit dans la proposition 1. La

figure 3.6 montre les paramètres acceptés (en rouge), c'est-à-dire l'espace $E$.

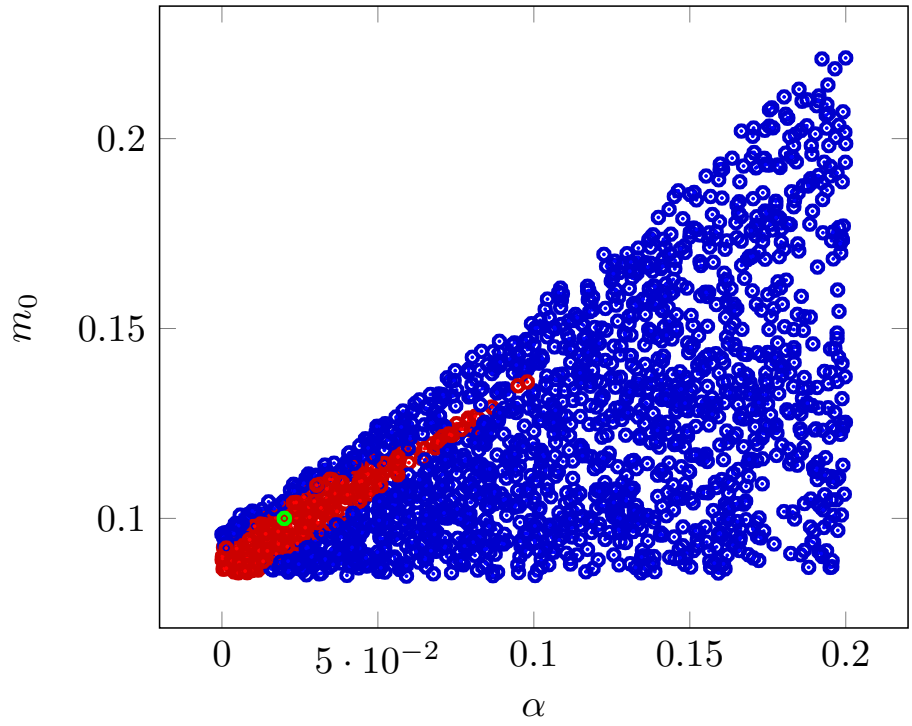



FIGURE 3.6 – Méthode de Monte Carlo appliquée à l'espace des paramètres $(\alpha, m_0)$. En rouge, les jeux de paramètres acceptés (l'erreur volumique est en dessous du seuil de tolérance pour chaque examen) et en bleu ceux refusés. Le point vert est le jeu de paramètres cible.

L'espace $E$ ainsi déterminé, nous calculons le jeu de paramètres retenu par la méthode. Les volumes au temps final $t = 30$ sont calculés pour chaque jeu de paramètres de $E$. On retient alors le jeu de paramètres donnant le volume médian parmi tous ces volumes. La figure 3.7 montre le faisceau de volumes obtenus avec les paramètres de $E$, ainsi que la courbe correspondant au jeu de paramètres retenu. La figure 3.8 montre l'histogramme des volumes finaux obtenus avec les paramètres de $E$, ainsi que le volume médian retenu et le volume réel.

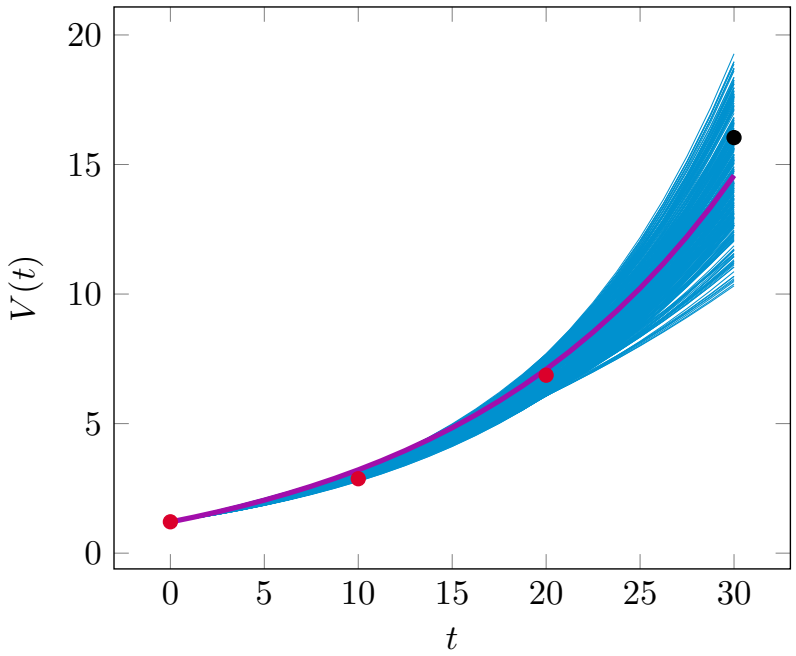



FIGURE 3.7 – Faisceau des volumes théoriques pour chaque jeu de paramètres de l'ensemble $E$. Les points rouges sont les données utilisées pour le calibrage volumique et le point noir est la donnée qui sert à comparer la prédiction et le volume réel. La courbe violette est celle obtenue avec le jeu de paramètres qui donne un volume final médian parmi ceux de $E$.

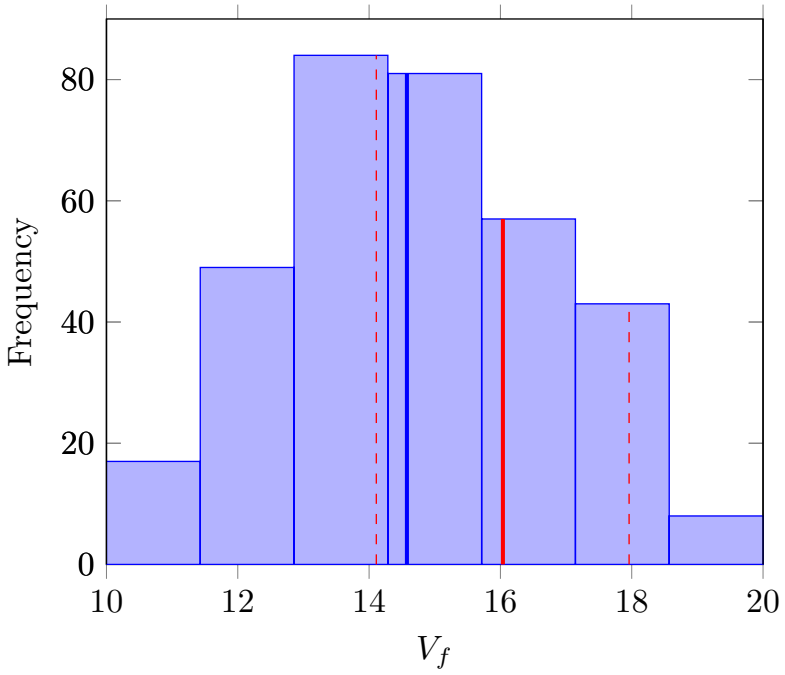



FIGURE 3.8 – Histogrammes des volumes finaux ($t = 30$) pour chaque jeu de paramètres de l'ensemble $E$. La droite bleue représente le volume médian parmi ceux de $E$, la droite rouge représente le volume réel, et les droites rouges pointillées montrent l'intervalle d'erreur de 12% avec le volume réel.

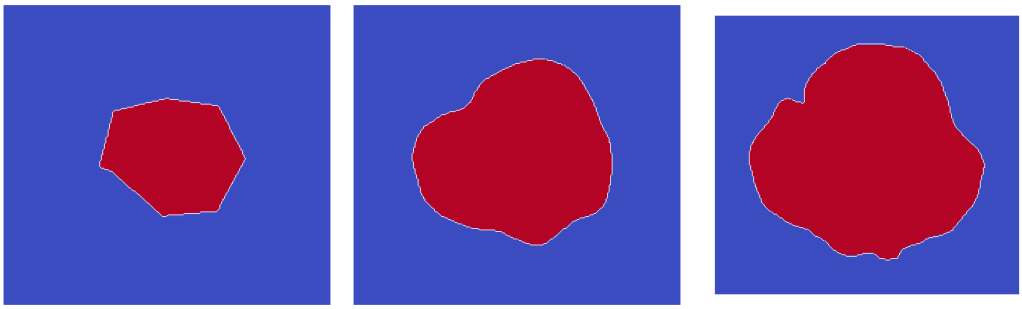
Le quatrième point n'étant pas utilisé pour la calibration, il permet ici de constater que la prédiction est satisfaisante : l'erreur relative entre le volume prédit et le volume observé est de 9.1%, soit plus petite que la marge d'erreur de 12% de segmentation. Par contre, le jeu de paramètres estimé vaut $(\alpha^e, m_0^e) = (0.052, 0.118)$ ce qui est loin du jeu de paramètres utilisé pour créer les données $(\alpha^t, m_0^t) = (0.02, 0.1)$. Cela montre que les deux paramètres sont corrélés et donc difficilement identifiables. En effet, l'évolution tumorale obtenue avec une faible quantité de nutriments, et une faible consommation de ces nutriments peut être très proche de l'évolution obtenue avec plus de nutriments, mais aussi plus de consommation. C'est ce qu'on remarque ici, avec des paramètres estimés tous les deux plus grands que ceux ciblés. Cependant, lorsque la valeur de $\alpha$ est fixée, le paramètre $m_0$ est moins variable et plus identifiable, comme on le remarque sur la figure 3.6. **La validation du modèle ne se fera donc pas en regardant si l'on retrouve les paramètres cibles, mais plutôt en vérifiant si la prédiction d'évolution est satisfaisante comme c'est le cas ici**. Une mesure de discordance entre la simulation et l'observation sera mise en place afin d'évaluer la qualité de la prédiction.

Notons qu'il aurait également été possible de rajouter un terme de pénalisation sur les pa-

ramètres, avec une norme $L^1$ ou $L^2$, afin de rendre l'estimation paramétrique plus robuste. Cela n'a pas été fait ici puisque l'objectif est de reproduire l'évolution.

### 3.3.4 Exemple sur un cas clinique 2D : métastase au poumon

Nous souhaitons valider cette calibration volumique dans un cas clinique. Nous disposons pour cela de données de métastases pulmonaires en 2D, fournies par le service radiologique de l'Institut de Cancérologie Bergonié. Sur la figure 3.9, ces données sont représentées pour trois temps. Le premier temps montre les limites de la segmentation : le contour tumoral est polygonal et peu réaliste. Les deux premiers temps sont utilisés pour le calibrage et le troisième temps permet de comparer la prédiction à l'observation. Comme précédemment, les couples $(t_i, V_i)$ sont calculés pour $i = 1$ et $i = 2$. La méthode décrite ci-dessous est alors appliquée à ces deux points. Notons que contrairement au cas synthétique précédent, seulement deux points sont utilisés pour la calibrage. Le nombre de jeux de paramètres dans l'espace $E$ est donc plus grand, et le champ des volumes prédits est donc plus étendu. Le faisceau des courbes volumiques obtenues avec les jeux de paramètres acceptés dans $E$ est représenté sur la figure 3.10. La courbe donnant un volume final médian parmi les volumes finaux de $E$ est représentée en violet. Le volume prédit par notre méthode est alors de 244 $cm^2$, soit une erreur relative de 11.9% avec le volume réel de 218 $cm^2$. L'erreur est plus importante que dans le cas précédent, ce qui peut s'expliquer par deux raisons : tout d'abord, on ne possède que 2 examens, contre 3 dans le cas précédent. De plus, dans le cas synthétique, la tumeur simulée était obtenue grâce au même modèle qui est utilisé pour le calibrage. Dans le cas clinique, les erreurs de modélisation se rajoutent donc aux erreurs d'estimation. Cependant, l'erreur relative reste inférieure à 12%, ce qui est acceptable compte tenu de la marge d'erreur liée à la segmentation.



(a) $t = 0$ (b) $t = 701$ (c) $t = 1434$

FIGURE 3.9 – Données cliniques de métastases au poumon 2D utilisées pour valider le calibrage volumique. Elles sont fournies par le service radiologique de l'Institut de Cancérologie Bergonié.

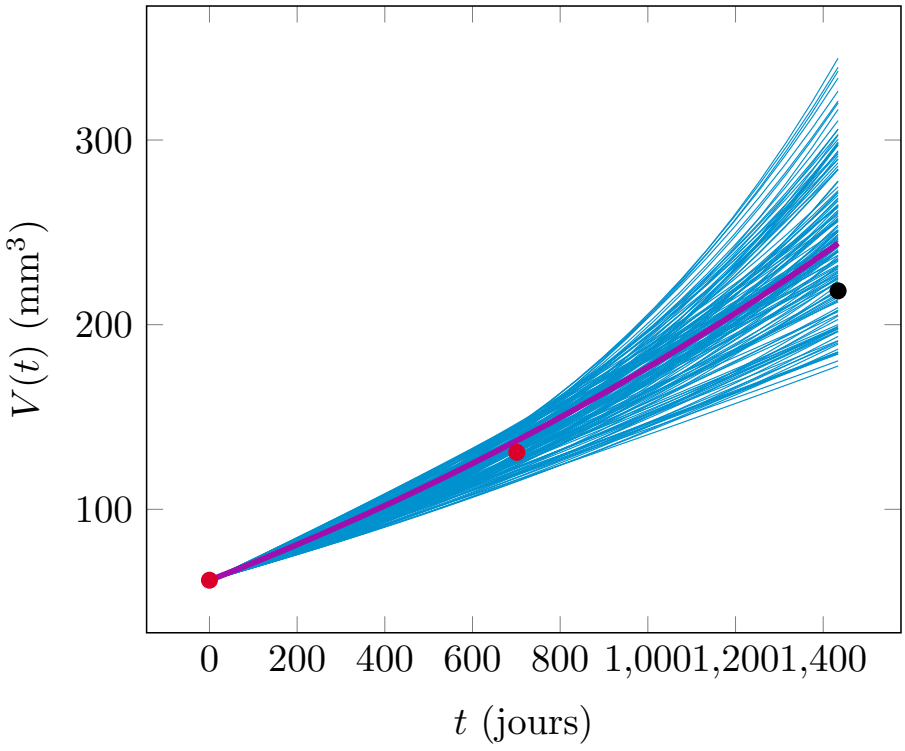
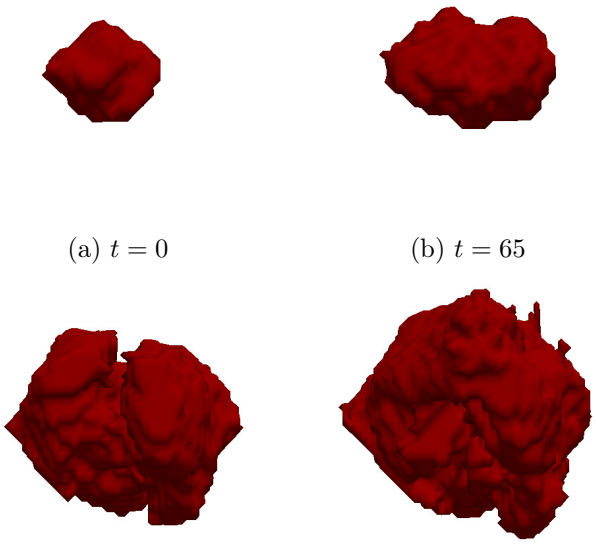



FIGURE 3.10 – Faisceau des volumes théoriques pour chaque jeu de paramètres de l'ensemble $E$, dans le cas d'une métastase pulmonaire 2D. Les points rouges sont les données utilisées pour le calibrage volumique et le point noir est la donnée qui sert à comparer la prédiction et le volume réel. La courbe violette est celle obtenue avec le jeu de paramètres qui donne un volume final médian parmi ceux de $E$.

### 3.3.5 Exemple sur un cas clinique 3D : métastase cérébrale

Les données de métastases cérébrales 3D sont représentées sur la figure 3.11 et sont fournies par Ana Ortiz de Mendivil Arrate (HM Hospitales, Espagne). La croissance volumique est représentée par les points rouges sur la figure 3.12. Comme précédemment, la courbe violette est celle donnant le volume médian parmi tous les volumes acceptés. Les paramètres retenus sont ceux correspondant à cette courbe. Seuls les trois premiers points sont utilisés pour la calibration volumique et on observe que le quatrième volume est parfaitement estimé par la calibration.



(a) $t = 0$ (b) $t = 65$

(c) $t = 275$ (d) $t = 324$

FIGURE 3.11 – Examens de métastase cérébrale 3D fournies par Ana Ortiz de Mendivil Arrate (HM Hospitales, Espagne). Les temps sont en jours.

### 3.3.6 Insuffisance du calibrage volumique

Comme montré dans les deux exemples précédents, le calibrage volumique est un outil fiable pour estimer le volume tumoral à un temps donné. Une fois les paramètres estimés, il est possible de relancer la simulation avec ce jeu de paramètres afin de visualiser la croissance tumorale. Dans le cas synthétique décrit ci-dessus, on obtient le résultat de la figure 3.13.

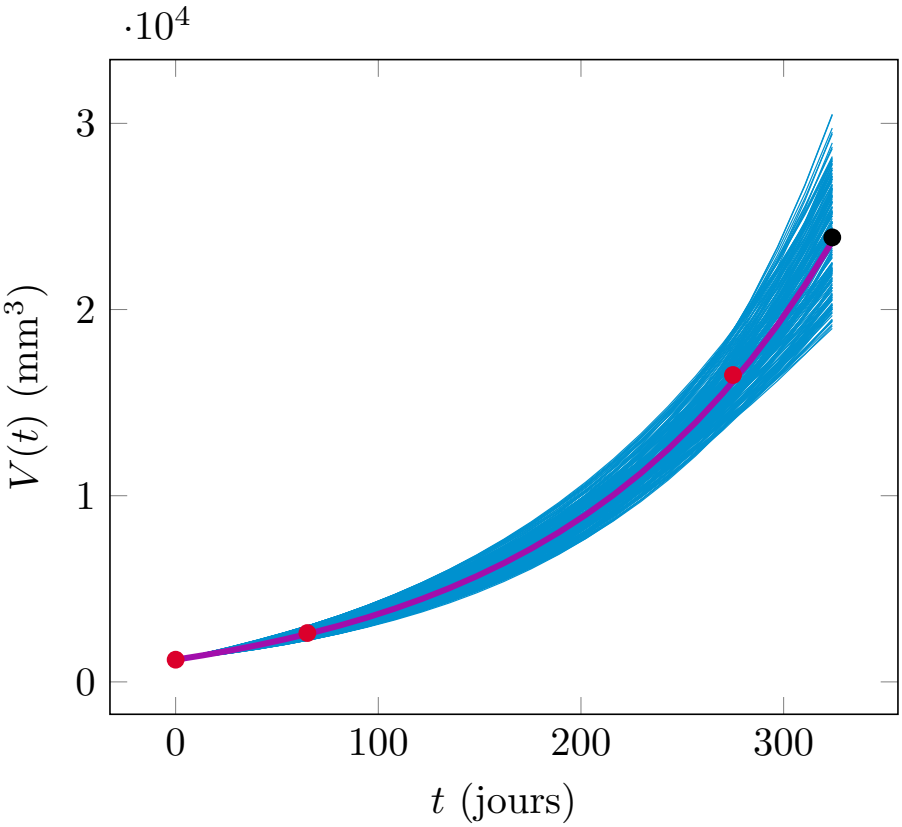



FIGURE 3.12 – Faisceau des volumes théoriques pour chaque jeu de paramètres de l'ensemble $E$, dans le cas d'une métastase cérébrale 3D. Les points rouges sont les données utilisées pour le calibrage volumique et le point noir est la donnée qui sert à comparer la prédiction et le volume réel. La courbe violette est celle obtenue avec le jeu de paramètres qui donne un volume final médian parmi ceux de $E$.

L'erreur commise lors de la calibration volumique se retrouve à la simulation : la tumeur estimée est légèrement plus petite que celle observée. Cependant, la forme prédite est conforme à la tumeur cible. Encore une fois, cela s'explique par le fait que la tumeur cible est obtenue à partir du modèle utilisé pour la calibration : dans les deux cas, la tumeur est générée à partir d'une vascularisation $M_0$ uniforme spatialement et la croissance se fait ainsi de manière isotrope. Comparons également la tumeur cible et celle simulée dans le cas clinique. Dans le cas des métastases pulmonaires, on obtient la figure 3.14. On remarque dans ce cas que l'estimation en forme est très mauvaise. En effet, la forme simulée garde la même structure qu'initialement, ce qui est dû au caractère isotrope du modèle. Notons qu'en plus, la tumeur initiale est petite et la segmentation comporte peu de points. L'incertitude est donc grande et l'erreur est ainsi propagée sur toute la simulation. Les deux premiers examens ont été utilisés pour le calibrage volumique, qui ne prend en compte que le volume tumoral. On se rend compte ici que le changement de forme apparu entre

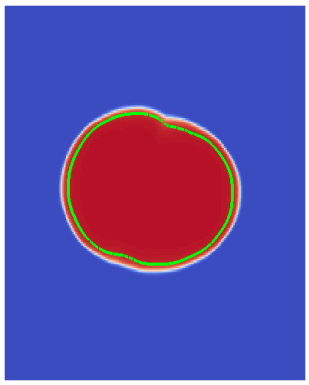


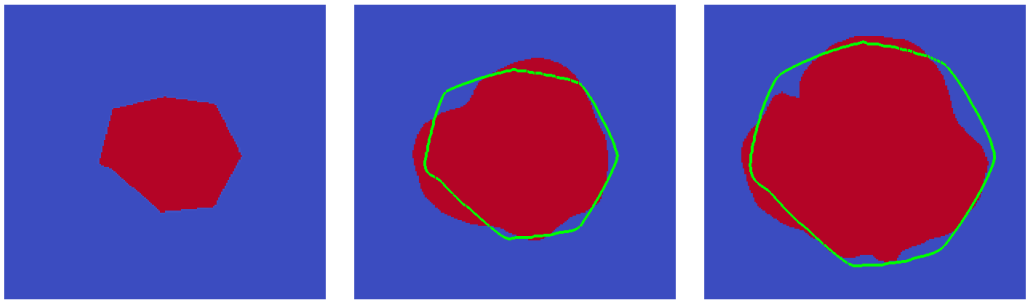
FIGURE 3.13 – Comparaison dans le cas synthétique entre la tumeur cible (en rouge) et la tumeur obtenue par la simulation avec les paramètres estimés par le calibrage volumique (contour vert) au temps $t = 300$.

le premier et le second examen n'est pas du tout pris en considération par ce calibrage. La forme simulée est donc plutôt éloignée de la forme cible au deuxième temps, ce qui est ainsi accentué au temps de prédiction.

Le même phénomène se produit en 3D sur la figure 3.15 : la simulation obtenue avec les paramètres de calibration volumique est très éloignée des observations. En effet, la tumeur réelle se développe vers une direction préférentielle tandis que la tumeur simulée croît isotropiquement.

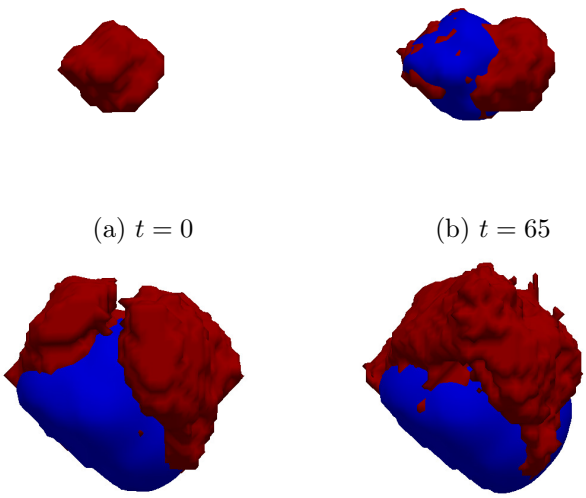
Cet exemple clinique montre les limites de notre calibrage volumique : il permet de prédire le volume tumoral à un temps futur mais la forme de la tumeur ne peut pas être estimée par ce procédé. Deux raisons expliquent cette mauvaise prédiction. La première est le fait que le modèle est simpliste : il a l'avantage de pouvoir s'appliquer aux métastases du cerveau et du poumon, mais il ne prend pas en compte les particularités de la pathologie. La seconde raison est qu'une hypothèse importante est faite dans ce procédé de calibration volumique : on suppose en effet que la vascularisation initiale est homogène spatialement. En réalité, les fibres et vaisseaux de l'organe considéré sont responsables des modifications de forme de la tumeur. Afin d'améliorer ces deux points, l'idée est d'utiliser des techniques d'assimilation de données, qui permettent à la fois de corriger la forme de la tumeur simulée, mais aussi d'apprendre et de reconstituer l'environnement tumoral en se servant des examens dont on possède. Dans notre exemple clinique 2D par exemple, l'utilisation des deux premiers examens peut permettre de deviner que la tumeur croît plus rapidement dans deux directions : vers la gauche et vers le haut. De la même manière, dans le cas 3D,

une direction semble être privilégiée. Prendre en compte ces changements de forme dans le modèle peut ainsi permettre d'améliorer les prédictions, et ainsi de donner une information plus précise aux cliniciens.



(a) $t = 0$ (b) $t = 701$ (c) $t = 1434$

FIGURE 3.14 – Comparaison entre les tumeurs observées (masse rouge) et les contours des tumeurs simulées (en vert) par la méthode du calibrage des paramètres par l'expression volumique, dans le cas de la métastate pulmonaire 2D. La forme polygonale du contour au temps $t = 0$ s'explique par le fait que la tumeur est petite et donc que la segmentation comporte peu de points.



(a) $t = 0$ (b) $t = 65$

(c) $t = 275$ (d) $t = 324$

FIGURE 3.15 – Comparaison entre les tumeurs observées (masse rouge) et les tumeurs simulées (masse bleue) par la méthode du calibrage des paramètres par l'expression volumique, dans le cas de la métastase cérébrale 3D.

## 3.4 Assimilation de données

### 3.4.1 Principe de l'assimilation de données

Le but de l'assimilation de données est de corriger la dynamique du modèle en utilisant des observations. La construction du modèle a nécessité de faire des simplifications et de négliger plusieurs aspects biologiques (particularités de la pathologie, environnement tumoral, etc...). En particulier, la structure du cerveau autour de la tumeur n'est pas prise en compte dans le modèle. Or les vaisseaux sanguins et les fibres peuvent modifier considérablement la croissance de la tumeur. Si l'on a vu dans la partie précédente que le modèle permettait d'estimer l'évolution du volume tumoral, l'estimation de la forme de la tumeur est cependant mauvaise. L'assimilation de données est un outil qui va nous permettre d'intégrer les informations de forme au modèle afin de corriger l'évolution et d'améliorer la prédiction.

Notons $\mathcal{M}$ notre modèle décrit précédemment. Notons que l'équation sur $M$ peut se réécrire sous la forme :

$$M(t) = M^0 \exp\left(\int_0^t P(s,x)ds\right). \tag{3.32}$$

On considère alors :

$$y(t) = \left(P(t,x), M(t,x), \mathbf{v}(t,x), \int_0^t P(s,x)ds, \theta(t)\right),$$

l'état étendu du système, comprenant les valeurs des différentes inconnues et des paramètres $\theta(t) = (\alpha, M_0)$. L'intérêt de prendre en compte $\int_0^t P(s,x)ds$ dans l'état étendu du système sera expliqué dans la section sur la résolution numérique de l'équation régissant $M$. Le modèle s'écrit ainsi :

$$\begin{cases} \dot{y}(t) = \mathcal{M}(\hat{y}, t), \\ y(0) = y_0. \end{cases} \tag{3.33}$$

en notant $y_0 = (P(0,x), M(0,x), \mathbf{v}(0,x), 0, \theta_0)$ la condition initiale. On remarque que dans ce modèle, la fonction $\theta(t)$ est constante égale à $\theta_0$.

Notons $z(t)$ les observations que l'on possède et qui seront utilisées afin de corriger l'état du système. Notons $\mathcal{Z}$ l'espace des observations muni des opérations classiques. Dans un premier temps, nous supposerons que ces observations sont acquises pour tout temps $t$. Nous verrons dans une partie suivante comment adapter le procédé lorsque les observations ne sont connues que ponctuellement. Comme évoqué précédemment, le modèle ne peut pas décrire exactement la trajectoire $z(t)$ à cause des simplifications de modélisation. De plus, ces observations peuvent être sujettes à des incertitudes de mesures. En particulier, $P(x,0)$

est initialisé à partir de la segmentation de la tumeur à $t = 0$, et est donc connu à une part d'incertitude près. De même, les paramètres du modèle $\theta_0$ sont initialisés à partir du calibrage volumique. Or ce calibrage ne donne qu'une estimation des paramètres, et est donc également soumis à des incertitudes. On déduit donc qu'en pratique, on a simplement un *a priori* sur $y_0$. On le décompose alors en deux parties : $y_0 = y_\bullet + \zeta$, où $y_\bullet$ est la partie connue et $\zeta$ est l'incertitude. Le problème complet se récrit donc :

$$\begin{cases} \dot{y}(t) = \mathcal{M}(\hat{y}, t), \\ y(0) = y_\bullet + \zeta. \end{cases} \tag{3.34}$$

L'assimilation de données consiste alors à utiliser les observations $z(t)$ afin d'estimer $\zeta$. Notons que $\zeta$ possède deux composantes : une qui correspond à l'incertitude sur l'état, c'est-à-dire la valeur de $P_0$, et l'autre qui correspond à l'incertitude sur les paramètres du modèle. La correction du modèle nécessite de pouvoir mesurer la différence entre l'état modélisé et l'observation. On définit donc un opérateur de discordance $D(z(t), y(t))$ qui mesure cet écart.

### 3.4.2 Approche variationnelle

L'approche variationnelle consiste à minimiser la fonction de coût $J_T(\zeta)$ définie par :

$$J_T(\zeta) = ||\zeta||^2_{\mathcal{X}} + \int_0^T ||D(z(t), y(t))||^2_{\mathcal{Z}} dt. \tag{3.35}$$

Elle comprend un premier terme qui mesure l'écart à l' *a priori* initial, ainsi qu'un second terme qui mesure l'écart aux données. Les espaces $\mathcal{X}$ et $\mathcal{Z}$ sont respectivement les espaces des conditions initiales et des observations. Ils sont munis de normes qui dépendent respectivement de la covariance de l'incertitude initiale et de la covariance des observations. Notons que le second terme dépend implicitement de $\zeta$ puisque la condition initale $y(0)$ dépend de $\zeta$. La minimisation de cette fonction de coût se fait en suivant le protocole suivant :

1. Initialisation : on pose $y_0 = y_\bullet$, soit $\zeta = 0$.

2. Simulation du modèle dans le sens direct.

3. Calcul de la fonction coût.

4. Intégration du modèle adjoint (l'adjoint du modèle linéaire tangent) à partir du résidu final, ce qui donne le gradient de la fonction coût.

5. Mise-à-jour de $\zeta$ en utilisant une méthode de descente de gradient.

Les étapes 2 à 5 se répètent jusqu'à ce que la fonction coût ait atteint une valeur inférieure à un seuil, ou qu'un nombre maximal d'étapes ne soit atteint. Notons qu'il est également possible de calculer le gradient directement, de façon approchée, en simulant $2p$ fois le modèle direct (où $p$ est le nombre de paramètres) et en calculant les dérivées partielles par différences finies. Les méthodes variationnelles nécessitent donc un nombre important d'itérations des modèles directs et adjoints, ce qui est coûteux en temps de calcul. Le schéma d'une telle méthode est représenté sur la figure 3.16.

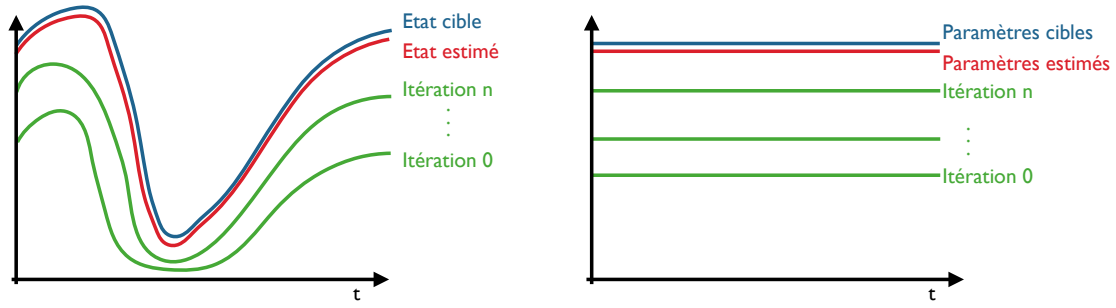



FIGURE 3.16 – Schéma des approches variationnelles en assimilation de données. Plusieurs simulations complètes sont nécessaires afin de corriger l'état et les paramètres.

Dans notre étude, nous n'utilisons pas cette approche variationnelle car elle nécessite de calculer le modèle adjoint et requiert un nombre important d'itérations des modèles directs et adjoints. Nous avons choisi d'utiliser une approche séquentielle.

### 3.4.3 Approche séquentielle

Les approches séquentielles n'effectuent pas d'itérations successives, mais corrigent pendant la simulation l'état et les paramètres. Cette correction s'effectue grâce à un retour d'informations obtenu par le calcul de l'opérateur de discordance. L'avantage de ce type d'approche est la réduction du temps de calcul - si le terme correctif se calcule rapidement - puisqu'une seule itération du modèle est nécessaire pour corriger l'état et les paramètres. Une seconde simulation effectuée avec les paramètres et l'état corrigé permet ensuite la prédiction du comportement futur. Lorsque l'objectif est simplement de prédire l'évolution future, il est également possible de laisser la première simulation continuer en arrêtant la

correction. Le schéma de la figure 3.17 montre ce type d'approche. Cela consiste donc à considérer le modèle suivant :

$$\begin{cases} \dot{\hat{y}}(t) = \mathcal{M}(\hat{y}, t) + K(D(z, \hat{y})), \\ \hat{y}(0) = y_\bullet, \end{cases} \tag{3.36}$$

où $K$ est l'opérateur de gain, également appelé filtre, qui est choisi de sorte que :

$$\lim_{t \to \infty} \hat{y} \longrightarrow y,$$

où $y$ est solution de l'équation (3.34). Un exemple de gain est $K(D(z, \hat{y})) = -\partial_{\hat{y}} D(z, \hat{y})$, ce qui est naturel puisque ce terme donne la direction de descente maximale de l'opérateur de discordance. Un autre exemple est le filtre de Kalman linéaire et ses extensions [29]. Dans le cas linéaire, il est équivalent au cas variationnel. L'inconvénient principal du filtre de Kalman est qu'il nécessite à chaque itération la résolution d'une EDP dont l'inconnue est la matrice de covariance (de taille $d$x$d$ avec $d$ le nombre de degrés de liberté), ce qui implique des grands temps de calcul.

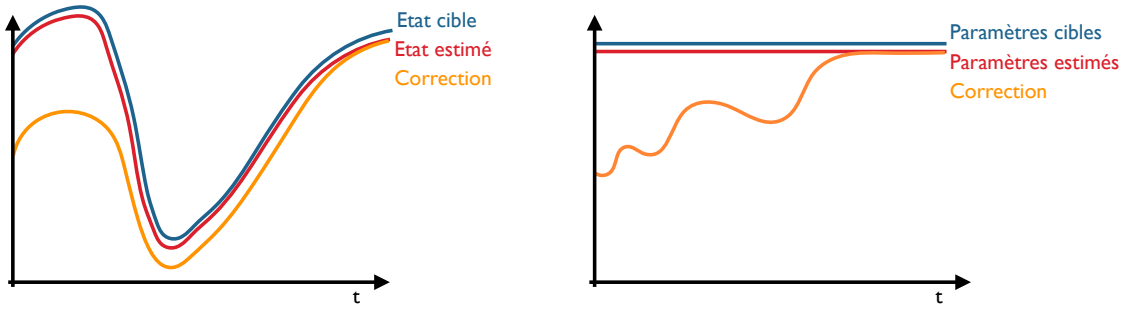



FIGURE 3.17 – Schéma des approches séquentielles en assimilation de données. L'état et les paramètres sont corrigés au cours de la simulation du modèle.

### 3.4.4 Mesure de similarité

Les méthodes d'assimilation de données nécessitent de définir une mesure de similarité, ou opérateur de discordance, pour comparer le front tumoral simulé à l'observation. La mesure que l'on choisit doit pouvoir indiquer dans quelles zones la tumeur est en avance et dans quelles zones elle est en retard par rapport à l'observation. Ainsi, une mesure simple basée sur le volume de l'erreur n'est pas satisfaisante. Comme initié dans [8] pour des modèles de réaction-diffusion, nous définissons une mesure basée sur la fonctionnelle de Chan Vese,

issue de la théorie du traitement d'images [23, 24], définie par :

$$\begin{aligned} D(z,P) &= \int_{\mathcal{B}} H(P-P_{th})(z-C_{\max}(z,P))^2 + (1-H(P-P_{th}))(z-C_{\min}(z,P))^2 dx, \\ &= \int_{\Omega_P^{\text{in}}} (z-C_{\max}(z,P))^2 dx + \int_{\mathcal{B}\setminus\overline{\Omega_P^{\text{in}}}} (z-C_{\min}(z,P))^2 dx, \end{aligned} \tag{3.37}$$

où $H$ est la fonction de Heaviside, et $C_{\max} = \max(C_1, C_2)$, $C_{\min} = \min(C_1, C_2)$ où :

$$C_1(z,P) = \frac{\displaystyle\int_{\Omega_P^{\text{in}}} z\, d\mathcal{B}}{\displaystyle\int_{\Omega_P^{\text{in}}} d\mathcal{B}}, \quad C_2(z,P) = \frac{\displaystyle\int_{\mathcal{B}\setminus\overline{\Omega_P^{\text{in}}}} z\, d\mathcal{B}}{\displaystyle\int_{\mathcal{B}\setminus\overline{\Omega_P^{\text{in}}}} d\mathcal{B}}. \tag{3.38}$$

La figure 3.18 représente comment les coefficients $C_1$ et $C_2$ sont calculés.
$C_1$ mesure la proportion de la tumeur simulée qui est correctement placée par rapport à l'observation. Lorsque $C_1 = 1$, la tumeur simulée n'a pas d'avance par rapport à l'observation.
$C_2$ mesure au contraire la proportion à l'extérieur de la tumeur qu'il reste à combler pour se rapprocher de l'observation. Lorsque $C_2 = 0$, la tumeur simulée n'a pas de retard par rapport à l'observation. Notons que lorsque la tumeur est petite par rapport au domaine de simulation, $C_2$ est petit et a peu d'influence sur la discordance.
Le premier et le second termes de l'équation (3.37) mesurent donc respectivement à quel point la tumeur simulée est en avance ou en retard par rapport à l'observation. Ce terme est minimum lorsque $P$ et $z$ ont exactement la même forme.

### 3.4.5 Observateur de Luenberger

Dans cette section, on suppose que les paramètres $\theta$ du modèle sont connus (en particulier la condition initiale $M_0$). Seule la condition initiale $P_0$ comporte des incertitudes. L'assimilation de donnée séquentielle est appliquée avec la mesure de similarité décrite ci dessus. En pratique, on peut toujours s'assurer que $C_{\max} = C_1$ et $C_{\min} = C_2$ en prenant un domaine de taille suffisamment grande. En effet, plus le domaine est grand, plus $C_2$ est petit, alors que $C_1$ n'est pas modifié. Dans la suite, nous supposerons alors $C_{\max} = C_1$ et $C_{\min} = C_2$. Nous supposons dans un premier temps que l'observation $z(x)$ ne dépend pas du temps.

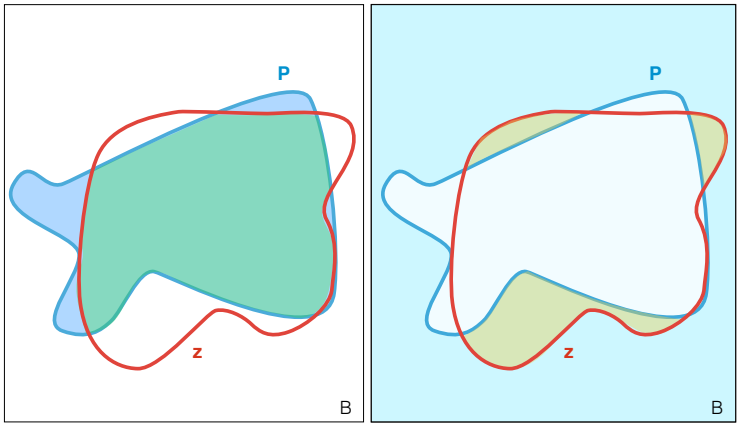



FIGURE 3.18 – Schéma pour le calcul des termes $C_1$ et $C_2$ dans la fonctionnelle de Chan Vese. $C_1$ est égal à la fraction entre l'aire de la zone verte et celle de la zone bleue. $C_2$ est la fraction entre l'aire de la zone jaune et celle de la zone bleu clair.

**Proposition 2.**

$$< -\frac{\partial D}{\partial P}, \psi >= \int_{\Gamma_P} \Big( -\big(z - C_1(z,P)\big)^2 + \big(z - C_2(z,P)\big)^2 \Big) \psi \; ds. \tag{3.39}$$

*Démonstration.* Soit $E$ l'opérateur définit par : $E(z, P, C_1(z,P), C_2(z,P)) = D(z,P)$. Alors :

$$\frac{\partial D}{\partial P} = \frac{\partial E}{\partial P} + \frac{\partial E}{\partial C_1}\frac{\partial C_1}{\partial P} + \frac{\partial E}{\partial C_2}\frac{\partial C_2}{\partial P}. \tag{3.40}$$

Or on remarque que :

$$\begin{aligned} \frac{\partial E}{\partial C_1} = -2\int_{\mathcal{B}} H(P - P_{th})\big(z - C_1(z,P)\big)dx &= -2\Big(\int_{\hat{\Omega}_P^{\text{in}}} z dx - C_1(z,P)\int_{\hat{\Omega}_P^{\text{in}}} dx\Big), \\ &= 0, \end{aligned} \tag{3.41}$$

par définition de $C_1$, et :

$$\begin{aligned} \frac{\partial E}{\partial C_1} = -2\int_{\mathcal{B}} H(P - P_{th})\big(z - C_2(z,P)\big)dx &= -2\Big(\int_{\mathcal{B}\setminus\overline{\Omega_P^{\text{in}}}} z dx - C_2(z,P)\int_{\mathcal{B}\setminus\overline{\Omega_P^{\text{in}}}} dx\Big), \\ &= 0, \end{aligned} \tag{3.42}$$

par définition de $C_2$.

En calculant formellement la dérivée de Fréchet de $D(z,P)$ par rapport à $P$ dans la direc-

tion $\psi$, on obtient donc [26] :

$$< -\frac{\partial D}{\partial P}, \psi >=< -\frac{\partial E}{\partial P}, \psi >= \int_{\Gamma_P} \Big( -\big(z - C_1(z,P)\big)^2 + \big(z - C_2(z,P)\big)^2 \Big) \psi \; ds. \quad (3.43)$$

□

En appliquant une méthode de descente de gradient, le terme choisi pour corriger l'état est donc le suivant :

$$\delta_{\hat{\Gamma}} \Big( -\big(z - C_1(z,P)\big)^2 + \big(z - C_2(z,P)\big)^2 \Big).$$

Dans le cas général où $z$ dépend du temps, le même terme correctif sera gardé [27]. Nous montrerons dans la section suivante que cela permet bien de corriger l'état. En notant avec un chapeau les solutions de l'observateur, et par $\hat{\Gamma}$ et $\hat{\Omega}^{\text{in}}$ respectivement le front tumoral et la région tumorale associée, on obtient le modèle suivant :

$$\begin{cases} \partial_t \hat{P} + \operatorname{div}(\hat{v}\,\hat{P}) &= \hat{M}\hat{P} + \lambda\, \delta_{\hat{\Gamma}} \Big( -\big(z - C_1(z,\hat{P})\big)^2 + \big(z - C_2(z,\hat{P})\big)^2 \Big), & \mathcal{B} \\ \partial_t \hat{S} + \operatorname{div}(\hat{v}\,\hat{S}) &= \lambda\, \delta_{\hat{\Gamma}} \Big( -\big(1 - z - C_1(1-z,\hat{S})\big)^2 + \big(1 - z - C_2(1-z,\hat{S})\big)^2 \Big), & \mathcal{B} \\ \partial_t \hat{M} &= -\alpha \hat{M}\hat{P}, & \mathcal{B} \\ \hat{v} &= -\nabla \hat{\pi}, & \mathcal{B} \end{cases}$$

où $\lambda$ est une constante positive. En utilisant le fait que :

$$\big(1 - z - C_1(1-z,S)\big) = -\big(z - C_2(z,P)\big),$$

et

$$\big(1 - z - C_2(1-z,S)\big) = -\big(z - C_1(z,P)\big),$$

on obtient en sommant les deux premières équations du modèle et le fait que $P + S = 1$ :

$$\nabla \cdot \hat{v} = -\Delta \hat{\pi} = \hat{M}\hat{P}.$$

Sur la figure 3.19, on distingue 3 types de corrections :

— Le point 1 correspond au cas où la tumeur simulée $P$ est en retard sur l'observation $z$. A ce point, on a $z = 1$ et le terme de correction vaut donc :

$$\text{Correction}_1 = -(1 - C_1)^2 + (1 - C_2)^2 = (C_1 - C_2)(2 - C_1 - C_2). \quad (3.44)$$

Or par définition de $C_1$ et $C_2$, on a :

$$\begin{aligned} 0 \leqslant C_1 \leqslant 1, \\ 0 \leqslant C_2 \leqslant 1, \end{aligned} \tag{3.45}$$

donc $(2 - C_1 - C_2) \geqslant 0$. Puisque $(C_1 - C_2) \geqslant 0$, le terme correctif est positif et l'ajout de matière le long du front va contribuer à la diminution de l'erreur entre $P$ et $z$.

— Le point 2 correspond au cas inverse : la tumeur simulée est en avance par rapport à l'observation. En ce point, on a $z = 0$. Le terme de correction vaut alors :

$$\text{Correction}_2 = -(0 - C_1)^2 + (0 - C_2)^2 = -(C_1 - C_2)(C_1 + C_2). \tag{3.46}$$

Puisque $C_1 + C_2 \geqslant 0$ et $C_1 \geqslant C_2$, ce terme correctif est négatif. Le retrait de matière en ce point 2 permet donc à la tumeur simulée de réduire son avance et de recoller progressivement à l'observation.

— Le point 3 correspond au cas idéal où la frontière de la tumeur simulée et de l'observation sont confondues. En ce point, $z = 0.5$ et le terme correctif vaut alors :

$$\text{Correction}_3 = -(0.5 - C_1)^2 + (0.5 - C_2)^2 = -(C_1 - C_2)(1 - C_1 - C_2). \tag{3.47}$$

Le terme $(C_1 - C_2)$ est positif, mais le terme $(1 - C_1 - C_2)$ peut être positif ou négatif selon la configuration globale. Le terme de correction n'est donc pas nul *a priori*. Cela signifie que même lorsque la frontière cible est atteinte en un point, la correction continue de se faire pour corriger la forme globale. L'équilibre n'est atteint que lorsque $C_1 = 1$ et $C_2 = 0$.

### 3.4.6 Justifications mathématiques

L'équation sur $P$ du modèle de croissance tumorale est une équation hyperbolique non linéaire. Puisque le domaine est borné est que la tumeur croît, il n'est pas possible d'avoir une existence globale des solutions. L'existence locale ainsi que l'unicité de la solution sont montrées dans la thèse [28]. Le temps d'existence de la solution dépend de la taille du domaine et des paramètres de croissance. Par contre rien n'a été prouvé sur l'existence et l'unicité du modèle de l'observateur.

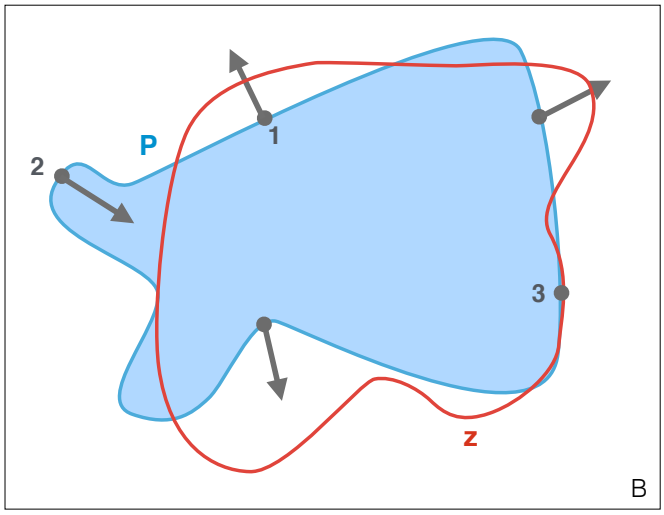



FIGURE 3.19 – Effets de la correction d'état sur le contour de $P$, étant donné l'observation $z$.

Nous souhaitons justifier que le terme de correction d'état permet bien de rapprocher l'état simulé de l'état observé. Le modèle considéré dans cette section est légèrement simplifié : l'équation régissant la vascularisation de la tumeur $M$ ne comporte plus de terme en $P$. Cela signifie biologiquement que la consommation d'oxygène se fait dans tout le domaine, et pas seulement là où les cellules proliférantes se trouvent. On obtient alors le modèle suivant :

$$\left\{\begin{array}{rcll} \partial_t P + \nabla\cdot(\mathbf{v}\,P) &=& MP, & \mathcal{B} \\ \nabla\cdot\mathbf{v} &=& MP, & \mathcal{B} \\ \partial_t M &=& -\alpha M, & \mathcal{B} \\ \mathbf{v} &=& -\nabla\pi, & \mathcal{B} \end{array}\right. \tag{3.48}$$

avec $\alpha > 0$ et les conditions initiales :

$$\left\{\begin{array}{rcl} P(0,x) &=& P_0(x) = \hat{P}_0(x) + \zeta_X, \\ M(0,x) &=& m_0, \end{array}\right. \tag{3.49}$$

avec $m_0 > 0$. Le modèle avec correction d'état correspondant s'écrit :

$$\left\{\begin{array}{rcll} \partial_t \hat{P} + \nabla\cdot(\hat{v}\,\hat{P}) &=& \hat{M}\hat{P} - \lambda f(P, \hat{P}-P), & \mathcal{B} \\ \nabla\cdot\hat{v} &=& \hat{M}\hat{P}, & \mathcal{B} \\ \partial_t \hat{M} &=& -\alpha\hat{M}, & \mathcal{B} \\ \hat{v} &=& -\nabla\hat{\pi}, & \mathcal{B} \end{array}\right. \tag{3.50}$$

avec des conditions initiales :

$$\begin{cases} \hat{P}(0,x) &= \hat{P}_0(x), \\ \hat{M}(0,x) &= m_0, \end{cases} \tag{3.51}$$

avec des conditions de Dirichlet homogènes au bord du domaine pour $\pi$ et $\hat{\pi}$.
Notons $P(t,x)$ la solution de l'équation (3.48) avec les conditions initiales (3.49), et par $\hat{P}(t,x)$ la solution de l'équation (3.50) avec conditions initiales (3.51). Le terme de correction est ici un terme de correction quelconque $f$ et nous allons voir quelles sont les conditions sur $f$ qui assurent une correction effective.

Notons $[0,T]$ la période de suivi clinique. Nous supposons que le domaine $\mathcal{B} \subset \mathbb{R}^n$ est suffisamment grand pour que $\forall t \in [0,T]$, $\text{supp}(P(t,x)) \subset \mathcal{B}$ et $\text{supp}(\hat{P}(t,x)) \subset \mathcal{B}$. En pratique, la dimension de l'espace considéré est $n = 2$ ou $n = 3$.

Notons que grâce à la simplification de modèle, $M_0$ est constant spatialement et donc $\forall t \in [0,T]$, $M(t,x) = \hat{M}(t,x) = m_0 \mathrm{e}^{-\alpha t}$ sur tout le domaine.

Notons $\tilde{P} = \hat{P} - P$, $\tilde{v} = \hat{v} - v$ et $\tilde{\pi} = \hat{\pi} - \pi$ les erreurs.
Notons $g(P, \tilde{P}) = \tilde{P} f(P, \tilde{P})$.

**Proposition 3.** *Soit $\varepsilon > 0$ et soit $T_1 \in [0,T]$ le temps du dernier examen clinique. Supposons que $f$ satisfasse les hypothèses suivantes $\forall (u,v) \in [0,1]^2$ :*

$$\begin{cases} f(u, v-u) &= -f(u, 1-v-u), \\ f(u, 1-u) &\geqslant 0, \\ f(u, -u) &\leqslant 0. \end{cases} \tag{3.52}$$

*Alors il existe une constante $C_0 \in \mathbb{R}$ telle que, si $||P_0||_{H^s} \leqslant C_0$ avec $s = \frac{n}{2} + 2$, alors :*

- *si $\forall (u,v) \in [0,1]\times]0,1]$, $g(u,v) > 0$ alors il existe un gain $\lambda_\varepsilon$ tel que $||\tilde{P}(T_1,x)||_{L^2} \leqslant \varepsilon$ avec une décroissance exponentielle.*
- *si $\int_{[0,1]^2} g(u,v)dudv > 0$ alors il existe $\lambda_\varepsilon$ tel que $||\tilde{P}(T_1,x)||_{L^2} \leqslant \varepsilon$ avec une décroissance linéaire.*

La démonstration utilise les lemmes et le théorème suivants, qui sont démontrés en annexe :

**Lemme 1.** *Supposons que $P_0 \in H^s$ avec $s > \frac{n}{2}$ et soit $C = 2||P_0||_{H^s}$. Alors il existe $t_0$ tel que le compact $K_C = \{P(t,x), ||P||_{L^\infty_{t_0}, H^s} \leqslant C\}$ soit stable par $\Phi \circ \Psi$.*

**Lemme 2.** *Soit $t_0$ le temps fixé par le lemme précédent. Il existe $t_1 \in [0, t_0]$ tel que l'application $\Phi \circ \Psi$ soit contractante sur $K_C = \{P(t,x), ||P||_{L^\infty_{[0,t_1]},H^s} \leqslant C\}$.*

**Théorème 1.** *Supposons que $P_0 \in H^s$ avec $s > \frac{n}{2}$. Il existe un temps $t_2$ dépendant de $||P_0||_{H^s}$ et une constante $C_{t_2}$ tels que la solution au système (3.48), notée $P$ satisfasse : $||P||_{L^\infty_{[0,t_2]},H^s} < C_{t_2}$.*

*Démonstration.* (proposition 3)
La première condition sur $f$, $\forall (u,v) \in [0,1]^2$,

$$f(u, v-u) = -f(u, 1-v-u), \tag{3.53}$$

assure que l'équation sur $\hat{P}$ et celle sur $\hat{S} = 1 - \hat{P}$ sont bien corrigées de façon opposée et assure ainsi cette expression $\nabla \cdot \hat{v} = \hat{M}\hat{P}$.
De plus, puisque :

$$\begin{cases} f(u, 1-u) & \geqslant 0, \\ f(u, -u) & \leqslant 0, \end{cases} \tag{3.54}$$

on a $\forall t \in [0,T]$, $\forall x \in \mathcal{B}$, $\hat{P}(t,x) \in [0,1]$. Notons que l'on a également $P \in [0,1]$ comme vu précédemment.

En soustrayant la première ligne du système (3.48) à la première ligne du système (3.50), on obtient :

$$\begin{aligned} \partial_t \tilde{P} &= -\nabla \cdot (\hat{v}\hat{P}) + \nabla \cdot (vP) + M\tilde{P} - \lambda f(P, \tilde{P}), \\ &= -\hat{v} \cdot \nabla \hat{P} - M\hat{P}^2 + \mathbf{v} \cdot \nabla P + MP^2 + M\tilde{P} - \lambda f(P, \tilde{P}), \end{aligned} \tag{3.55}$$

ce qui donne, en utilisant $\hat{v} \cdot \nabla \hat{P} = \hat{v} \cdot \nabla \tilde{P} + \hat{v} \cdot \nabla P$ :

$$\partial_t \tilde{P} = -\hat{v} \cdot \nabla \tilde{P} - \tilde{v} \cdot \nabla P + M\tilde{P}(1 - P - \hat{P}) - \lambda f(P, \tilde{P}). \tag{3.56}$$

Puisque l'on a, par la formule de la divergence :

$$\begin{aligned} \int_{\mathcal{B}} \hat{v} \cdot \nabla \tilde{P} \tilde{P} dx &= \frac{1}{2} \int_{\mathcal{B}} \hat{v} \cdot \nabla (\tilde{P}^2) dx, \\ &= -\frac{1}{2} \int_{\mathcal{B}} M\hat{P}\tilde{P}^2 dx, \end{aligned} \tag{3.57}$$

on obtient :

$$\begin{aligned}\int_{\mathcal{B}} \partial_t \tilde{P}\tilde{P}dx = &- \int_{\mathcal{B}} \tilde{v}\cdot\nabla P\tilde{P}dx \\ &+ M\int_{\mathcal{B}} \tilde{P}^2\left(1-\frac{3}{2}P-\frac{1}{2}\tilde{P}\right)dx - \lambda\int_{\mathcal{B}} f(P,\tilde{P})\tilde{P}dx,\end{aligned} \tag{3.58}$$

et enfin,

$$\frac{1}{2}\frac{d}{dt}\left(||\tilde{P}||_2^2\right) = -\int_{\mathcal{B}} \tilde{v}\cdot\nabla P\tilde{P}dx + M\int_{\mathcal{B}} \tilde{P}^2\left(1-\frac{3}{2}P-\frac{1}{2}\tilde{P}\right)dx - \lambda\int_{\mathcal{B}} f(P,\tilde{P})\tilde{P}dx. \tag{3.59}$$

Notons alors

$$g(P,\tilde{P}) = f(P,\tilde{P})\tilde{P}. \tag{3.60}$$

Soit $t_0$ un réel. Notons $\Phi$ et $\Psi$ les applications suivantes, avec $M(t) = m_0 \mathrm{e}^{-\alpha t}$ :

$$\begin{aligned}&\Phi : (L^\infty_{[0,t_0]}, H^{s+1}) \to (L^\infty_{[0,t_0]}, H^s) \\ &\qquad u \longmapsto P \text{ tel que } \begin{cases} \partial_t P + \nabla\cdot(u,P) = MP, \\ P(0,x) = P_0, \end{cases}\end{aligned} \tag{3.61}$$

$$\begin{aligned}&\Psi : (L^\infty_{[0,t_0]}, H^{s}) \to (L^\infty_{[0,t_0]}, H^{s+1}) \\ &\qquad P \longmapsto \mathbf{v} = -\nabla\pi \text{ tel que } \begin{cases} -\Delta\pi = MP \text{ sur } \Omega, \ \cdot \\ \pi = 0 \text{ sur } \partial\Omega. \end{cases}\end{aligned} \tag{3.62}$$

Posons $C_0 = \frac{1}{2\sqrt{TC'}}$, , où $C'$ est la constante définie dans la preuve du théorème précédent. Supposons que $||P_0||_{H^s} \leqslant C$ avec $s = \frac{n}{2}+2$, alors il existe d'après le théorème précédent $C_T \in \mathbb{R}$ tel que $||P(t,x)||_{L^\infty_T, H^s} \leqslant C_T$. En utilisant le plongement continu $H^{s-1} \hookrightarrow L^\infty$, on obtient l'existence d'une constante $C_\nabla$ vérifiant $\forall t \in [0,T]$, $||\nabla P||_{L^\infty} \leqslant C_\nabla$. On obtient alors :

$$\int_{\mathcal{B}} \tilde{v}\cdot\nabla P\tilde{P}dx \leqslant C_\nabla ||\tilde{v}||_{L^2}||\tilde{P}||_{L^2}. \tag{3.63}$$

Puisque $\Delta\tilde{\pi} = -M\tilde{P}$, cela donne :

$$\begin{aligned}\int_{\mathcal{B}} (\nabla\pi)^2 dx &\leqslant m_0 \exp(-\alpha t)\left|\int_{\mathcal{B}} \tilde{P}\pi dx\right|, \\ &\leqslant m_0\exp(-\alpha t)||\tilde{P}||_{L^2}||\tilde{\pi}||_{L^2}.\end{aligned} \tag{3.64}$$

Puisque $\tilde{\pi}$ satisfait des conditions au bord de Dirichlet homogènes, l'inégalité de Poincaré donne :

$$||\tilde{\pi}||_{L^2} \leqslant C_{\mathcal{B}}||\nabla\tilde{\pi}||_{L^2}, \tag{3.65}$$

où $C_{\mathcal{B}}$ ne dépend que de la taille du domaine $\mathcal{B}$, et puisque $\tilde{v} = -\nabla\tilde{\pi}$, on obtient :

$$||\tilde{v}||_{L^2}^2 \leqslant C_{\mathcal{B}} M ||\tilde{P}||_{L^2} ||\tilde{v}||_{L^2}, \tag{3.66}$$

et donc,

$$||\tilde{v}||_{L^2} \leqslant C_{\mathcal{B}} M ||\tilde{P}||_{L^2}. \tag{3.67}$$

Finalement, on obtient :

$$\int_{\mathcal{B}} \tilde{v} \cdot \nabla P \tilde{P} dx \leqslant C_{\nabla} C_{\mathcal{B}} M ||\tilde{P}||_{L^2}^2. \tag{3.68}$$

En combinant l'équation (3.68) et l'équation (3.69), on obtient :

$$\frac{d}{dt}\left(||\tilde{P}||_2^2\right) \leqslant \int_{\mathcal{B}} 2\tilde{P}^2 M \left(1 - \frac{3}{2}P - \frac{1}{2}\tilde{P} + C_{\nabla} C_{\mathcal{B}}\right) - \lambda g(P, \tilde{P}) dx. \tag{3.69}$$

— Dans le cas où $g(P, \tilde{P}) > 0$ lorsque $\tilde{P} \neq 0$, c'est-à-dire lorsque la correction se fait en tous les points où il est nécessaire de le faire, et dans le bon sens, alors on pose :

$$\lambda_{\varepsilon} = \max_{t\in[0,T_1], \tilde{P}(t,x)\neq 0} \frac{\tilde{P}^2 M \left(1 - \frac{3}{2}P - \frac{1}{2}\tilde{P} + C_{\nabla} C_{\mathcal{B}}\right) + \eta_{\varepsilon}}{g(P, \tilde{P})}, \tag{3.70}$$

où $\eta_{\varepsilon}$ est choisi tel que :

$$||\tilde{P}(T_1, x)||_2^2 \leqslant ||\tilde{P}_0||_2^2 \exp(-2\eta_{\varepsilon} T_1) \leqslant \varepsilon, \tag{3.71}$$

ce qui donne :

$$\eta_{\varepsilon} \geqslant \frac{1}{2T_1} \ln\left(\frac{||\tilde{P}_0||_2^2}{\varepsilon}\right). \tag{3.72}$$

Une telle valeur de $\lambda_{\varepsilon}$ permet donc une convergence exponentielle de l'erreur vers 0.

— Dans le cas où la seule condition sur $g$ est la suivante $\int_{\mathcal{B}} g(P, \tilde{P}) > 0$, c'est-à-dire lorsque la correction se fait dans le bon sens, mais pas nécessairement partout où il est nécessaire de corriger, alors on définit :

$$\lambda_{\varepsilon} = \max_{t\in[0,T_1]} \frac{\eta_{\varepsilon} + \int_{\mathcal{B}} \tilde{P}^2 M \left(1 - \frac{3}{2}P - \frac{1}{2}\tilde{P} + C_{\nabla} C_{\mathcal{B}}\right) dx}{\int_{\mathcal{B}} g(P, \tilde{P}) dx}, \tag{3.73}$$

avec $\eta_{\varepsilon}$ choisi tel que :

$$||\tilde{P}(T_1, x)||_2^2 \leqslant ||\tilde{P}_0||_2^2 - \eta_{\varepsilon} T_1 \leqslant \varepsilon, \tag{3.74}$$

c'est-à-dire :

$$\eta_\varepsilon \geqslant \frac{||\tilde{P}_0||_2^2 - \varepsilon}{T_1}. \tag{3.75}$$

Une telle valeur de $\lambda_\varepsilon$ assure une converge linéaire de l'erreur vers 0.

□

Dans notre modèle, la valeur de $f$ choisie remplit les conditions du deuxième cas, ce qui assure que l'on peut trouver une valeur de $\lambda$ qui assure une erreur plus petite que $\varepsilon$ au temps $T_1$, avec une décroissance linéaire de l'erreur.

Notons que les fonctions qui assurent une convergence exponentielle de l'erreur vers 0 ne sont pas utilisables dans notre modèle. En effet, on ne souhaite corriger que dans la zone de la tumeur simulée afin d'assurer une croissance réaliste de la tumeur. Par exemple, avec une fonction $f(P, \tilde{P}) = \tilde{P}$, la correction d'état crée de la matière partout là où $P = 1$ et $\hat{P} = 0$, ce qui n'a plus de sens biologiquement.

**Remarque 1.** *La preuve peut être adaptée afin de montrer la convergence de* $||\tilde{P}||_{H^s}$ *vers* 0*, mais il faut alors des conditions de régularité sur la fonction* $f$*.*

Sur la figure 3.20, l'évolution de $||\tilde{P}||_2^2$ est représentée pour deux corrections : celle utilisée dans notre étude et une correction $f(P, \tilde{P}) = \tilde{P}$. On observe bien une décroissance linéaire avec la correction de Luenberger et une décroissance exponentielle avec l'autre correction, comme attendu. On remarque que l'erreur ne tend pas complètement vers 0 et recroît légèrement en temps long. Cela s'explique ici par le maillage de l'espace de calcul et à la discrétisation de la masse du Dirac (voir section sur la résolution numérique) : la correction ne permet d'obtenir la forme cible qu'à une maille près. Puisque la tumeur croît, son aire aussi et l'erreur due au maillage est aussi croissante.

### 3.4.7 Filtre de Kalman réduit à l'espace des paramètres

#### Estimation paramétrique

Le filtre de Luenberger décrit précédemment permet de corriger l'état, mais les paramètres ont été supposés connus. En pratique, on ne connaît qu'un *a priori* des paramètres et il est nécessaire de les corriger afin de coller aux observations. Dans notre cas du modèle de croissance de tumeurs, cet *a priori* est obtenu à partir de la calibration volumique. Or, dans

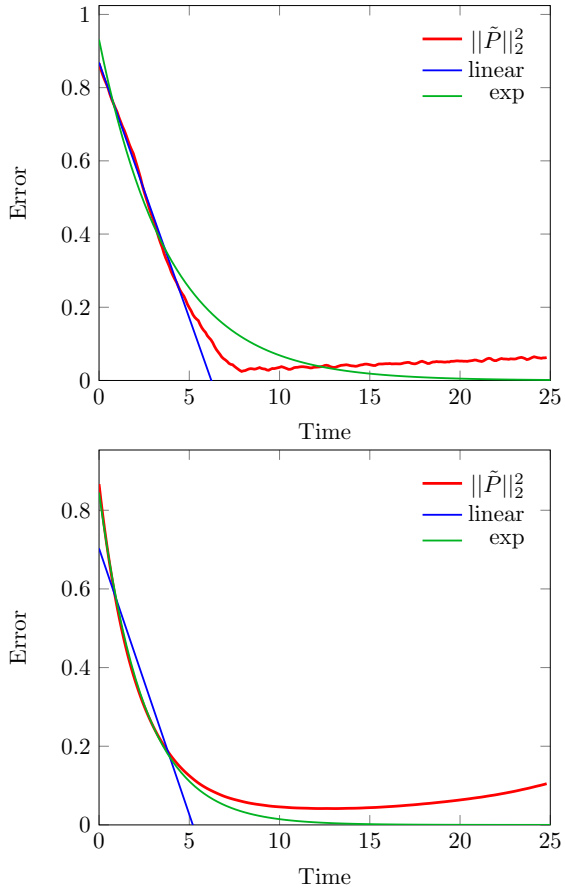



FIGURE 3.20 – Evolution de l'erreur $||\tilde{P}||_2^2$ (en rouge) en fonction du temps, pour deux types de correction d'état : première ligne avec la correction de Luenberger, seconde ligne avec la correction $f(P, \tilde{P}) = \tilde{P}$. Afin de comparer les deux décroissances, les régressions sur l'intervalle $[0, 5]$ avec un modèle linéaire (en bleu) et exponentiel (en vert) sont représentées.

cette calibration, le champ de vascularisation initial $M_0$ est supposé constant spatialement. En pratique, ce champ dépend du milieu, en particulier des vaisseaux sanguins, et il est donc nécessaire de corriger ce champ et de pouvoir l'estimer spatialement si l'on souhaite prédire la forme de la tumeur.
Dans cette section, nous allons commencer par introduire les filtres de Kalman, puis nous parlerons de filtre de Kalman réduit aux paramètres.

### Filtres de Kalman

Le filtre de Kalman [29] vise à corriger l'état étendu du système 3.33. Initialement, l'algorithme de Kalman a été introduit dans le cas linéaire, c'est-à-dire dans le cas où l'opérateur du modèle $\mathcal{M}$ ainsi que l'opérateur de discordance $D$ sont linéaires. Le gain de Kalman $K$

est alors choisi de sorte à minimiser la fonctionelle :

$$J_T(\zeta) = ||\zeta||^2_{\mathcal{X}} + \int_0^T ||D(z(t), y(t))||^2_{\mathcal{Z}} dt,$$

comme dans le cas variationnel, mais de manière séquentielle. Ce gain $K$ dépend des métriques de l'espace des observations et de l'espace de correction, c'est-à-dire qu'il dépend de la confiance aux données et de la confiance aux *a priori*. On a alors le modèle suivant :

$$\begin{cases} \dot{\hat{y}}(t) = \mathcal{M}(\hat{y}, t) + K(D(z, \hat{y})), \\ \hat{y}(0) = y_\bullet, \end{cases} \tag{3.76}$$

où on rappelle que $y_\bullet$ est l'*a priori* sur la condition initiale $y_0$.
Le gain $K$ est générique et le filtre de Kalman est donc valable quel que soit $\mathcal{M}$ ou $D$. Il dépend de la matrice de covariance de taille $d$x$d$ (où $d$ est le nombre de degrés de liberté) qui doit être inversée à chaque itération ce qui rend le calcul très coûteux. Dans le cas général où les opérateurs ne sont pas linéaires, une approximation de ce filtre est calculé en linéarisant les opérateurs non linéaires. Ce nouveau filtre s'appelle le filtre de Kalman étendu (Extended Kalman Filter [30]). Cependant, le calcul de ces opérateurs tangents peut s'avérer coûteux et nous allons utiliser ici une autre alternative. Le principe du filtre de Kalman particulaire (Unscented Kalman Filter [31]) est d'utiliser des particules, appelées *sigma points*, qui vont remplacer les opérateurs tangents. Ces particules ont pour but d'évaluer une itération du modèle avec un jeu de paramètres légèrement décalé du jeu initial afin de déterminer quelle correction des paramètres permet de réduire l'écart aux observations. Ces *sigma points* sont calculés à partir de l'état existant, et à partir des *sigma points* unitaires sp$[i]$ qui vérifient :

$$\begin{aligned} \sum_{i=1}^{r} \kappa_i \text{sp}[i] &= 0, \\ \sum_{i=1}^{r} \kappa_i \text{sp}[i] \cdot \text{sp}[i]^T &= I_p, \end{aligned} \tag{3.77}$$

où $p$ est le nombre de paramètres à estimer, $r$ est le nombre de particules, $I_p$ est la matrice identité de taille $p$ et les $\kappa_i$ sont des coefficients associés aux *sigma points* vérifiant $\sum_{i=1}^{p} \kappa_i = 1$. Les particules sont donc d'espérance nulle, et de matrice de covariance unitaire. Les *sigma points* points les plus couramment utilisés en pratique sont les deux suivants :

— *sigma points* points canoniques : $r = 2p$ points, avec des coefficients $\kappa_i = \frac{1}{2p}$ pour

tout $i \in [\![1, r]\!]$ :

$$\mathrm{sp}[i] = \begin{cases} \sqrt{p} e_i, & \text{pour } 1 \leqslant i \leqslant p \\ -\sqrt{p} e_{i-p}, & \text{pour } p+1 \leqslant i \leqslant 2p. \end{cases}$$

où $e_i$ est le $i^{\text{ème}}$ vecteur de la base canonique de $\mathbb{R}^p$.

— *sigma points* du simplexe : $r = p+1$ points, avec des coefficients $\kappa_i = \frac{1}{p+1}$ pour tout $i$ :

$$\mathrm{sp}[i] = \sqrt{p} S_{p+1}^{(i)},$$

où les vecteurs $S_{p+1}^{(i)}$ sont les colonnes des matrices $[S_{p+1}^*]$ :

$$\begin{cases} [S_1^*] &= \left( -\frac{1}{\sqrt{2\kappa}} \quad \frac{1}{\sqrt{2\kappa}} \right), \ \kappa = \frac{p}{p+1} \\ [S_d^*] &= \begin{pmatrix} & & & 0 \\ & [S_{d-1}^*] & & \vdots \\ & & & 0 \\ \frac{1}{\sqrt{\kappa d(d+1)}} & \cdots & \frac{1}{\sqrt{\kappa d(d+1)}} & \frac{-d}{\sqrt{\kappa d(d+1)}} \end{pmatrix}, \ 2 \leqslant d \leqslant p. \end{cases}$$

Un exemple pour $p = 2$ de tels *sigma points* unitaires est représenté sur la figure 3.27.

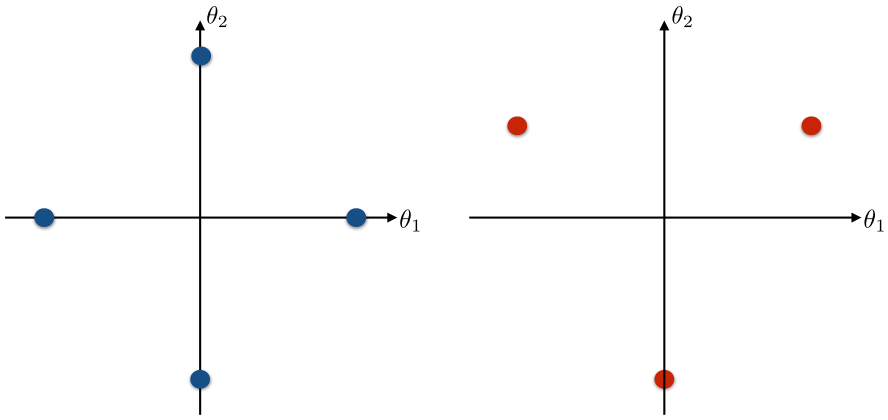



FIGURE 3.21 – Tracé des *sigma points* unitaires canoniques (à gauche) et du simplexe (à droite) lorsque $p = 2$.

Les *sigma points* canoniques sont les points qui n'ont qu'une composante non nulle. Ces points évaluent donc l'effet des $p$ paramètres indépendemment, dans les deux directions. $2p$ points sont donc nécessaires pour parcourir toutes les directions. Les *sigma points* du simplexe sont les sommets d'un simplexe de taille $p$, centré en 0. Il n'y a donc que $p+1$ particules qui permettent d'estimer les directions privilégiées. Puisque moins de particules doivent être lancées, moins d'itérations du modèle sont nécessaires, et les *sigma points*

du simplexe ont donc l'avantage de réduire le coût de calcul par rapport aux $2p$ *sigma points* canoniques. Nous verrons cependant dans une prochain section que les *sigma points* canoniques permettent d'obtenir des résultats plus précis et ont donc été préférés.
Le schéma de la figure 3.22 montre les différentes étapes de l'estimation de paramètres grâce aux particules. Les particules sont créées à partir de l'état étendu et des *sigma points* unitaires décrits ci-dessus. Chaque particule subit alors une itération du modèle, avec les paramètres propres à la particule. Le nouvel état intermédiaire atteint par chaque particule est alors comparé aux observations. Les variations paramétriques permettant de réduire la discordance entre la simulation et l'observation sont celles gardées lors de la correction paramétrique. Un nouvel état étendu est alors obtenu, avec les nouveaux paramètres estimés et avec l'état général reconstitué à partir des états des particules.

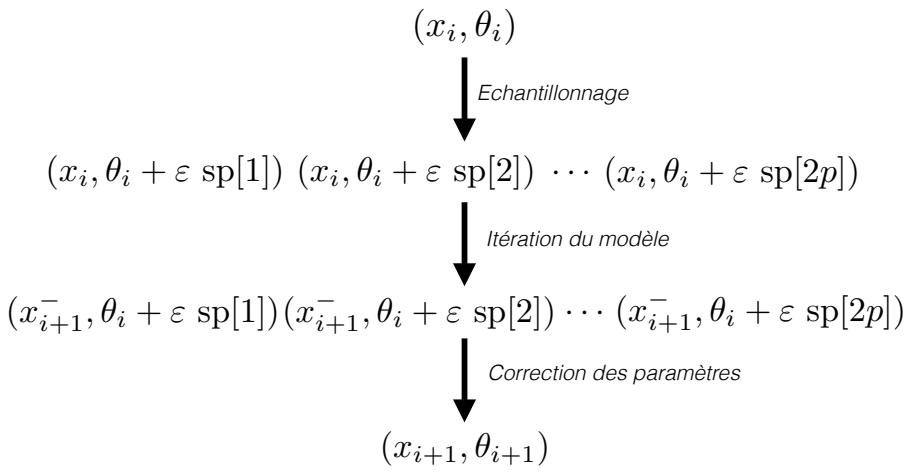



FIGURE 3.22 – Schéma de la correction paramétrique par filtre de Kalman particulaire. La variable $\varepsilon$ correspond à l'écart-type des paramètres.

L'itération de la correction paramétrique dans un cas simple, avec $p = 2$ et des *sigma points* canonique, donne le schéma de la figure 3.23.

### Estimation jointe état-paramètres

Les filtres décrits ci-dessus corrigent l'état étendu du système, c'est-à-dire l'état ainsi que les paramètres. Ils présentent l'avantage de ne pas dépendre de la dynamique et ainsi d'être utilisables pour n'importe quel modèle à corriger. Cependant, en pratique, la dimension de l'espace est trop grande pour appliquer ce genre de méthodes, qui sont trop coûteuses en temps de calcul et en mémoire. L'idée est dont d'utiliser un filtre réduit (Reduced Order

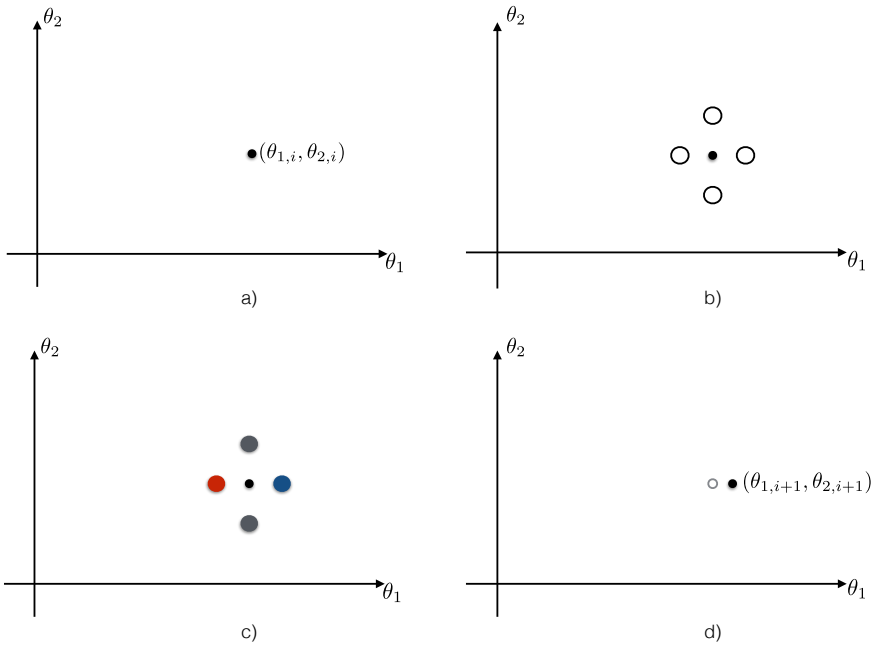



FIGURE 3.23 – Exemple de correction paramétrique pour $p = 2$ et avec des *sigma points* canoniques. A partir de l'état $(\theta_{1,i}, \theta_{1,i})$ (a), les 4 *sigma points* sont construits (b). Une itération du modèle est lancée avec les paramètres de chacune des particules, et la discordance de chaque particule avec les observations est calculée (c). Les points rouges, bleus et gris symbolisent respectivements une discordance forte, faible et moyenne. Le nouvel état $(\theta_{1,i+1}, \theta_{1,i+1})$ est alors calculé en corrigeant l'état initial dans la direction qui minimise la discordance (d).

Unscented Kalman Filter) qui ne corrige que les paramètres. L'état est ainsi corrigé par le filtre de Luenberger présenté précédemment [14], qui a l'avantage d'être moins coûteux, mais qui dépend du choix de la mesure de similarité. Puisque l'espace des paramètres est de dimension bien inférieure à celle de l'espace de l'état, le coût de calcul est drastiquement réduit. Afin d'effectuer cette réduction, notons $p$ le nombre de paramètres à estimer, et $d$ la dimension de l'espace de l'état. On a donc $p << d$. La matrice de covariance globale de la méthode UKF, de taille $dxd$ est alors décomposée sous la forme : $LU^{-1}L^{-1}$, où $U$ est une matrice inversible de taille $p$ et $L$ l'opérateur d'extension. Le fait de travailler sur $U$ et $L$ au lieu de la matrice complète permet de réduire drastiquement les coûts de calcul.

La correction de l'état et des paramètres se fait alors de façon jointe, avec deux filtres différents : un filtre de Luenberger pour l'état, et un filtre réduit RoUKF pour les paramètres. On a alors le schéma de la figure 3.24 d'estimation jointe suivant.

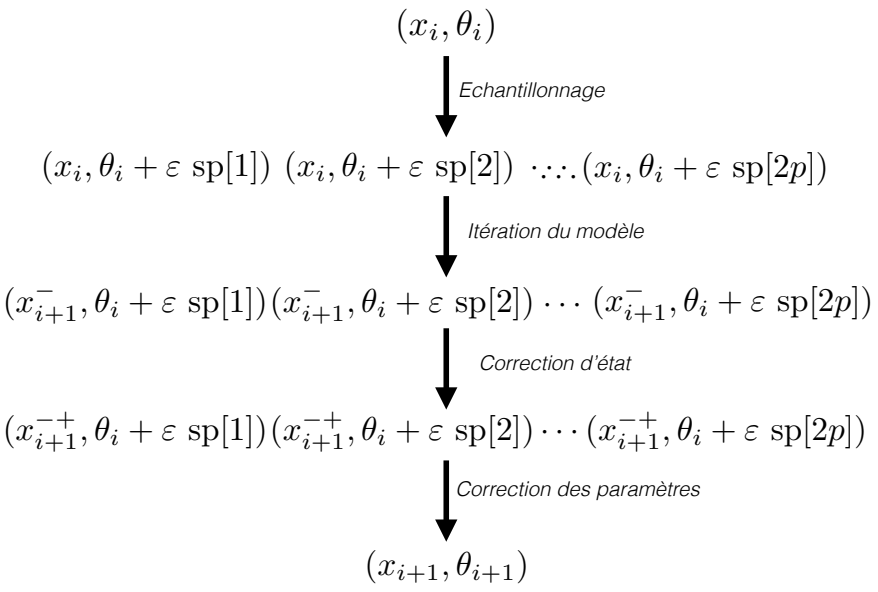



FIGURE 3.24 – Schéma de correction des paramètres grâce à la création de *sigma points*. La variable $\varepsilon$ correspond à l'écart-type des paramètres.

Il ne diffère du schéma précédent qu'à une étape près, qui est la correction d'état de chaque particule par le filtre de Luenberger.

Notons $\mathrm{Cov}_\theta$ la matrice de covariance initiale des paramètres, et $M$ la matrice de covariance des observations. Rappelons que l'on note $\mathcal{M}$ le modèle, $D$ la discordance, $z$ les observations, sp le vecteur de*sigma points* et $\mathrm{D}_\kappa$ la matrice diagonale comportant les coefficients associés aux *sigma points*. Nous noterons avec une étoile la matrice obtenue par concaténation des vecteurs pour $i = 1...r$. La matrice $C = \sqrt{U^{-1}}$ correspond à la décomposition de Choleski de $U^{-1}$. L'algorithme RoUKF utilisé quand on considère les *sigma points* du simplexe est proposé ci-dessous [15].

**Remarque 2.** *Pour l'utilisation des sigma points canoniques, le rang de la matrice de covariance peut ne pas rester égal à p. Dans ce cas là, une décomposition en valeurs singulières doit être appliquée sur la matrice de covariance ce qui complexifie l'algorithme [16]. Pour le moment, nous avons considéré le même algorithme pour les deux types de sigma points, mais il est clair que c'est une comparaison à effectuer à l'avenir.*

**Initialisation**

$$\begin{cases} U_0 &= \mathrm{Cov}_\theta^{-1} \\ L_n^X &= 0 \\ L_n^\theta &= \mathbb{1} \end{cases}$$

**Itération** 1. Echantillonnage

$$\left\{ \begin{array}{lcl} C_n & = & \sqrt{U_n^{-1}} \\ \begin{pmatrix} x_n^{[i]+} \\ \theta_n^{[i]+} \end{pmatrix} & = & \begin{pmatrix} x_n^{+} \\ \theta_n^{+} \end{pmatrix} + \begin{pmatrix} L_n^X \\ L_n^\theta \end{pmatrix} C_n^T \mathrm{sp}^{[i]} \text{ pour } i = 1...r. \end{array} \right.$$

2. Itération du modèle

$$\left\{ \begin{pmatrix} x_{n+1}^{[i]-} \\ \theta_{n+1}^{[i]-} \end{pmatrix} = \mathcal{M} \begin{pmatrix} x_n^{[i]+} \\ \theta_n^{[i]+} \end{pmatrix} \text{ pour } i = 1...r. \right.$$

3.Correction d'état

$$\left\{ \begin{array}{lcl} x_{n+1}^{[i]+-} & = & x_{n+1}^{[i]+-} + \lambda D_{n+1}(z_{n+1}, x_{n+1}^{[i]+-}) \text{ pour } i = 1...r, \\ x_{n+1}^{+-} & = & \sum_{i=1}^r \kappa_i x_{n+1}^{[i]+-}. \end{array} \right.$$

4. Correction des paramètres

$$\left\{ \begin{array}{lcl} L_{n+1}^X & = & x_{n+1}^{[*]+-} \mathrm{D}_\kappa \mathrm{sp}^* \\ L_{n+1}^\theta & = & \theta_{n+1}^{[*]+-} \mathrm{D}_\kappa \mathrm{sp}^* \\ \Gamma_{n+1} & = & D(z_{n+1}, x_{n+1}^{[*]+-}) \mathrm{D}_\kappa \mathrm{sp}^* \\ U_{n+1} & = & 1 + \Gamma_{n+1}^T M_{n+1} \Gamma_{n+1} \\ x_{n+1}^{+} & = & x_{n+1}^{+-} + L_{n+1}^X U_{n+1}^{-1} \Gamma_{n+1}^T M_{n+1} D_{n+1}(z_{n+1}, x_{n+1}^{+-}) \\ \theta_{n+1}^{+} & = & \theta_{n+1}^{+-} + L_{n+1}^\theta U_{n+1}^{-1} \Gamma_{n+1}^T M_{n+1} D_{n+1}(z_{n+1}, x_{n+1}^{+-}). \end{array} \right.$$

Remarquons que dans cet algorithme, la correction paramétrique se retrouve modifiée par le filtre d'état. Notons également que les termes $\mathrm{Cov}_\theta$ et $M$, de covariance des paramètres et des observations respectivement, correspondent à la formulation discrète des normes des espaces $\mathcal{X}$ et $\mathcal{Z}$ de l'approche variationnelle. C'est l'inversion de la matrice $U$ qui coûte cher en temps de calcul, d'où l'intérêt de la réduction.

## 3.5 Résolution numérique

Contrairement à la section théorique précédente, nous ne supposons plus que $P_0 \in H^s$. Dans cette section et en pratique, nous définissons une tumeur initiale $P_0$ qui vaut 1 à l'intérieur de la tumeur et 0 à l'extérieur.

### 3.5.1 Discrétisation du Dirac

Le terme de correction d'état comporte le terme $\delta_{\hat{\Gamma}}$ qui permet de ne corriger que sur le front de la tumeur simulée. Numériquement, la masse de Dirac est approchée par la fonction suivante [32] :

$$\frac{1}{\epsilon}\zeta\left(\frac{\phi(x,t)}{\epsilon}\right)|\nabla\phi|,$$

où la fonction lisse $\zeta$ est définie par :

$$\zeta(y) = \begin{cases} \frac{1}{2}(1+\cos(\pi y)) & \text{si } |y|<1, \\ 0 & \text{sinon,} \end{cases}$$

et où $\phi(x,t)$ est la fonction redistanciée de $\phi^0(x,t) = 2P(x,t) - 1$ (afin d'obtenir une fonction signée avec l'interface tumorale représentée par le niveau 0). La fonction $\zeta$ est représentée sur la figure 3.25.

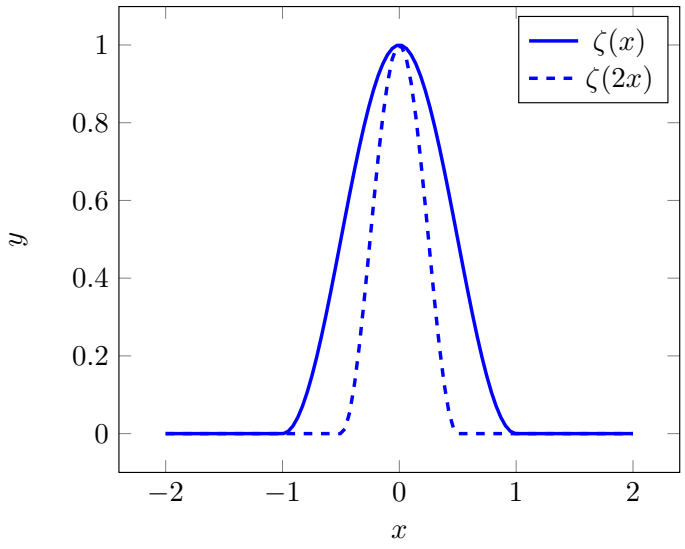



FIGURE 3.25 – Tracés des fonctions $\zeta(x)$ et $\zeta(2x)$.

La méthode décrite dans [32] ne peut pas être utilisée directement sur $\phi_0$ ici car le gradient

de $P$ peut être très élevé sur le front $\hat{\Gamma}$. La redistanciation au temps $t$ est effectuée en résolvant l'équation suivante sur $\Phi(x,t,T)$ :

$$\begin{aligned} \frac{\partial \Phi}{\partial T} &= \mathrm{sgn}(\phi^0)(1-|\nabla\Phi|), \\ \Phi(x,t,0) &= \phi^0(x,t). \end{aligned} \tag{3.78}$$

La méthode numérique utilisée est un schéma décentré amont (*upwind*) d'ordre 1 comme expliqué dans [33]. Les caractéristiques de l'équation 3.78 se déplacent à une vitesse 1, ce qui signifie que la fonction $\Phi(x,t,T_0)$ est une fonction distance signée pour tous les points à une distance $T_0$ de l'interface. Puisque la redistanciation est coûteuse en temps, et puisque seulement le front de la tumeur simulée $\hat{\Gamma}$ doit être suffisamment lisse, la redistanciation n'est effectuée que pendant un temps fixe $T_{\max} = 5\ \max(dx, dy, dz)$, où $dx$, $dy$ et $dz$ sont les dimensions d'une maille. On obtient alors $\phi(x,t) = \Phi(x,t,T_{\max})$.

### 3.5.2 Interpolation des données

Le modèle de correction d'état et de paramètre fait intervenir une mesure de similarité qui est calculée à partir des observations en tout temps $t$. En pratique, ces données ne sont pas une séquence continue d'images mais une succession de plusieurs examens (3 ou 4) espacés de plusieurs mois. Il serait possible de ne corriger que lorsque l'on possède des données [34], mais le nombre d'examens disponibles est trop petit dans notre cas pour que cette méthode soit efficace. **Afin de calculer la discordance entre la tumeur simulée et la tumeur observée, nous allons reconstruire les observations intermédiaires à partir des examens acquis**. Supposons que l'on dispose d'examens aux temps $t_1$ et $t_2$ et que l'on souhaite corriger l'état au temps $t \in [t_1, t_2]$. Afin d'interpoler $z(t)$, les observations $z(t_1)$ et $z(t_2)$ sont redistanciées afin d'obtenir $\Phi_1$ et $\Phi_2$, les distances signées au front tumoral. On définit alors $\Phi(t)$ par :

$$\Phi(t) = \eta(t)\Phi_1 + (1-\eta(t))\Phi_2, \tag{3.79}$$

où $\eta(t)$ est à valeurs dans $[0,1]$ et correspond au poids que l'on souhaite donner à $t_1$ par rapport à $t_2$. La manière la plus simple d'interpoler l'image est de considérer une interpolation linéaire :

$$\eta(t) = \frac{t_2 - t}{t_2 - t_1}.$$

Cependant, la croissance tumorale n'est pas linéaire, comme on l'a vu en introduction dans la section sur la modélisation par modèle EDO, et cette interpolation n'est donc

pas réaliste. **L'idée est donc d'utiliser l'équation sur le volume tumoral présenté à la section sur la calibration volumique**. Rappelons que lorsque $M_0$ est constant spatialement, de valeur $m_0$, alors le volume tumoral se calcule explicitement :

$$V(t) = V_0 \left(1 + \frac{m_0}{m_0 - \alpha} \left(\exp(m_0 - \alpha)t - 1\right)\right). \tag{3.80}$$

Même si le volume tumoral ne respecte pas exactement cette évolution, en particulier parce que l'hypothèse "$M_0$ constant" n'est pas toujours vérifiée, cette équation est une bonne approximation de la croissance du volume. En particulier, elle est plus réaliste qu'une interpolation linéaire. De plus, la calibration volumique permet d'estimer les paramètres $m_0$ et $\alpha$ intervenant dans cette égalité. Nous définissons donc :

$$\Phi(t) = \frac{V(t_2) - V(t)}{V(t_2) - V(t_1)}\Phi_1 + \frac{V(t) - V(t_1)}{V(t_2) - V(t_1)}\Phi_2. \tag{3.81}$$

Puisque $\Phi_1$ et $\Phi_2$ sont respectivements les distances signées des contours de $z_1$ et $z_2$, le contour de $z(t)$ que l'on souhaite interpoler correspond à la ligne de niveau 0 de $\Phi(t)$. On définit donc :

$$z(t) = \begin{cases} 1 & \text{si} \Phi(t) > 0, \\ 0 & \text{sinon.} \end{cases} \tag{3.82}$$

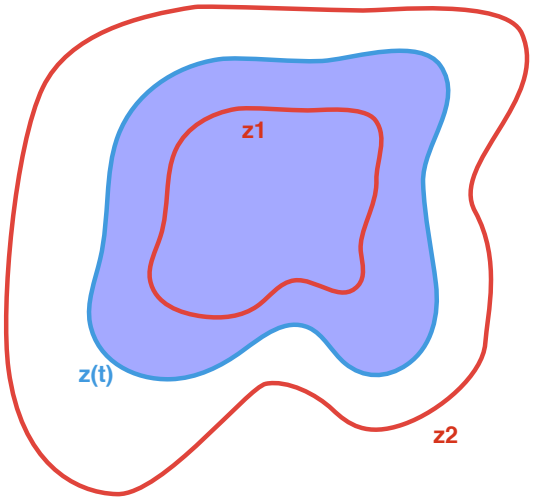



FIGURE 3.26 – Schéma de l'interpolation de l'observation $z(t)$ connaissant $z_1$ and $z_2$.

### 3.5.3 Choix des *sigma points*

L'algorithme du filtre de Kalman d'ordre réduit particulaire nécessite de définir des particules, appelées *sigma points*. Deux types de *sigma points* ont été définis précédemment. La figure 3.27 rappelle la construction de ces particules lorsque $p = 2$.

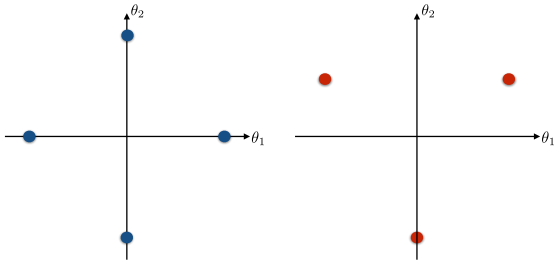



FIGURE 3.27 – Tracé des *sigma points* unitaires canoniques (à gauche) et du simplexe (à droite) lorsque $p = 2$.

Les *sigma points* du simplexe ont l'avantage d'être de dimension $p + 1$ où $p$ est le nombre de paramètres à estimer, contre $2p$ pour les *sigma points* canoniques. Cependant, l'inconvénient de ces *sigma points* du simplexe est que l'ordre des paramètres à estimer modifie les *sigma points* et modifie ainsi l'estimation paramétrique. En particulier, un paramètre peut évoluer même s'il n'influe pas du tout sur l'état, et donc sur la discordance. Dans notre modèle, ce cas se présente lorsqu'on estime $M_0$ spatialement : une zone de $M$ pas atteinte par la tumeur ne doit pas voir son paramètre modifié lors de l'estimation paramétrique. Dans l'exemple suivant à 2 paramètres, nous allons voir pourquoi l'ordre des paramètres a une importance dans le cas des *sigma points* du simplexe, et pas pour les *sigma points* canoniques.

Supposons que seuls deux paramètres sont à estimer : $\theta_1$ et $\theta_2$. D'après la définition des *sigma points* du simplexe (définis à la section 3.4.7), la matrice associée aux trois particules unitaires est de la forme suivante :

$$\mathrm{sp}^* = \begin{pmatrix} a & -a & 0 \\ b & b & -2b \end{pmatrix} \text{ avec } a, b \in \mathbb{R}. \tag{3.83}$$

A partir de l'état $n$ de l'état, trois itérations du modèle sont lancées avec les paramètres associés à ces trois particules. Notons $\mathrm{Disc}(a, b)$ la discordance de la particule obtenue à partir de la particule unitaire $(a, b)$. Supposons que le premier paramètre $\theta_1$ a une influence sur l'état, et donc sur la discordance, mais pas le second paramètre $\theta_2$. Nous avons donc $A := \mathrm{Disc}(a, b) \neq B := \mathrm{Disc}(-a, b) \neq C := \mathrm{Disc}(0, -2b)$. Dans le schéma de correction jointe

état-paramètre décrit ci-dessus, la matrice de discordance des particules se calcule par la formule :

$$\Gamma_{n+1} = \text{Disc}^* \text{D}_\kappa \text{sp}^*, \tag{3.84}$$

où $\text{Disc}^*$ correspond à la ligne des discordances des particules. Puisque les scalaires $\kappa_i = \frac{1}{p+1}$ sont tous égaux dans le cas de *sigma points* du simplexe, on obtient :

$$\Gamma_{n+1} = \frac{1}{p+1} \text{Disc}^* \text{sp}^*. \tag{3.85}$$

On a alors :

$$\Gamma_{n+1} = \begin{pmatrix} A & B & C \end{pmatrix} \begin{pmatrix} a & b \\ -a & b \\ 0 & -2b \end{pmatrix} = \Big( a(A-B), b(A+B-2C) \Big).$$

Puisque le terme $b(A+B-2C)$ n'est pas nécessairement nul, le second paramètre $\theta_2$ va évoluer dans la correction paramétrique même s'il n'a pas d'influence sur la discordance. Supposons maintenant que $\theta_2$ a une influence sur la discordance, mais pas $\theta_1$. On obtient alors $A := \text{Disc}(a,b) = \text{Disc}(-a,b)$ and $A \neq C := \text{Disc}(0,-2b)$. Cela donne :

$$\Gamma_{n+1} = \begin{pmatrix} A & A & C \end{pmatrix} \begin{pmatrix} a & b \\ -a & b \\ 0 & -2b \end{pmatrix} = \Big( 0, b(A+B-2C) \Big).$$

Dans ce cas, puisque le premier terme de $\Gamma_{n+1}$, est nul, le paramètre $\theta_1$ n'évolue pas à ce stade de la correction paramétrique. Cela est cette fois cohérent avec le fait que $\theta_1$ n'a pas d'influence sur la discordance. L'ordre d'écriture des paramètres $\theta_1$ et $\theta_2$ modifie donc le résultat de la correction paramétrique. Ce phénomène s'explique par le fait que les points du simplexe comportent des composantes non nulles selon plusieurs directions et les paramètres associés sont donc corrélés entre eux. Cela ne se produit pas dans le cas de *sigma points* canoniques, puisque cette fois chaque paramètre est traité de manière indépendante. Dans le cas de deux paramètres, on a en effet :

$$\text{sp}^* = \begin{pmatrix} a & 0 & -a & 0 \\ 0 & b & 0 & -b \end{pmatrix} \text{ avec } a, b \in \mathbb{R}. \tag{3.86}$$

et en notant $A := \text{Disc}(a,0)$, $B := \text{Disc}(0,b)$, $C := \text{Disc}(-a,0)$ et $D := \text{Disc}(0,-b)$, on ob-

tient :

$$\Gamma_{n+1} = \begin{pmatrix} A & B & C & D \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & b \\ -a & 0 \\ 0 & -b \end{pmatrix} = \Big( a(A-C), b(B-D) \Big) .$$

ce qui montre que lorque le paramètre $\theta_1$ n'est pas influent, alors $A = C$ et le premier terme est nul, et la même chose se produit avec le second paramètre. Deux solutions sont possibles pour contourner ce problème :

— utiliser les *sigma points* canoniques, même si le temps de calcul est plus long.

— ordonner les paramètres selon leur influence présumée, afin que les paramètres influents ne modifient pas les paramètres non influents.

La méthode retenue est la première : il est en effet difficile d'estimer un ordre d'influence des paramètres. Afin de limiter le temps de calcul, l'échantillonnage et l'itération du modèle des particules canoniques sont lancés en parallèle.

### 3.5.4 Résolution de l'équation sur $M$

Rappelons l'équation sur la vascularisation $M$ :

$$\partial_t M = -\alpha M P. \tag{3.87}$$

Cette équation est résolue par un schéma exponentiel classique dans le cas du modèle sans correction : le passage du temps $t_{i+1}$ au temps $t_i$ se fait par le schéma :

$$M^{i+1} = M^i \exp(P^i dt). \tag{3.88}$$

Cependant, on remarque qu'alors la vascularisation initiale $M_0$ n'est utilisée qu'au premier pas de temps. Puisque l'on souhaite corriger ce terme $M_0$, il est nécessaire qu'à tout temps $t_i$, $M^i$ soit calculé à partir de $M_0$. Le schéma de résolution de l'équation devient alors le schéma suivant :

$$M^{i+1} = M^0 \exp\left( \int_{t=0}^{t_i} P(t,x) dt \right). \tag{3.89}$$

Cela permet alors de mettre à jour $M$ lorsque $M_0$ est corrigé. Puisque $\int_{t=0}^{t_i} P(t,x)dt$ évolue au cours du temps, il fait partie de l'état étendu du sytème.

## 3.6 Validation de la méthode sur données synthétiques

### 3.6.1 Correction d'état

Afin de valider la correction d'état par le filtre de Luenberger, nous utilisons des données synthétiques. Sur la figure 3.29, les observations sont représentées en rouge à trois temps $t_0 = 0$, $t_1 = 20$ et $t_2 = 40$. Comme précédemment, les corrections se font à partir des temps $t_0$ et $t_1$, le dernier temps $t_2$ permettant de comparer l'observation à la simulation et ainsi de valider ou non la correction du système. La vascularisation initiale $M_0$ n'est plus supposée constante spatialement mais varie selon que la tumeur se trouve dans la matière blanche ou dans la matière grise du cerveau. En effet, [35, 36] montrent que la propagation tumorale dans la matière blanche est plus rapide que dans la matière grise. Les segmentations de la matière blanche et de la matière grise utilisées pour construire les cas synthétiques sont représentées sur la figure 3.28.

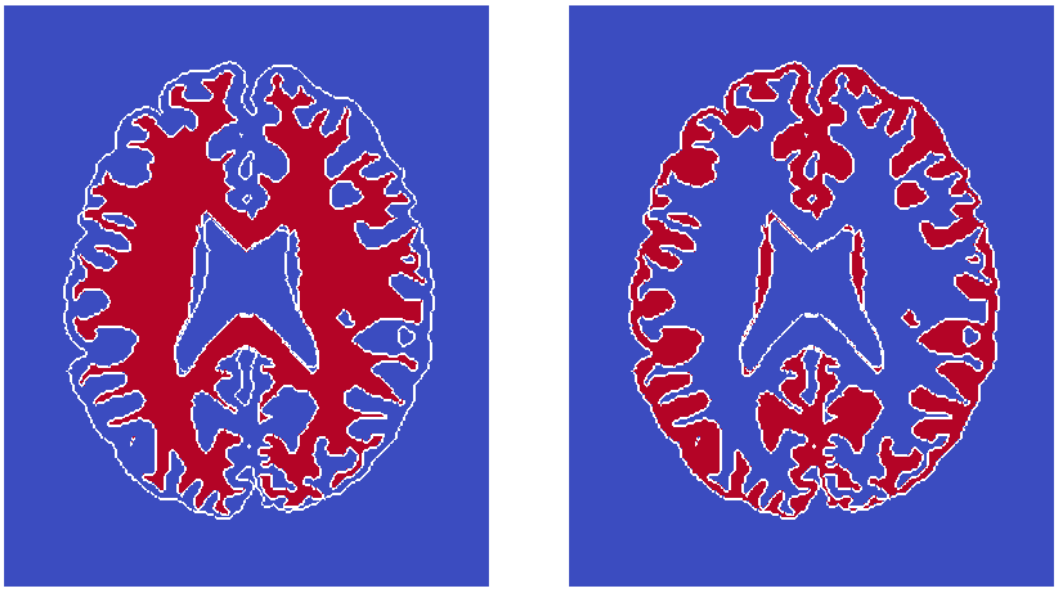
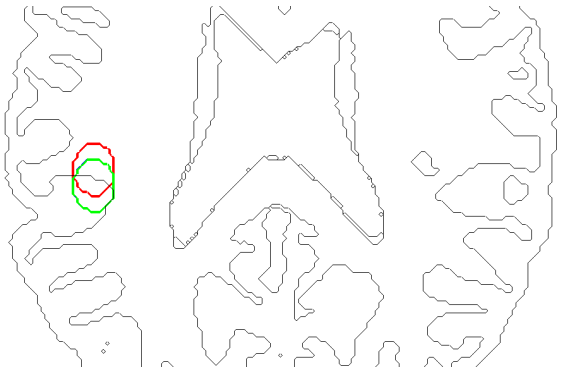


FIGURE 3.28 – Segmentation de la matière blanche (à gauche) et grise (à droite) en 2D. Le contour du cerveau est représenté en blanc. Au centre du cerveau, on observe les ventricules.

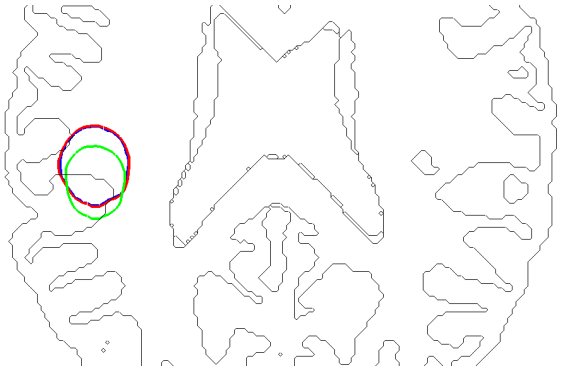
Puisque le but de cette partie est de corriger l'état et pas les paramètres, le même jeu de paramètres est utilisé pour toutes les simulations : $(\alpha, m_0, m_1) = (0.02, 0.1, 0.03)$, où $m_0$ est la valeur de $M_0$ dans la matière blanche, plus élevée que $m_1$, la valeur dans la matière

grise. Sur la simulation cible de la figure 3.29, on observe que la zone tumorale située dans la matière grise croît plus lentement que la zone située dans la matière blanche, ce qui est conforme à notre choix d'initialisation des paramètres. Afin de valider la correction d'état, on choisit de légèrement décaler la condition initiale de la condition initiale cible. Cette nouvelle condition initiale est représentée en vert au temps $t_0$. La simulation verte est celle obtenue lorsqu'aucune correction d'état n'est effectuée. Puisque la tumeur est initialement décalée dans la matière grise, sa croissance est plus lente que la simulation cible, et on remarque aux temps $t_1$ et $t_2$ que le volume et la forme tumorale restent éloignés de la cible. Nous souhaitons alors utiliser la méthode de correction d'état pour corriger cette erreur de condition initiale et retrouver une simulation proche de la simulation cible.

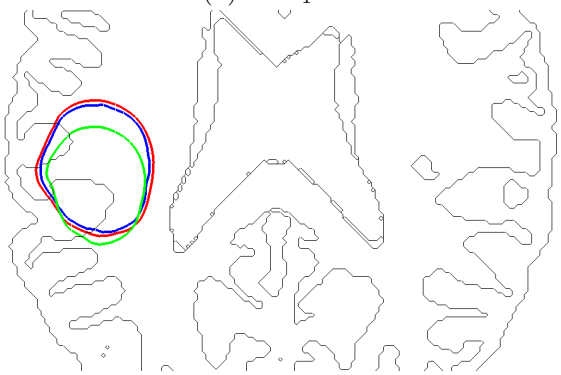
La simulation bleue est obtenue en corrigeant l'état entre $t_0$ et $t_1$, et en partant de la même condition initiale verte. Le paramètre de correction d'état choisi est $\lambda = 0.03$. Dans une section suivante, nous verrons comment choisir $\lambda$ afin de corriger efficacement l'état sans que la correction ne soit prépondérante sur le modèle. Sur la figure, on remarque qu'au temps $t_1$, le contour bleu est superposé au contour cible. Lorsque cette correction est relâchée, le contour bleu reste très proche du contour cible, comme on peut le voir au temps $t_2$. La légère erreur de contour observée à $t_2$ provient du fait que la matrice extracellulaire $M_0$ n'a pas été consommée de la même manière dans les simulations rouges et bleues, due à la différence de condition initiale.



(a) $t = t_0$



(b) $t = t_1$



(c) $t = t_2$

FIGURE 3.29 – Comparaison des formes tumorales dans trois cas : la cible (en rouge), la simulation sans correction d'état (en vert) et avec correction d'état (en bleu) au temps initial $t_0$ (ligne a), au temps intermédiaire $t_1$ (ligne b) et au temps final $t_2$ (ligne c). La correction d'état ne se fait qu'entre les temps $t_0$ et $t_1$. Les conditions initiales $P_0$ des cas avec et sans correction sont décalées de la condition initiale $P_0$ de la simulation cible. Le jeu de paramètres utilisé est le même pour les trois simulations : $(\alpha, m_0, m_1) = (0.02, 0.1, 0.03)$, avec $m_0$ et $m_1$ les valeurs dans la matière blanche et grise respectivement. La simulation avec correction est obtenue avec un paramètre $\lambda = 0.03$.

La figure 3.30 montre l'évolution de la discordance avec la cible, dans les cas avec et sans correction d'état. Dans le cas avec correction, la discordance est décroissante entre $t_0$ et

$t_1$, ce qui est cohérent avec le fait que l'on corrige l'état, puis remonte légèrement entre $t_1$ et $t_2$. Dans le cas sans correction, la discordance diminue un peu également, à cause de la croissance tumorale, mais la simulation reste tout de même éloignée de la simulation cible.

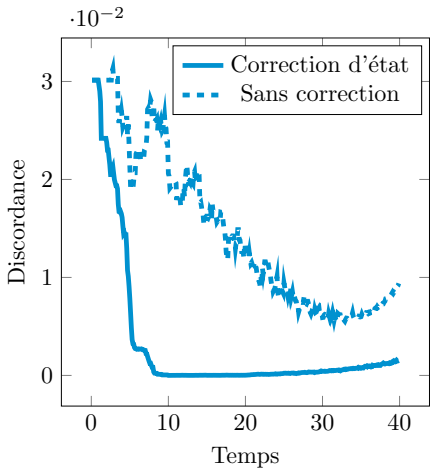



FIGURE 3.30 – Evolution de la discordance dans le cas d'une simulation avec correction d'état (ligne bleue) et sans correction d'état (pointillés bleus). La correction d'état se fait entre $t_0 = 0$ et $t_1 = 20$.

Ce premier résultat valide ainsi la méthode de correction d'état par filtre de Luenberger.

### 3.6.2 Correction jointe état-paramètres

Dans cette section, le but est de valider la méthode de correction de paramètre par filtre de Kalman. Les conditions initiales seront supposées parfaitement connues. Comme vu dans la section sur la calibration volumique, les paramètres $\alpha$ et $M_0$ sont corrélés et donc non identifiables. Dans un premier temps, nous allons supposer $\alpha$ connu afin de valider le modèle sur un cas synthétique. Dans un second temps, $\alpha$ ne sera plus supposé connu et il sera initialisé à partir du calibrage volumique. Dans un troisième temps, $\alpha$ est ajouté à la liste des paramètres à estimer par le filtre de Kalman. Ces trois méthodes sont testées afin de valider le principe d'étude, mais aussi de voir quelle est la meilleure qui sera appliquée aux données réelles.

#### Cas $\alpha$ connu

Dans cette sous-section, nous supposons $\alpha$ connu : la même valeur de ce paramètre sera utilisée dans la simulation cible ainsi que dans les simulations de corrections de paramètre,

à savoir $\alpha = 0.001$. Les seuls paramètres à estimer sont donc $m_0$ et $m_1$, les valeurs de $M_0$ dans la matière blanche et dans la matière grise, comme dans le cas précédent. La simulation cible représentée en rouge sur la figure 3.32 est obtenue avec les paramètres cibles $(m_0^t, m_1^t) = (0.08, 0.01)$. On extrait alors quatre données aux temps $t_0 = 0$, $t_1 = 10$, $t_2 = 20$ et $t_3 = 30$ et seuls les trois premiers temps sont utilisés pour la correction. Le calibrage volumique est appliqué à ces trois premières données, sauf que la valeur de $\alpha$ est forcée à 0.001. On obtient alors le faisceau de la figure 3.31, qui donne une erreur volumique relative de 10%.

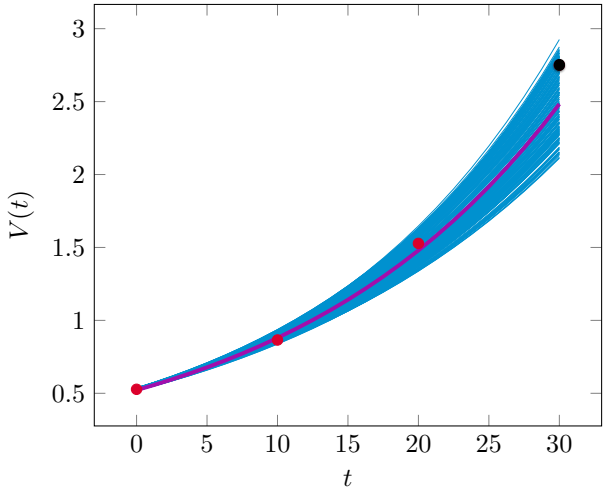
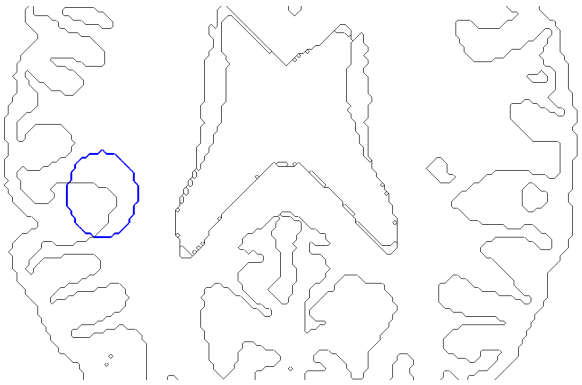
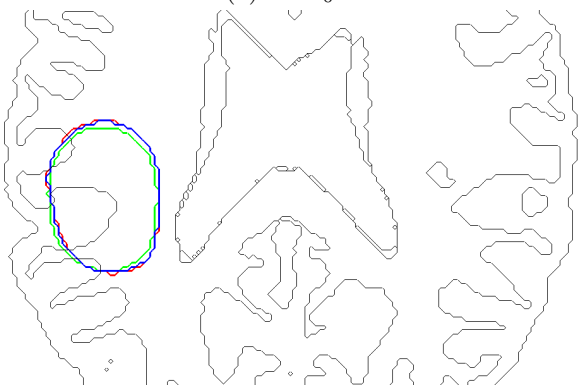



FIGURE 3.31 – Calibration volumique dans le cas synthétique où $\alpha$ est connu et vaut 0.001. Les trois premières données (points rouges) sont utilisées pour calibrer le modèle et la dernière donnée (point noir) est gardée pour être comparée à la prédiction. La courbe violette est celle donnant le volume médian parmi les jeux acceptés. Elle est obtenue avec $m_{\text{unif}} = 0.052$.

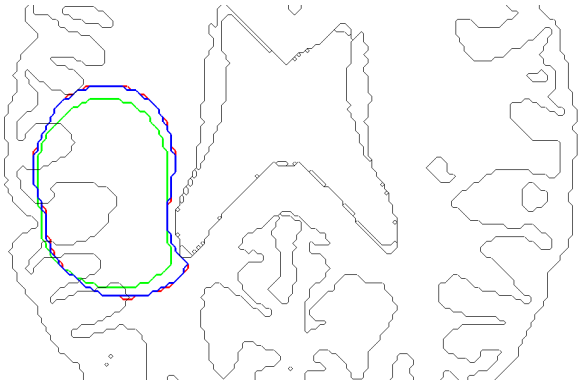
Le calibrage sert également à trouver une valeur $m_{\text{unif}}$ qui sert d'*a priori* à l'estimation. Ici, $m_{\text{unif}} = 0.052$. La simulation verte représentée sur la figure 3.32 est la simulation obtenue avec $M_0 = m_{\text{unif}}$. Puisque $M_0$ est alors uniforme, la simulation est éloignée de la simulation cible obtenue avec une distinction de valeur dans la matière blanche et grise. Le filtre de Kalman permet alors de corriger les paramètres et donne la simulation bleue. Au temps $t_2$, la simulation est confondue avec la cible, et cela reste le cas au temps $t_3$ : les paramètres ont été corrigés efficacement. Ces figures sont obtenues avec une faible correction d'état $\lambda = 0.0001$.



(a) $t = t_0$



(b) $t = t_2$



(c) $t = t_3$

FIGURE 3.32 – Comparaison des formes tumorales dans trois cas : la cible (en rouge), la simulation sans correction d'état (en vert) et avec correction d'état (en bleu) au temps initial $t_0$ (ligne a), au temps intermédiaire $t_2$ (ligne b) et au temps final $t_3$ (ligne c). La correction d'état et de paramètres ne se fait qu'entre les temps $t_0$ et $t_1$. Le jeu de paramètres cible est le jeu $(m_0, m_1) = (0.08, 0.01)$, avec $m_0$ et $m_1$ les valeurs dans la matière blanche et grise respectivement. La simulation verte est obtenue avec le jeu uniforme $(m_0, m_1) = (0.052, 0.052)$ calculé par calibrage volumique. La simulation avec correction est obtenue avec une erreur de covariance 0.01, avec $\lambda = 1e{-}4$ et avec des paramètres initialisés avec le jeu uniforme.

L'évolution de la discordance représentée sur la figure 3.33 confirme la nécessité de la correction, puisque la discordance croît rapidement sans correction, alors qu'elle reste très faible avec correction.

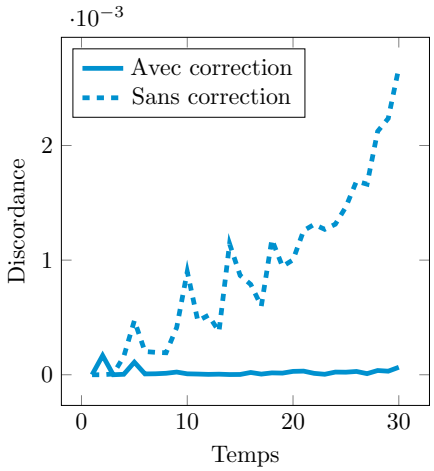



FIGURE 3.33 – Evolution de la discordance dans le cas d'une simulation avec correction état-paramètres (ligne bleue) et sans correction (pointillés bleus). La correction jointe état-paramètres se fait entre $t_0 = 0$ et $t_2 = 20$.

De plus, la méthode est validée par la figure 3.34 où les paramètres $(m_0, m_1)$ sont initialisés à $m_{\text{unif}}$ et convergent finalement vers le jeu de paramètres cible $(m_0^t, m_1^t)$. Cette convergence permet de valider la méthode jointe état-paramètres dans le cas où $\alpha$ est connu.

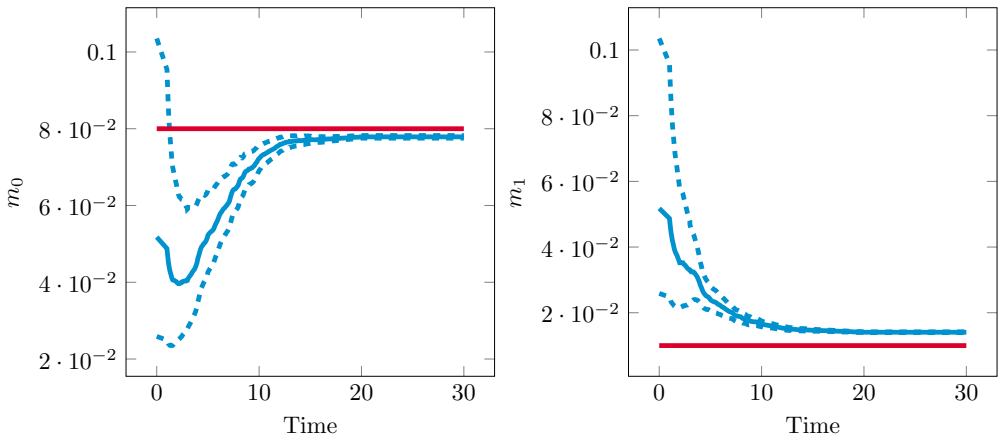
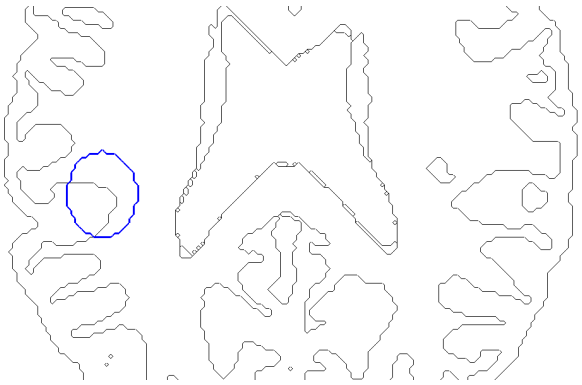



FIGURE 3.34 – Evolution des paramètres dans la simulation avec correction des paramètres. Initialement, $(m_0, m_1) = (0.052, 0.052)$ puis les paramètres convergent vers le jeu cible $(m_0, m_1) = (0.08, 0.01)$. Les lignes rouges représentent les cibles, les lignes bleues l'évolution des paramètres et les lignes pointillées l'évolution de l'écart type.

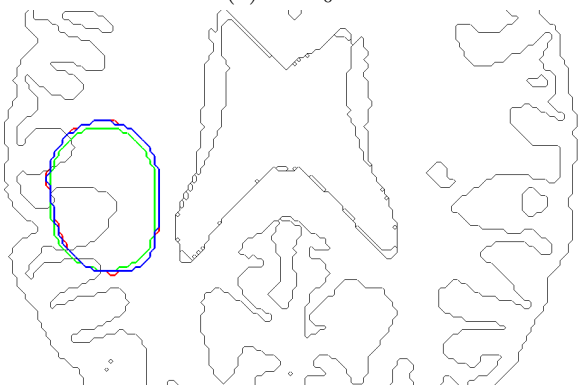
**Cas $\alpha$ estimé par calibration volumique**

En pratique, la valeur de $\alpha$ n'est pas mesurable. Elle doit donc être estimée également. Puisque $M_0$ et $\alpha$ sont corrélés, une valeur de $\alpha$ différente de celle utilisée pour la simulation cible conduira à une valeur différente de $M_0$. Les paramètres $(m_0, m_1)$ ne convergent donc pas nécessairement vers les valeurs cibles, mais c'est la comparaison à la simulation au temps $t_3$ qui permet de valider ou non l'estimation paramétrique. Afin d'estimer $\alpha$, deux méthodes sont possibles. La première est de le fixer grâce à la calibration volumique. Le second moyen est de l'intégrer dans la liste des paramètres à estimer par filtre de Kalman, ce qui sera fait dans la section ci-dessous.

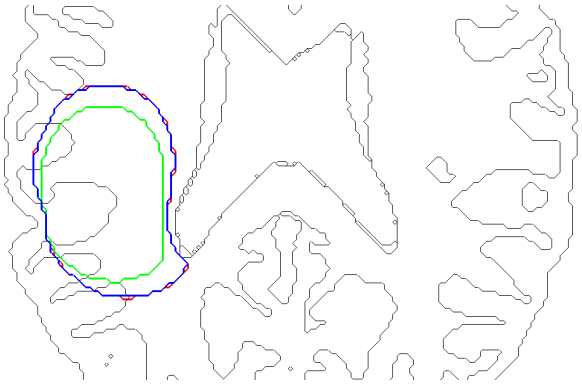
Le même exemple que précédemment est utilisé. La calibration volumique donne le jeu de paramètres suivant : $(\alpha, m_{\text{unif}}) = (0.034, 0.065)$. On applique alors le même procédé que précédemment, avec $\alpha = 0.034$, et en initialisant les paramètres $(m_0, m_1)$ à l'*a priori* $m_{\text{unif}}$. On obtient alors les figures 3.35. Comme précédemment, une correction d'état avec $\lambda = 0.0001$ est appliquée. On observe alors que la correction des paramètres est efficace puisque le contour bleu de la simulation avec correction est très proche de la cible aux temps $t_2$ et $t_3$, ce qui n'est pas le cas de la simulation verte sans correction. La figure 3.36 montre que les paramètres convergent vers le jeu $(m_0, m_1) = (0.088, 0.017)$. Comme évoqué précédemment, il est logique que l'on ne retrouve pas les paramètres utilisés pour créer la simulation cible, puisque la valeur de $\alpha$ n'est pas la même. On remarque que les valeurs limites sont légèrement plus grandes que $(m_0^t, m_1^t)$ ce qui est cohérent avec le fait que $\alpha$ soit passé de 0.001 à 0.034.



(a) $t = t_0$



(b) $t = t_2$



(c) $t = t_3$

FIGURE 3.35 – Comparaison des formes tumorales dans trois cas : la cible (en rouge), la simulation sans correction d'état (en vert) et avec correction d'état (en bleu) au temps initial $t_0$ (ligne a), au temps intermédiaire $t_2$ (ligne b) et au temps final $t_3$ (ligne c). La correction d'état et de paramètres ne se fait qu'entre les temps $t_0$ et $t_2$. Le jeu de paramètres utilisé pour la construire la simulation cible est $(\alpha, m_0, m_1) = (0.001, 0.08, 0.01)$, avec $m_0$ et $m_1$ les valeurs dans la matière blanche et grise respectivement. La simulation verte est obtenue avec le jeu uniforme $(\alpha, m_0, m_1) = (0.034, 0.065, 0.065)$ calculé par calibrage volumique. La simulation avec correction est obtenue avec une erreur de covariance 0.01, avec $\lambda = 0.0001$, et avec des paramètres initialisés avec le jeu uniforme.

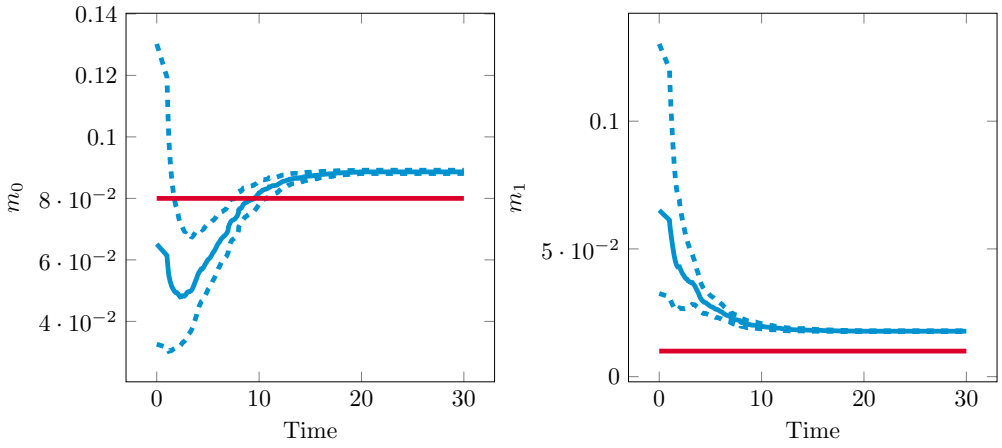
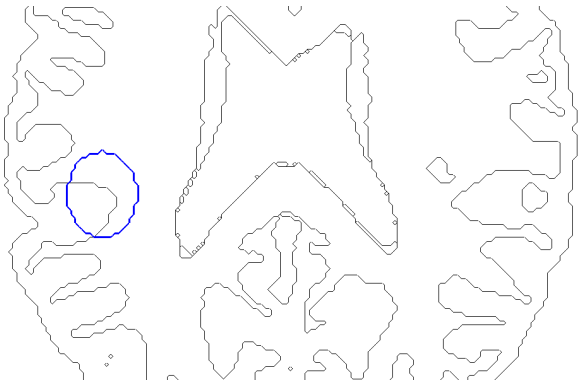


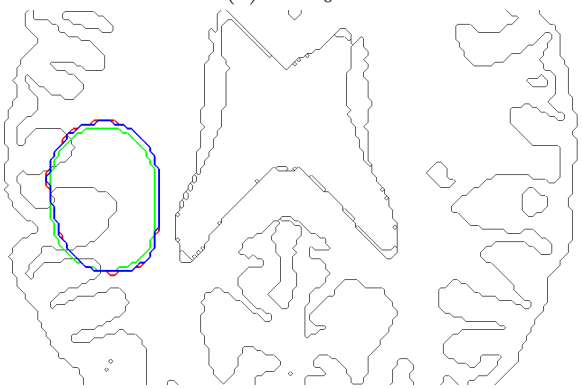

FIGURE 3.36 – Evolution des paramètres dans la simulation avec correction des paramètres. Initialement, $(m_0, m_1) = (0.065, 0.065)$ puis les paramètres convergent vers un jeu $(m_0, m_1) = (0.088, 0.017)$. Les lignes rouges représentent les cibles, les lignes bleues l'évolution des paramètres et les lignes pointillées l'évolution de l'écart type.

### Cas $\alpha$ estimé par filtre de Kalman

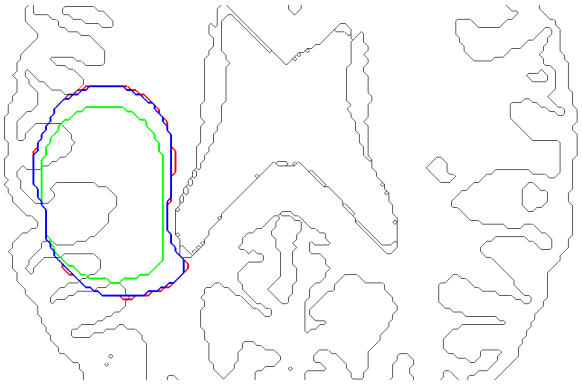
Nous allons ici ajouter $\alpha$ à la liste des paramètres à estimer par filtre de Kalman. Trois paramètres seront donc à estimer : $(\alpha, m_0, m_1)$. Ces trois paramètres sont initialisés grâce à l'estimation paramétrique donnée dans la section précédente, à savoir $(\alpha, m_{\text{unif}}) = (0.034, 0.065)$. On obtient alors l'évolution paramétrique de la figure 3.38. On remarque que les paramètres $(m_0, m_1)$ convergent vers des valeurs proches de celles de la section précédente. En revanche, le paramètre $\alpha$ n'évolue que très peu. Cela s'explique par la corrélation qu'il y a entre les trois paramètres : $\alpha$ n'a pas besoin d'évoluer puisque $m_0$ et $m_1$ compensent la mauvaise estimation de $\alpha$. Puisque le paramètre $\alpha$ ne converge pas vers la valeur cible, il n'est pas nécessaire de chercher à l'estimer. En pratique, on préfèrera donc utiliser la méthode précédente, qui utilise le calibrage volumique afin de fixer $\alpha$. En effet, la convergence des paramètres permet d'obtenir des résultats stables. De plus, cette méthode est moins coûteuse puisqu'un paramètre de moins est à estimer. **L'estimation paramétrique dans le cas de données cliniques se fera donc en fixant $\alpha$ par calibrage volumique et en estimant $M_0$ par filtre de Kalman.**



(a) $t = t_0$



(b) $t = t_2$



(c) $t = t_3$

FIGURE 3.37 – Comparaison des formes tumorales dans trois cas : la cible (en rouge), la simulation sans correction d'état (en vert) et avec correction d'état (en bleu) au temps initial $t_0$ (ligne a), au temps intermédiaire $t_2$ (ligne b) et au temps final $t_3$ (ligne c). La correction d'état et de paramètres ne se fait qu'entre les temps $t_0$ et $t_2$. Le jeu de paramètres utilisé pour la construire la simulation cible est $(\alpha, m_0, m_1) = (0.02, 0.1, 0.03)$, avec $m_0$ et $m_1$ les valeurs dans la matière blanche et grise respectivement. La simulation verte est obtenue avec le jeu uniforme $(m_0, m_1) = (0.034, 0.065, 0.065)$ calculé par calibrage volumique. La simulation avec correction est obtenue avec une erreur de covariance 0.008, avec $\lambda = 0.0001$ et avec des paramètres initialisés avec le jeu uniforme.

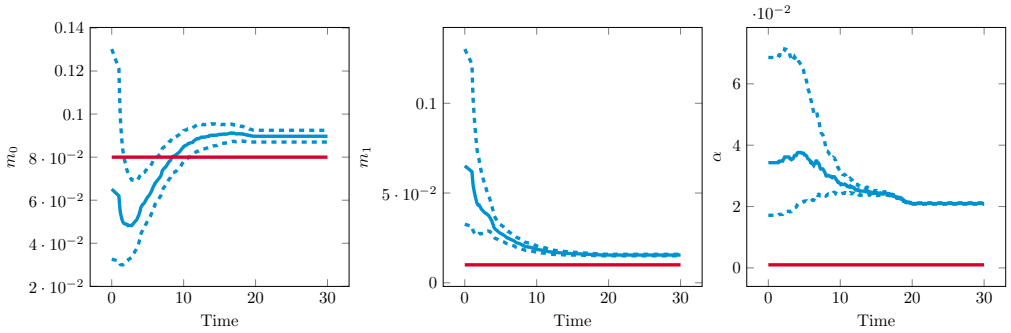
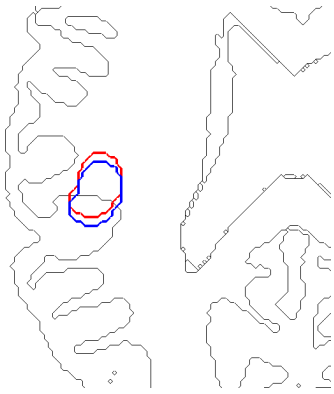


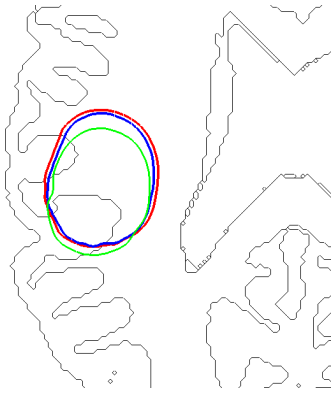

FIGURE 3.38 – Evolution des paramètres dans la simulation avec correction des paramètres. Initialement, $(\alpha, m_0, m_1) = (0.034, 0.065, 0.065)$. Le paramètre $\alpha$ varie peu, et $m_0$ et $m_1$ convergent vers des valeurs proches de celles du cas précédent. Les lignes rouges représentent les paramètres utilisés pour la simulation cible, les lignes bleues l'évolution des paramètres et les lignes pointillées l'évolution de l'écart type.

### 3.6.3 Correction jointe état-paramètres avec condition initiale décalée

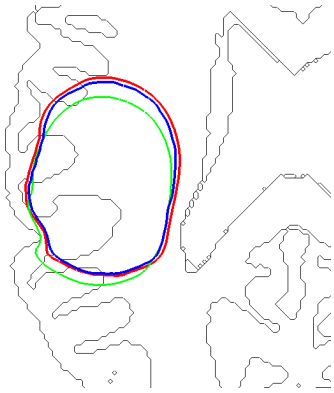
Contrairement au cas précédent où la condition initiale était supposée parfaitement connue, la condition initiale sera ici légèrement décalée de celle utilisée pour la simulation cible. Ce cas se rapproche donc du cas clinique, où la segmentation de la tumeur comporte des incertitudes. L'objectif de cette section est de montrer que l'estimation paramétrique fonctionne même lorsque la condition initiale n'est pas connue parfaitement, grâce à la correction d'état. Sur la figure 3.39, la simulation cible est représentée en rouge. Elle est obtenue avec le jeu de paramètres $(\alpha, m_0, m_1) = (0.02, 0.1, 0.04)$. Les simulations avec et sans correction ont une condition initiale légèrement décalée par rapport à la simulation cible. Seules les trois premières données synthétiques sont utilisées pour la correction d'état et de paramètres. Comme vu précédemment, la méthode retenue pour l'estimation paramétrique est de fixer $\alpha$ grâce à la calibration volumique et d'estimer $M_0$ par filtre de Kalman. Ici, la calibration volumique donne le jeu de paramètres $(\alpha, m_{\text{unif}}) = (0.03, 0.081)$. La simulation verte est celle obtenue avec ce jeu de paramètres fixe. En bleu, le contour simulé est obtenu par correction d'état avec un coefficient $\lambda = 0.01$ et une correction de paramètres avec une erreur de covariance égale à 0.004. Sur la figure 3.39, la correction jointe état-paramètre permet de corriger la forme tumorale à $t = t_2$ et de prédire correctement la forme à $t = t_3$. L'évolution des paramètres représentée sur la figure 3.40 montre que les paramètres convergent bien vers des valeurs légèrement supérieures aux valeurs cibles, ce qui est cohérent puisque la valeur de $\alpha$ est supérieure. On remarque cependant une fluctuation des paramètres au début de la simulation, ce qui s'explique par le décalage des conditions initiales : l'état doit d'abord être suffisamment bien corrigé pour que les paramètres soient correctement estimés.



(a) $t = t_0$



(b) $t = t_2$



(c) $t = t_3$

FIGURE 3.39 – Comparaison des formes tumorales dans trois cas : la cible (en rouge), la simulation sans correction d'état (en vert) et avec correction d'état (en bleu) au temps initial $t_0$ (ligne a), au temps intermédiaire $t_2$ (ligne b) et au temps final $t_3$ (ligne c). La correction d'état et de paramètres ne se fait qu'entre les temps $t_0$ et $t_2$. Le jeu de paramètres utilisé pour la construire la simulation cible est $(\alpha, m_0, m_1) = (0.02, 0.1, 0.04)$, avec $m_0$ et $m_1$ les valeurs dans la matière blanche et grise respectivement. Les conditions initiales bleues et vertes sont décalées par rapport à la condition initiale de la cible. La simulation verte est obtenue avec le jeu uniforme $(m_0, m_1) = (0.03, 0.081, 0.081)$ calculé par calibrage volumique. La simulation avec correction est obtenue avec une erreur de covariance 0.004, avec $\lambda = 1e - 2$ et avec des paramètres initialisés avec le jeu uniforme.

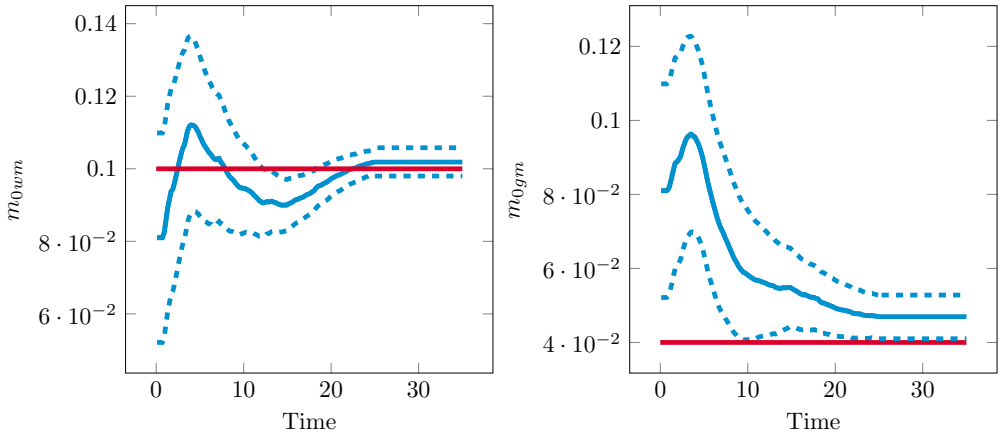



FIGURE 3.40 – Evolution des paramètres dans la simulation avec correction des paramètres. Initialement, $(m_0, m_1) = (0.081, 0.081)$, puis les paramètres convergent vers le jeu $(m_0, m_1) = (0.102, 0.047)$. Les lignes rouges représentent les paramètres utilisés pour la simulation cible, les lignes bleues l'évolution des paramètres et les lignes pointillées l'évolution de l'écart type.

### 3.6.4 Importance du choix de $\lambda$

Dans tous les cas précédents, la correction d'état permet à la fois de corriger la forme tumorale mais également de mieux estimer les paramètres. Le choix de la valeur $\lambda$ est essentiel : s'il est trop grand, la correction est prépondérante sur le modèle et les paramètres ne peuvent pas être estimés correctement. A l'inverse, si ce scalaire est trop petit, l'état n'est pas suffisamment corrigé et la forme cible n'est jamais atteinte. Ce problème est représenté sur la figure 3.41. L'erreur calculée sur le graphique de gauche est celle issue de la mesure de similarité. A droite, l'évolution du volume tumoral est représenté. Les courbes représentent plusieurs cas synthétiques, pour deux valeurs de $m_0$ et trois valeurs de $\lambda$. La cible est représentée en rouge sur la courbe volumique. Elle est obtenue avec un paramètre $m_0 = 0.1$. Quand $\lambda$ est petit ($\lambda = 0.02$), la correction n'est pas suffisante et la discordance entre simulation et observation ne diminue pas. Dans les deux autres cas, la valeur de $\lambda$ est suffisante pour assurer la convergence vers la cible. Dans le cas où $\lambda = 0.3$, on remarque que la convergence vers la cible est obtenue quelque soit la valeur de $m_0$ : la correction est prépondérante sur le modèle. La bonne valeur de $\lambda$ dans ce cas est donc la valeur intermédiaire $\lambda = 0.02$ qui permet de ne corriger l'état que lorsque $m_0$ a la valeur cible.

L'objectif est de proposer une façon de choisir automatiquement une valeur acceptable pour $\lambda$. La valeur proposée dans la justification mathématique de la convergence du terme

corrigé vers le terme cible n'est pas utilisable en pratique, puisque les majorations effectuées ne sont pas optimales. L'idée est de comparer le terme de croissance tumorale $G = \int MPdx$ et le terme de correction d'état $SC = \delta_{\hat{\Gamma}} F(|\nabla P|)\Big(-\big(z - C_1(\hat{\Omega}^{\text{in}})\big)^2 + \big(z - C_2(\hat{\Omega}^{\text{in}})\big)^2\Big)$ au temps initial, et de calculer $\lambda$ tel que le terme de correction ne soit pas trop grand devant le terme de croissance, ni négligeable. On cherche donc un gain $\lambda$ de la forme :

$$\lambda = r\frac{G}{\max(SC)}, \tag{3.90}$$

où $r$ est un ratio à choisir. La division par $\max(SC)$ permet de normaliser le terme de correction. On remarque qu'au temps initial, $G = m_0 V_0$ dans le cas où $M_0 = m_0$, et en notant $V_0$ le volume tumoral initial. On a alors :

$$\lambda = r\frac{m_0 V_0}{\max(SC)}. \tag{3.91}$$

Lorsque $m_0$ est grand, la croissance tumorale est rapide, ce qui nécessite bien une forte correction, donc un fort gain. De plus, la dépendance de $\lambda$ par rapport à $V_0$ peut s'interpréter par le fait que l'erreur de segmentation est plus petite lorsque la tumeur est grosse, et donc que la confiance aux données est plus grande.
Nos simulations sur les données synthétiques et réelles montrent que de bons résultats sont obtenus lorsque le ratio $r$ entre les deux termes est entre $\frac{1}{3}$ et $\frac{1}{5}$. Dans le cas de la figure 3.41, les valeurs de $\lambda = 0.02, 0.1$ et $0.3$ ont été obtenues avec des ratios respectifs $r = 0.066, 0.33$ et $1.0$, et seule la seconde valeur était acceptable. Bien évidemment, cet ordre de grandeur donné ici dépend fortement de la complexité de la forme tumorale et est à adapter selon les cas cliniques rencontrés.

### 3.6.5 Estimation de $M_0$ sur une grille

Les sections précédentes concernent le cas où la différence de croissance dans la matière blanche et la matière grise est la seule raison au changement de forme observé. En pratique, d'autres facteurs peuvent entrer en jeu, comme les fibres ou les vaisseaux entourant la tumeur. De plus, la segmentation de la matière blanche et de la matière grise du cerveau est coûteuse et source d'incertitudes. L'idée ici est de ne partir avec aucun *a priori* sur la grille $M_0$ et de l'estimer par correction paramétrique. $M_0$ est alors estimé sur une grille, comme celle représentée à la figure 3.42.

Chaque case de la grille correspond donc à une valeur $m_0$ à estimer. La grille choisie est beaucoup moins raffinée que la grille de calcul, puisque le filtre de Kalman est un algorithme

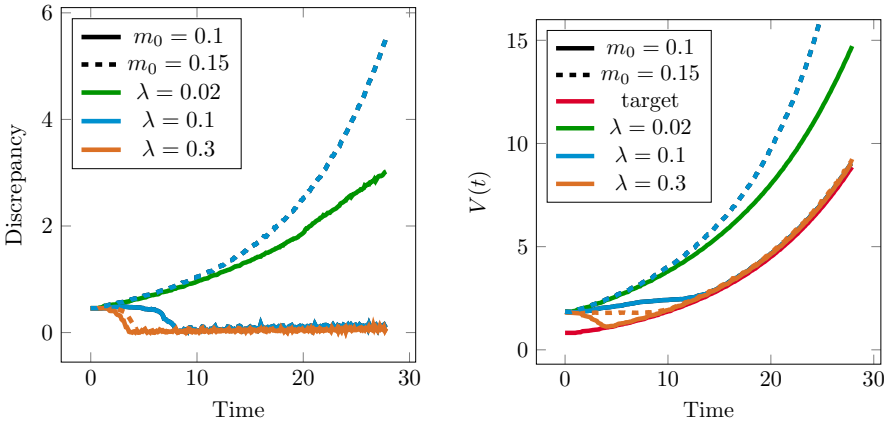



FIGURE 3.41 – Evolution de la discordance entre la simulation et l'observation (gauche) et du volume tumoral simulé (droite). La cible est obtenue avec une vascularisation constante égale à $m_0 = 0.1$, et est représentée en rouge. Les simulations avec correction d'état sont obtenues avec des valeurs $m_0 = 0.1$ ou $0.15$ et avec trois valeurs différentes de $\lambda$.

coûteux. C'est pour cette raison que le filtre de Kalman n'est appliqué qu'aux paramètres et non pour la correction de l'état. Afin de simuler la croissance d'une tumeur le long d'un vaisseau, une simulation cible est lancée avec $M_0$ prenant une valeur élevée à l'intérieur d'un vaisseau synthétique. Cette simulation est représentée sur la figure 3.43. On observe que la tumeur s'étire bien dans la direction du vaisseau, dont le contour est représenté en blanc. Les paramètres utilisés sont les suivants : $(\alpha, m_{\text{in}}, m_{\text{out}}) = (0.02, 0.3, 0.08)$, où $m_{\text{in}}$ et $m_{\text{out}}$ correspondent respectivement à la valeur de $M_0$ à l'intérieur et à l'extérieur du vaisseau. L'objectif ici est de prédire la forme tumorale mais aussi de comparer la valeur de $M_0$ réelle à celle estimée par notre modèle. Seulement trois examens sont donc conservés. La figure 3.44 montre la calibration volumique sur ces trois examens.

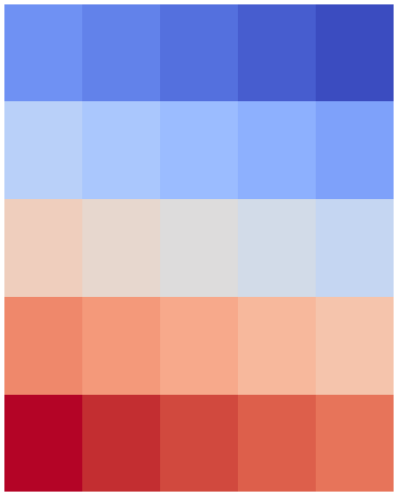


FIGURE 3.42 – Grille de discrétisation de $M_0$ : chaque case correspond à un paramètre à estimer.

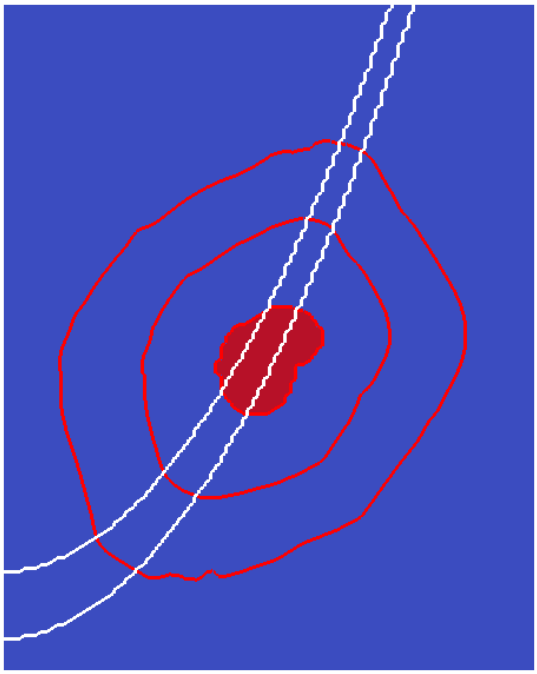


FIGURE 3.43 – Données synthétiques obtenues aux temps $t_0 = 0$ (masse rouge au centre), $t_1 = 15$ (premier contour rouge) et $t_2 = 25$ (second contour rouge). Cette simulation est obtenue avec $M_0$ prenant une valeur $m_{\text{in}} = 0.3$ à l'intérieur du vaisseau synthétique (contour blanc), et $m_{\text{out}} = 0.08$ à l'extérieur du vaisseau.

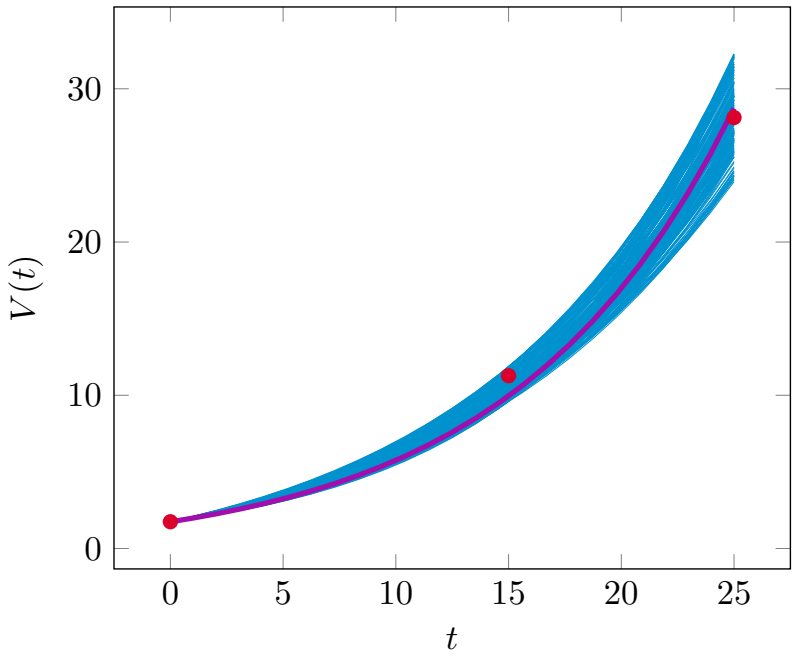



FIGURE 3.44 – Faisceau des volumes théoriques utilisés pour la calibration volumique. Les trois premières données (points rouges) sont utilisées pour calibrer le modèle et la dernière donnée (point noir) est gardée pour être comparée à la prédiction. La courbe violette est celle obtenue avec le jeu de paramètres qui donne un volume final médian parmi les volumes acecptés. Les paramètres retenus sont $(\alpha, m_{\text{unif}}) = (0.029, 0.132)$.

Les paramètres retenus sont alors $(\alpha, m_{\text{unif}}) = (0.029, 0.132)$. Le modèle de correction jointe état-paramètre lancé avec $M_0 = m_{\text{unif}}$ constant spatialement, donne l'estimation paramétrique de la figure 3.45. On remarque que la grille 5x5 comporte trois types de zones. Les zones grises sont celles non atteintes par la tumeur : les paramètres n'ont donc pas pu être corrigés, et sont restés à la valeur initiale. Les zones rouges sont celles où le paramètre a augmenté par rapport à la valeur initiale. A l'inverse, les zones bleues sont celles où le paramètre a diminué, et correspondent aux zones où la tumeur ralentit sa croissance. Dans toutes les simulations, un paramètre $\lambda = 0.001$ et une erreur de covariance d'observation de 0.08 sont utilisés. Il est intéressant de constater que les zones rouges correspondent bien aux zones traversées par le vaisseau. Cela montre qu'il est possible de tirer de l'information sur la vascularisation initiale $M_0$, même lorsque l'on n'a aucun *a priori*.

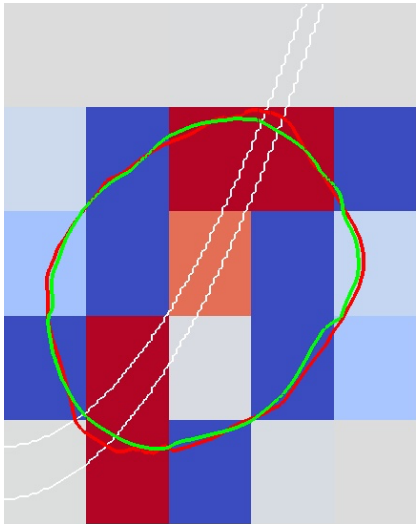


FIGURE 3.45 – Estimation paramétrique de $M_0$ sur la grille : les valeurs rouges sont celles où le paramètre a augmenté par rapport à sa valeur initiale, tandis que les zones bleues sont celles où il a diminué. Les zones grises correpondent aux zones où le paramètre n'a pas évolué. Le contour blanc représente le vaisseau synthétique cible.

Afin d'améliorer la prédiction, la grille est ensuite raffinée de la manière suivante :

— les zones rouges où le paramètre a augmenté sont raffinées,

— les zones bleues et grises sont unifiées en une seule zone.

Cela permet donc de raffiner les zones intéressantes. L'unification des autres zones permet de réduire le nombre de paramètres à estimer et ainsi de ne pas trop augmenter les coûts de calcul lors du raffinage. On obtient alors successivement les grilles de calcul représentées sur la figure 3.47.

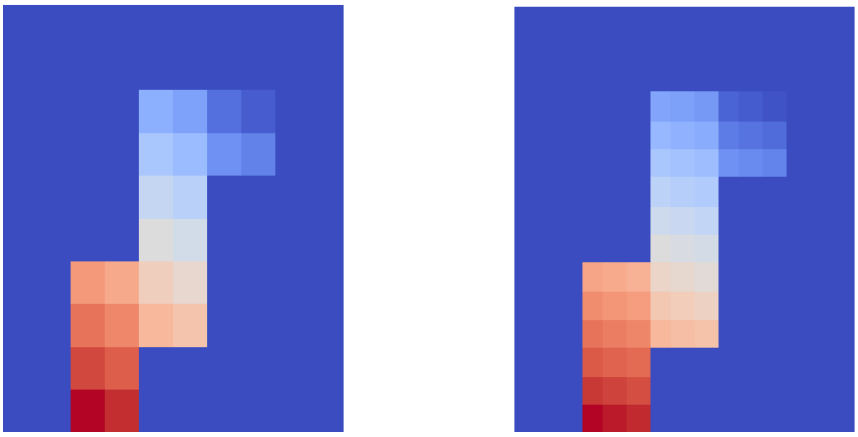


FIGURE 3.46 – Grilles obtenues à partir de la précédente en raffinant les zones d'intérêt et en unifiant les autres zones. Les cases de la grille initiale sont divisées en 2 (à gauche) et en 3 (à droite).

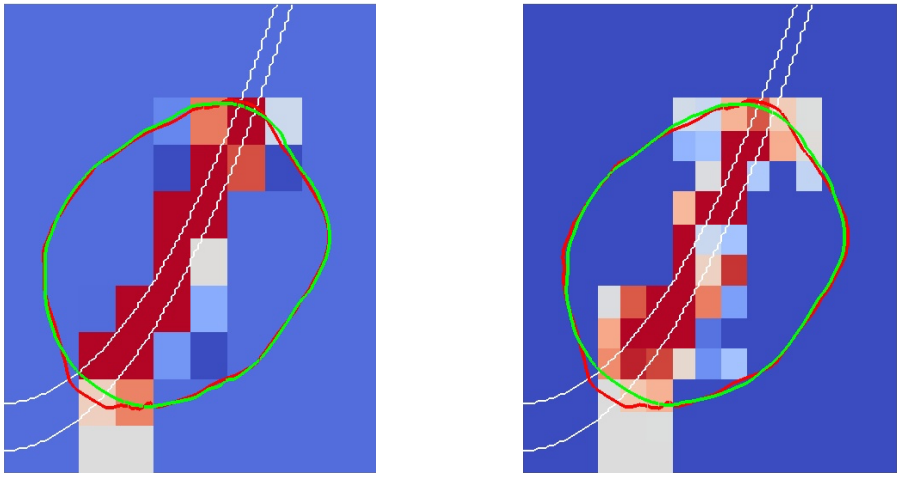


FIGURE 3.47 – Estimation des paramètres dans les grilles raffinées. Le $M_0$ cible est le vaisseau représenté en blanc. Les contours rouges et verts correspondent respectivement aux contours cibles et simulés.

Les paramètres estimés sur ces grilles sont représentés sur la figure 3.47. On remarque de même que précédemment que les zones rouges correspondent bien à celles situées dans la direction du vaisseau. On constate cependant que le maillage le plus fin n'est pas forcément

celui qui donne le meilleur résultat. Chaque zone étant très fortement corrélé à ses voisines, cette méthode ne permet pas d'obtenir une carte de vascularisation très précise. Par contre, la méthode est efficace pour estimer un comportement plus global comme c'est le cas ici. Il est alors possible de reconstituer le vaisseau par lissage puis seuillage du motif obtenu :

— **Lissage.** Notons $M_{0,\text{grille}}$ l'estimation de $M_0$ sur la grille au temps final. On construit alors $M_{0,\text{lisse}}$ par la relation de convolution :

$$M_{0,\text{lisse}}(x) = \int_{\mathcal{B}} M_{0,\text{grille}}(y) exp\left(\frac{-(x-y)^2}{\sigma^2}\right) dy,$$

où l'on choisit $\sigma = \max(dX, dY)$, avec $dX$ et $dY$ les dimensions d'une grosse maille de la grille de calcul de $M_0$.

— **Seuillage.** Le seuillage est ensuite réalisé à la valeur :

$$m_{\text{seuil}} = \max_x ||\nabla M_{0,\text{lisse}}(x)||,$$

c'est-à-dire que le seuil choisi est celui où le gradient de $M_{0,\text{lisse}}$ est le plus grand. On définit alors :

$$M_{0,\text{reconstruit}}(x) = \begin{cases} 1 & \text{si } M_{0,\text{lisse}}(x) > m_{\text{seuil}}, \\ 0 & \text{sinon.} \end{cases} \tag{3.92}$$

Cette méthode permet ainsi d'obtenir une structure plus réaliste. Le vaisseau finalement obtenu $M_{0,\text{reconstruit}}$ est représenté sur la figure 3.48. Notons que la méthode peut facilement s'adapter en 3D et sera utilisée dans la section suivante sur la donnée de métastase cérébrale.

La forte corrélation entre les cases de la grille ne rendent pas les valeurs des paramètres exploitables : la grille ne sert qu'à distinguer les zones où le paramètre augmente des zones où il diminue. La reconstitution du vaisseau à partir de cette grille permet ensuite d'estimer les paramètres à l'intérieur et à l'extérieur de ce vaisseau. La figure 3.48 représente la simulation obtenue par correction paramétrique sur le vaisseau reconstitué. Les paramètres correspondants sont représentés sur la figure 3.49. La valeur de $\alpha$ prédite par la calibration volumique étant plus grande que la valeur cible, les paramètres $m_{\text{in}}$ et $m_{\text{out}}$ estimés devraient converger vers des valeurs légèrement supérieures aux valeurs cibles. On remarque cependant que $m_{\text{in}}$ converge vers une valeur inférieure : cela s'explique par le fait que le vaisseau estimé est plus large que le vaisseau synthétique cible. Cependant, la convergence des paramètres ainsi que la bonne estimation de forme obtenue valide cette méthode.

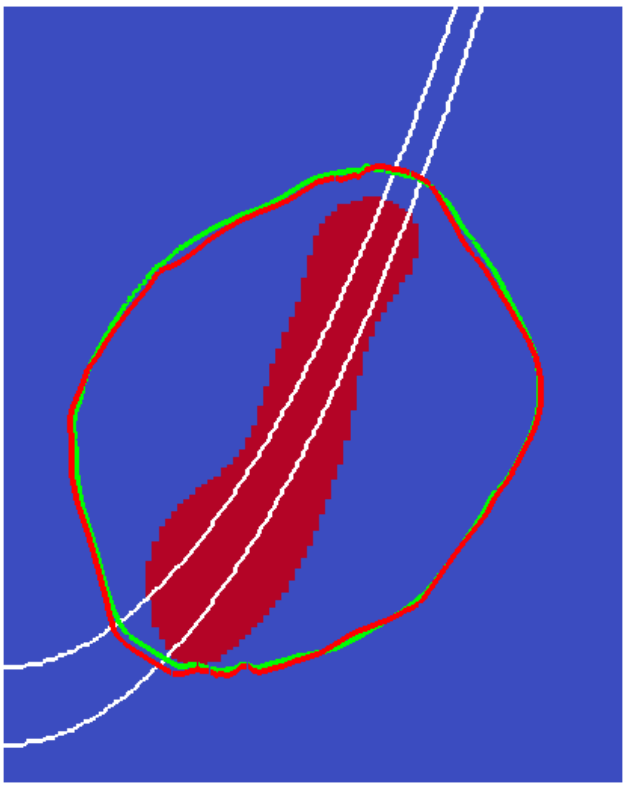


FIGURE 3.48 – $M_0$ est reconstitué par diffusion puis seuillage à partir des paramètres estimés sur la grille (masse rouge). En blanc, le vaisseau cible est représenté. Les contours rouges et verts correspondent respectivement aux contours cibles et simulés.

Dans cette section, les filtres de Luenberger et de Kalman ont été validés sur des exemples synthétiques : le premier permet une bonne correction de l'état, tandis que le second a permis de retrouver les paramètres cibles utilisés. De plus, la calibration volumique permet d'avoir un *a priori* sur les paramètres et de fixer $\alpha$. Enfin, la méthode détaillée ci-dessus permet de reconstruire la vascularisation autour de la tumeur à partir des changements de forme de cette dernière. Ces cas synthétiques permettent donc de valider la méthode d'assimilation utilisée. Elle est appliquée à des cas cliniques dans la prochaine section.

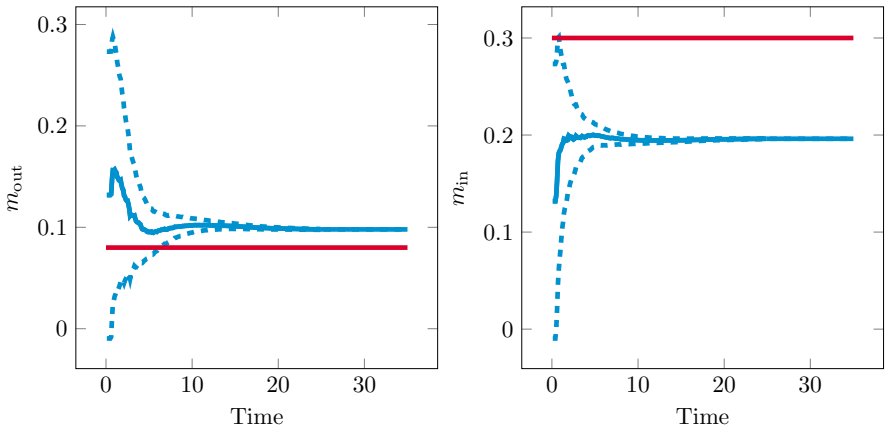
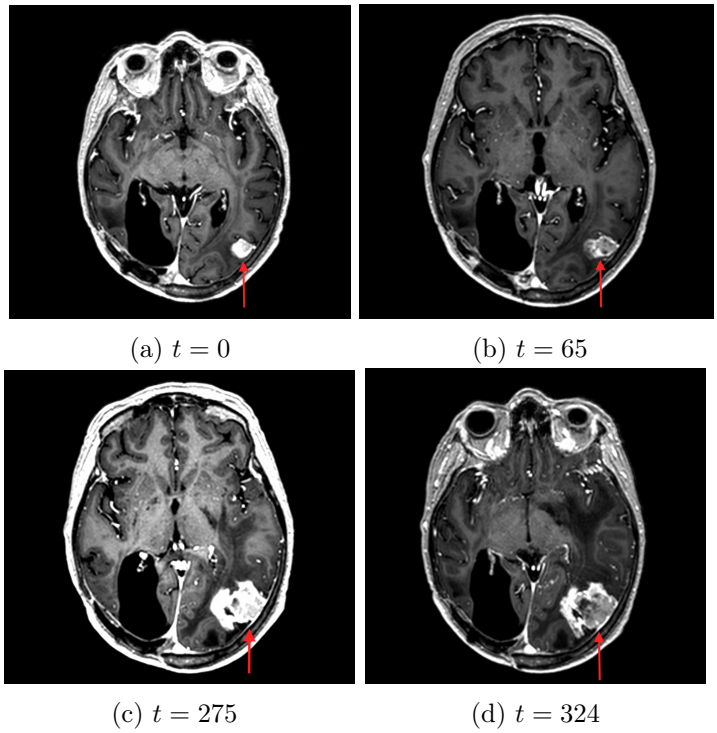



FIGURE 3.49 – Estimation des paramètres $m_{\text{in}}$ et $m_{\text{out}}$ lors de la correction jointe état paramètre. Initialisés à $m_{\text{unif}} = 0.132$, et avec $\alpha = 0.029$, les paramètres convergent vers le couple $(m_{\text{in}}, m_{\text{out}}) = (0.196, 0.097)$. Les paramètres cibles sont représentés en rouge.

## 3.7 Application aux données réelles

Dans cette section, nous appliquons la correction d'état et de paramètres à l'exemple de métastase cérébrale décrit précédemment. Les images médicales ainsi que les données segmentées sont visibles sur les figures 3.50 et 3.51

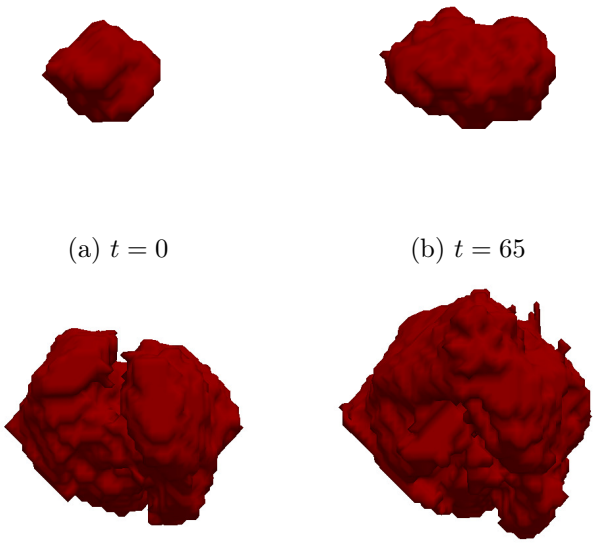


(a) $t = 0$ (b) $t = 65$

(c) $t = 275$ (d) $t = 324$

FIGURE 3.50 – IRM successives de la métastase cérébrale, à 4 temps différents (modalité T1). La coupe 2D donnant une aire maximale est représentée ici. Les simulations utilisent ensuite la reconstruction 3D de la tumeur. Ces données sont fournies par Ana Ortiz de Mendivil Arrate (HM Hospitales, Espagne).

La calibration volumique appliquée à ces données (représentée à la figure 3.12) montre que le volume prédit est très proche du volume observé au temps $t_3 = 324$ jours. Les paramètres estimés par cette calibration sont : $(\alpha, m_{\text{unif}}) = (0.0064, 0.0139)$. Comme vu précédemment, la simulation avec ces paramètres et sans correction est très éloignée de la forme cible. En effet, la tumeur simulée croît isotropiquement et en restant collée au bord du cerveau, comme on peut le voir sur la figure 3.52. Par contre, la tumeur observée se développe vers l'intérieur du cerveau, vers la droite sur la vue de dessous (deuxième image). L'objectif est alors d'améliorer la prédiction au temps $t_3$ et de reconstruire la vascularisation qui permet à la tumeur de se développer plus fortement dans cette direction.



(a) $t = 0$ (b) $t = 65$

(c) $t = 275$ (d) $t = 324$

FIGURE 3.51 – Segmentations 3D successives de la métastase cérébrale.

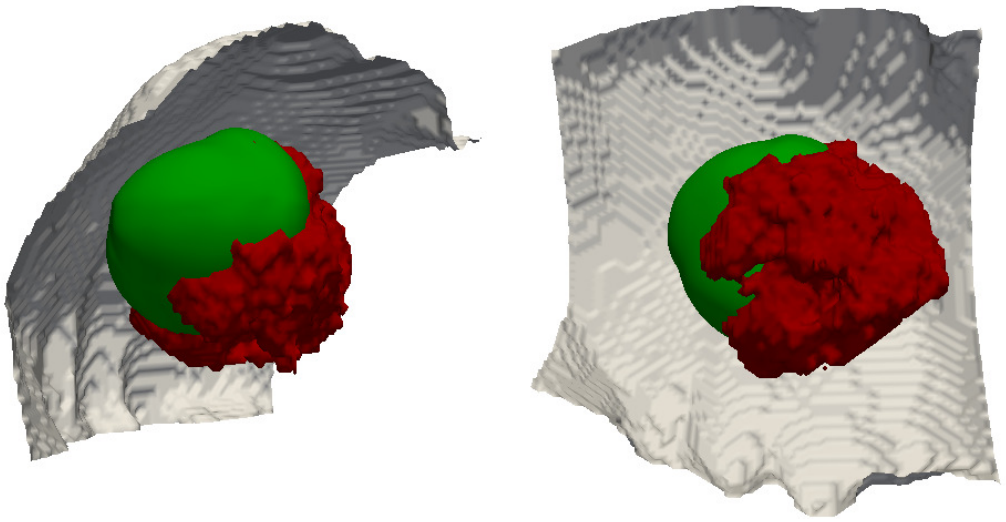
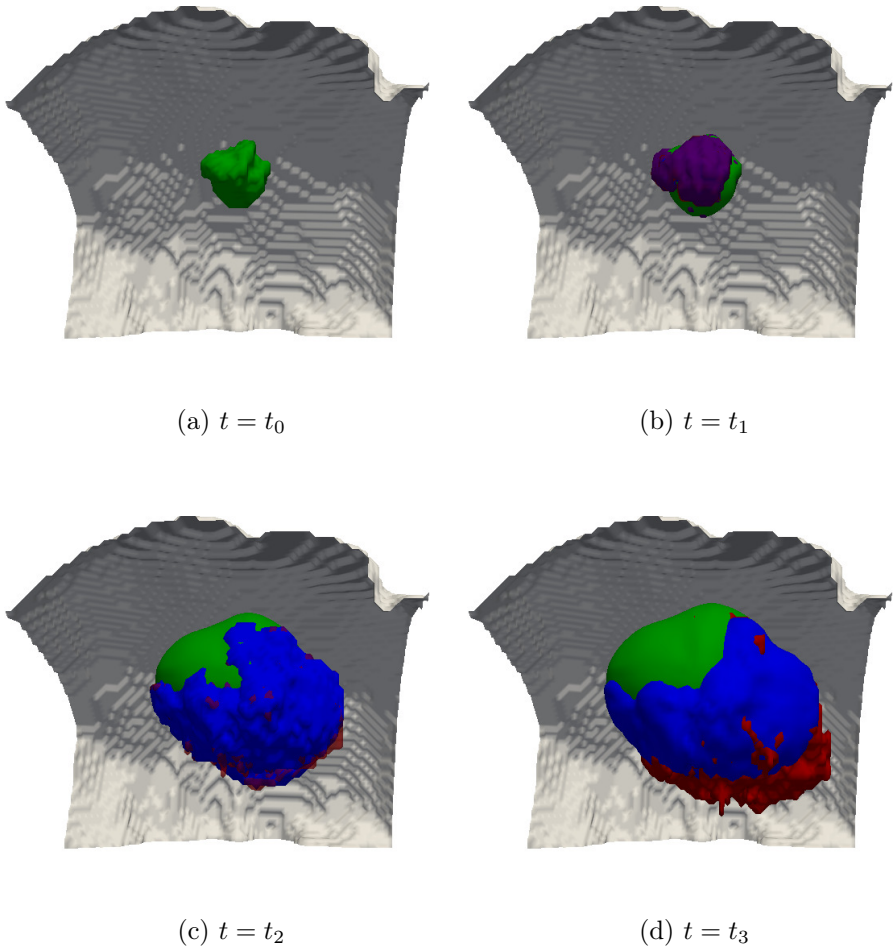


FIGURE 3.52 – Tumeur simulée (en vert) et observations (en rouge). Le bord du cerveau est représenté en gris. La tumeur simulée (sans correction) croît isotropiquement et reste collée au bord du cerveau, tandis que la tumeur observée se développe vers la droite (dans le plan de la deuxième image).

La méthode de correction d'état par filtre de Luenberger, et de correction des paramètres par filtre de Kalman permet d'obtenir la simulation bleue corrigée de la figure 3.53. La

correction n'est appliquée qu'entre les temps $t_0$ et $t_2$, et le modèle sans correction permet d'obtenir la dernière simulation de prédiction. Le paramètre $M_0$ est estimé sur une grille grossière 5x5x5, soit 250 particules à lancer. On remarque que les simulations au temps $t_1$ et $t_2$ sont très proches de l'observation. De plus, la forme prédite au temps $t_3$ est largement plus proche de la cible que celle obtenue sans correction. La tumeur observée semble s'être développée encore plus fortement dans la direction préférentielle que notre simulation. La valeur estimée de $M$ à la fin de la correction paramétrique est représentée sur la figure 3.54. Les zones rouges sont celles où le paramètre a augmenté par rapport à la valeur initiale. Ces zones correspondent bien à l'intuition que l'on a sur l'imagerie, à savoir que la direction préférentielle de la tumeur se situe vers le bas et vers la droite.



(a) $t = t_0$ (b) $t = t_1$

(c) $t = t_2$ (d) $t = t_3$

FIGURE 3.53 – Evolution de la tumeur observée (en rouge), simulée sans correction (en vert) et simulée avec correction (en bleu). La correction d'état et de paramètre ne se fait qu'entre $t_0$ et $t_2$, $M_0$ est estimé sur une grille de taille 5x5x5 avec les paramètres suivants : $\lambda = 0.05$ et une erreur de covariance de 0.01.

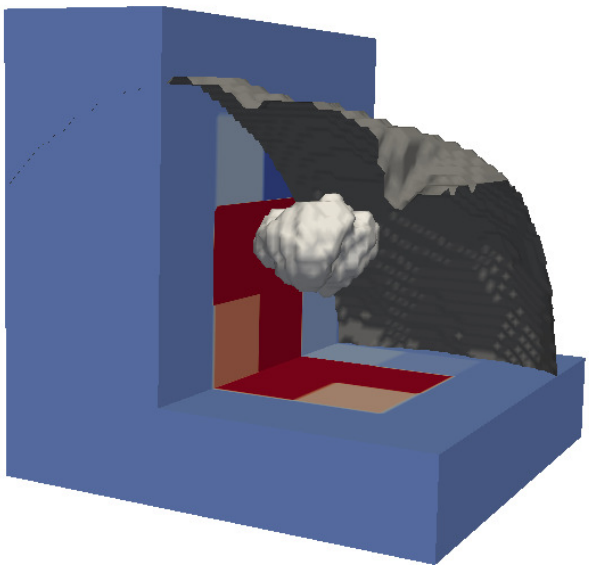
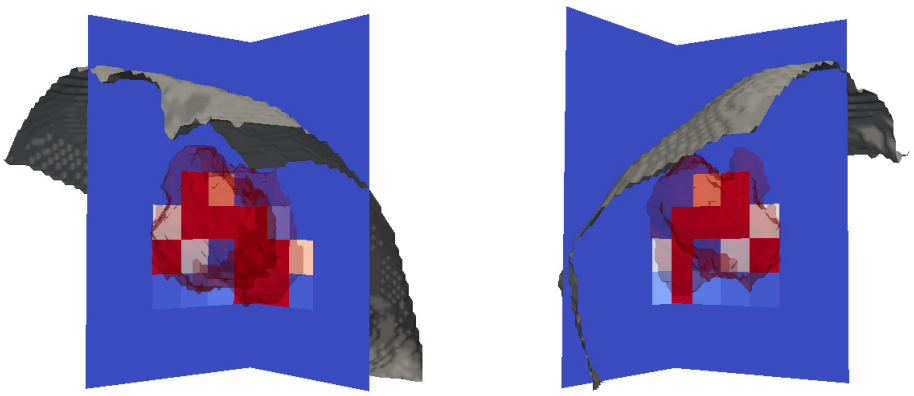


FIGURE 3.54 – Estimation finale de la vascularisation initiale $M_0$, sur une grille 5x5x5. La tumeur initiale est représentée en blanc et le bord du cerveau en gris. La grille $M_0$ est initialisée à la valeur $m_{\text{unif}}$ dans toutes les cases. Les zones rouges sont celles où le paramètre a augmenté.

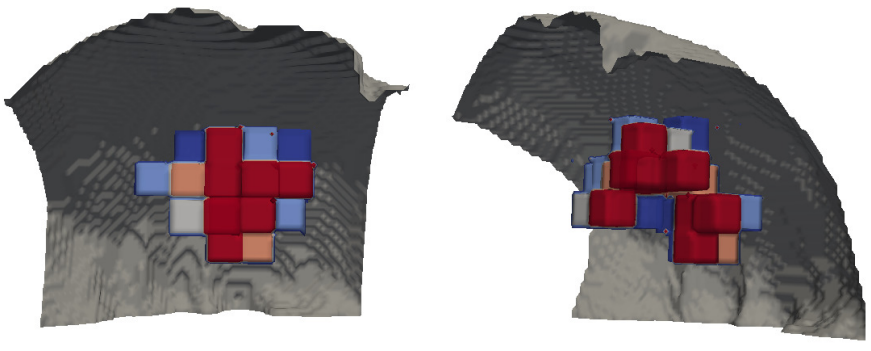
Comme précédemment dans le cas synthétique du vaisseau, on souhaite raffiner cette grille afin de reconstruire $M_0$ plus précisément. On ne garde que les zones où le paramètre a augmenté depuis sa valeur initiale, et on raffine ces zones 2 fois plus précisément. Les autres zones sont réunies pour ne former qu'une grande zone. La zone où le paramètre a augmenté est représentée sur la figure 3.55.

La grille de $M_0$ obtenue au temps final est représentée sous plusieurs angles sur la figure 3.56. On observe que des directions plus précises sont obtenues, en rouge. En particulier, sur la vue de droite en 3D, on remarque deux directions privilégiées, toujours vers la droite, qui sont cohérentes étant donnée la forme de la tumeur en $t_2$.

FIGURE 3.55 – Zones où le paramètre a augmenté dans la première correction des paramètre. Cette zone est donc raffinée deux fois plus précisément que précédemment.



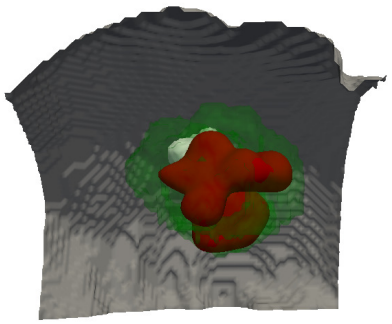
(a) Vue de droite (b) Vue de gauche



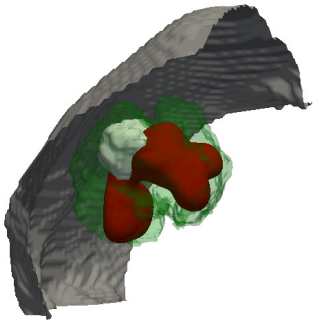
(c) Vue de face (d) Vue de droite

FIGURE 3.56 – $M_0$ estimé au temps final. Vue de droite et de gauche en coupes 2D (première ligne), avec la tumeur au temps $t_2$ en contour transparent rouge. Vue de face et de droite du réseau vasculaire reconstitué en 3D.

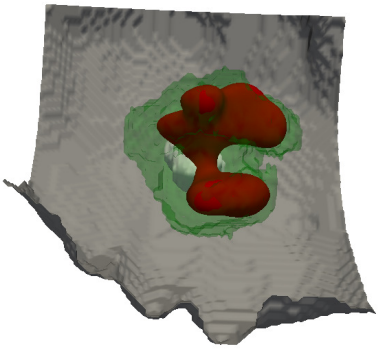
Comme dans le cas synthétique, on reconstruit alors un champ $M_0$ plus réaliste par lissage puis seuillage de la grille finale obtenue. La méthode utilisée est détaillée dans la section précédente. On obtient alors le résultat de la figure 3.59, vu sous plusieurs angles.



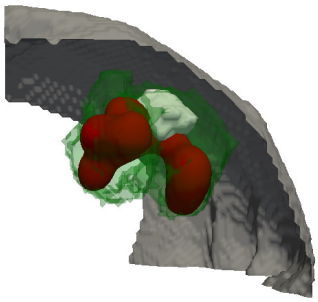
(a) Vue de face



(b) Vue de gauche



(c) Vue de droite



(d) Vue de dessous

FIGURE 3.57 – $M_0$ reconstitué par diffusion et seuillage de la grille précédemment obtenue. Le contour de la tumeur au temps $t_2$ est représenté en vert transparent, et le contour de la tumeur initiale est représenté en blanc.

Les différentes branches de la vascularisation obtenue peuvent être interprétées comme des fibres alimentant la tumeur. Bien que le réseau estimé soit grossier, dû à la taille initiale du maillage de $M_0$, il semble cohérent avec le comportement global de la tumeur. Il est alors possible d'estimer les paramètres plus précisément, à l'intérieur et à l'extérieur de la vascularisation estimée. Puisqu'il n'y a plus que deux paramètres, l'estimation de ces derniers est plus précise. On obtient l'estimation paramétrique de la figure 3.58.

Le paramètre à l'intérieur du vaisseau reconstruit converge bien vers une valeur supérieure

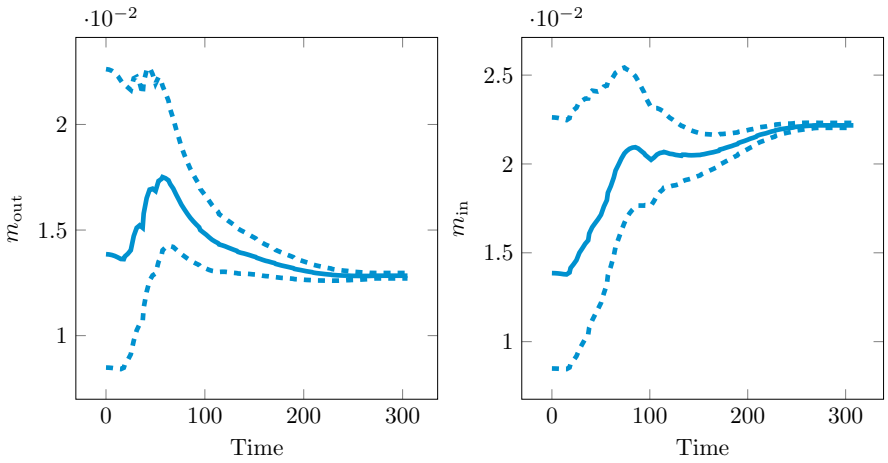
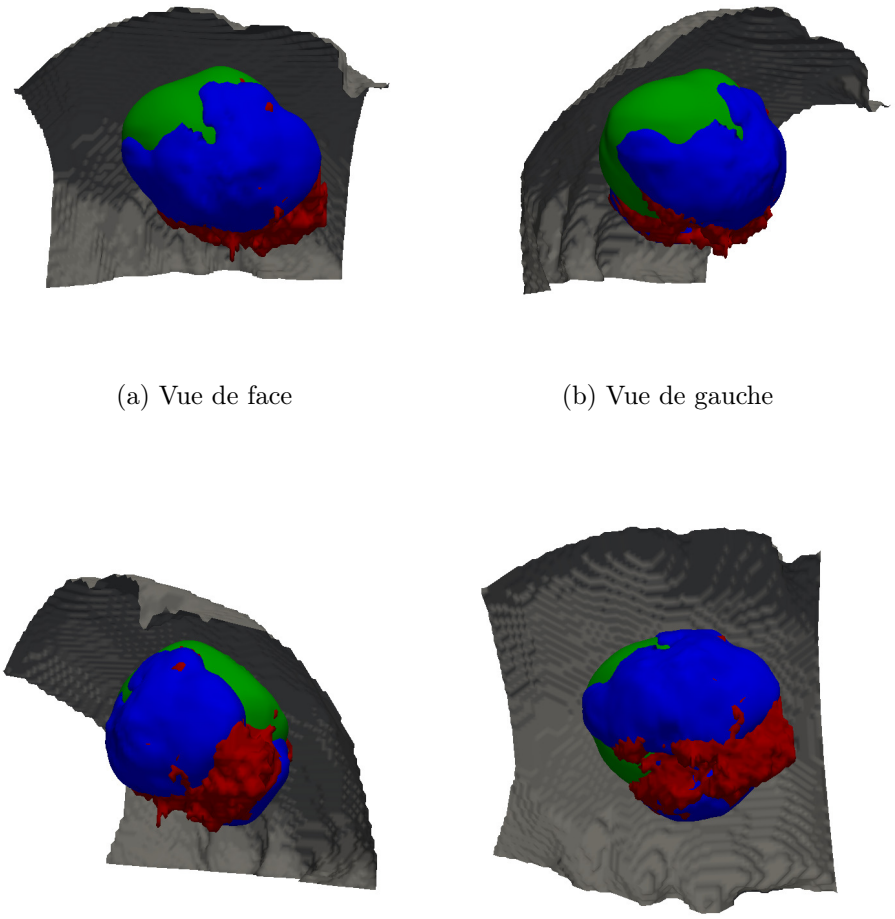



FIGURE 3.58 – Estimation des paramètres $m_{\text{in}}$ et $m_{\text{out}}$, respectivement à l'intérieur et à l'extérieur de la vascularisation estimée. Initialisés à $m_{\text{unif}} = 0.0138$, et avec $\alpha = 0.006$, les paramètres convergent vers le couple $(m_{\text{in}}, m_{\text{out}}) = (0.0221, 0.0128)$.

à celle du paramètre à l'extérieur du vaisseau. La simulation lancée avec ces paramètres donne le comportement représenté sur la figure 3.59.



(a) Vue de face (b) Vue de gauche

(c) Vue de droite (d) Vue de dessous

FIGURE 3.59 – Estimation finale (en bleu) à comparer avec la tumeur observée (en rouge) et à la simulation sans aucune correction (en vert), vue sous 4 angles différents.

On observe comme précédemment que l'on améliore grandement la prédiction par rapport à la simulation sans correction. La direction principale de la tumeur est bien capturée, avec une extension en bas à droite (sur la vue de face). Cependant, la tumeur semble s'être étendue encore plus fortement que ce que la simulation prédit. Cela s'explique par le fait que le vaisseau que l'on estime se limite à l'intérieur de la tumeur, c'est-à-dire là où on l'a corrigé. Etendre le vaisseau reconstruit à tout le domaine pourrait permettre d'améliorer la prédiction. On observe également sur la vue de gauche et la vue de dessous que la tumeur s'est légèrement étendue entre $t_2$ et $t_3$ vers le bas. C'est un autre inconvénient de notre méthode, qui ne peut pas anticiper les nouveaux changements de direction, lorsque l'on ne possède plus de données.

## 3.8 Conclusion

Nous avons vu dans cette partie que la correction d'un modèle simple de croissance tumorale permet **d'améliorer la prédiction de l'évolution morphologique de la tumeur**. Le modèle initialement utilisé est efficace pour estimer le volume de la lésion mais se limiter à une vascularisation initiale homogène en espace nous empêche d'intégrer des changements de forme. Dans le cas où la structure autour de la tumeur est connue (par exemple matière grise et blanche), nous avons **corrigé et calibré le modèle en estimant les paramètres dans chaque zone**. Dans le cas où l'on ne possède pas de données d'imagerie permettant d'estimer l'environnement tumoral, **nous avons corrigé ce modèle et reconstitué cet environnement en n'utilisant que la forme de la tumeur**. La méthode combine une correction d'état par filtre de Luenberger, ainsi qu'une correction des paramètres du modèle par filtre de Kalman réduit particulaire. **L'étude théorique du filtre d'état de Luenberger montre la convergence de la solution vers la cible lorsque la tumeur initiale est suffisamment régulière. Les données synthétiques générées nous permettent de valider la méthode**, puisque l'on est capable de retrouver les paramètres utilisés pour construire la cible. Une telle validation est nécessaire afin de s'assurer que le principe est robuste et peut être appliqué à des données réelles. En effet, il n'est pas possible de comparer la prédiction des paramètres avec les paramètres cibles lors d'un test sur données cliniques. Cette première étape a permis en particulier de déceler quelques subtilités liées à l'implémentation, comme **le choix des *sigma points*, le choix de $\lambda$, ainsi que l'interpolation des données à partir des examens fournis**.

L'application de la méthode globale aux données cliniques montre son efficacité. La calibration volumique est tout d'abord essentielle puisque les paramètres sont initialisés à partir de l'*a priori* de cette calibration, et car l'évolution du volume sert également pour l'interpolation des données. Dans le cas réel, cette calibration marche parfaitement avec un volume final estimé très proche de la réalité. La force de la méthode réside dans le fait qu'aucun *a priori* sur la structure de $M_0$ n'est fait initialement - ce qui est conforme au cas réel - mais qu'il est possible de **reconstruire la vascularisation de la tumeur qui lui a conféréé la forme observée**. Evidemment, la structure reconstruite est grossière, puisque les coûts de calculs ne permettent pas de lancer un nombre trop important de particules, mais le comportement global semble cohérent avec l'évolution observée. **La simulation obtenue avec les paramètres finaux ainsi qu'avec correction de l'état donne une évolution en 3D de la tumeur qui est largement plus proche de la réalité que l'évolution obtenue avec le modèle initial**.

On observe cependant une limitation à cette méthode : la correction paramétrique n'étant possible que lorsque la tumeur passe dans la zone à estimer, **il est impossible de corriger l'environnement extérieur**. Or, il semble nécessaire d'avoir une estimation des réseaux de fibres et de vaisseaux autour de la tumeur si l'on souhaite prédire la forme. L'une des perspectives de ce travail consiste donc à trouver un moyen d'étendre à l'extérieur de la tumeur la vascularisation estimée à l'intérieur. Un des moyens d'y parvenir serait de distinguer une ou plusieurs directions privilégiées de la structure $M_0$ reconstruite. Dans le cas de la métastase cérébrale, on remarque que la tumeur au temps final semble d'être développée encore plus fortement dans la direction principale estimée. Une telle extension de la vascularisation pourrait donc permettre d'améliorer encore la forme prédite.

Une autre possibilité serait de coupler cette approche avec une imagerie fonctionnelle de vascularisation de basse résolution, ce qui donnerait un *a priori*. Notre méthode permettrait alors d'améliorer la résolution de cette imagerie.

La condition initiale choisie pour nos simulations en cas réel est obtenue par segmentation de la tumeur sur la premier examen clinique. Dans le cas de tumeurs diffuses, la délimitation du contour est sujette aux incertitudes. Une des perspectives est d'adapter notre méthode à de telles tumeurs. L'idée est de remplacer la condition initiale obtenue par plusieurs conditions initiales probables. Ces nouvelles conditions initiales sont calculées à partir de l'image initiale en calculant une distance géodésique qui dépend à la fois de la distance euclidienne mais aussi du gradient de l'image [37]. Ce choix s'explique par le fait que les zones de l'image à faible gradient sont les zones où la segmentation est la plus incertaine.

# Bibliographie

## Bibliographie


[1] Danilo Babin, Aleksandra Pižurica, Jonas De Vylder, Ewout Vansteenkiste, and Wilfried Philips. Brain blood vessel segmentation using line-shaped profiles. *Phys. Med. Biol.*, 58(22) :8041, 2013.

[2] Rolf H. Reichle. Data assimilation methods in the Earth sciences. *Advances in Water Resources*, 31 :1411–1418, November 2008.

[3] Mélanie Rochoux, Annabelle Collin, Cong Zhang, Arnaud Trouve, Didier Lucor, and Philippe Moireau. Front shape similarity measure for shape-oriented sensitivity analysis and data assimilation for Eikonal equation. January 2017.

[4] M.C. Rochoux, B. Cuenot, S. Ricci, A. Trouvé, B. Delmotte, S. Massart, and other authors. Data assimilation applied to combustion. *Comptes Rendus Mecanique*, 2014.

[5] Jan Mandel, Lynn S. Bennethum, Jonathan D. Beezley, Janice L. Coen, Craig C. Douglas, Minjeong Kim, and Anthony Vodacek. A Wildland Fire Model with Data Assimilation. *Math. Comput. Simul.*, 79(3) :584–606, December 2008.

[6] Philippe Arbogast, Olivier Pannekoucke, Laure Raynaud, Renaud Lalanne, and Etienne Mémin. Object-oriented processing of CRM precipitation forecasts by stochastic filtering. *Q.J.R. Meteorol. Soc.*, 142(700) :2827–2838, October 2016.

[7] Scorer R. S. Atmospheric data analysis, Roger Daley, Cambridge Atmospheric and Space Science Series, Cambridge University Press, Cambridge, 1991. No. of pages : xiv + 457. Price : £55–00, US$79–50 (hardback) ISBN 0521 382157. *International Journal of Climatology*, 12(7) :763–764, November 2006.

[8] Annabelle Collin, Dominique Chapelle, and Philippe Moireau. A Luenberger observer for reaction–diffusion models with front position data. *Journal of Computational Physics*, 300(Supplement C) :288–307, November 2015.

[9] D. Chapelle, M. Fragu, V. Mallet, and P. Moireau. Fundamental principles of data assimilation underlying the Verdandi library : applications to biophysical model personalization within euHeart. *Med Biol Eng Comput*, 51(11) :1221–1233, November 2013.

[10] J. Sainte-marie A, D. Chapelle A, R. Cimrman C, and M. Sorine A. *Modeling and estimation of the cardiac electromechanical activity*. 2005.

[11] L. Li, F. X. Le Dimet, J. Ma, and A. Vidard. A Level-Set-Based Image Assimilation Method : Potential Applications for Predicting the Movement of Oil Spills. *IEEE Transactions on Geoscience and Remote Sensing*, 55(11) :6330–6343, November 2017.

[12] M. L. Martins, S. C. Ferreira, and M. J. Vilela. Multiscale models for the growth of avascular tumors. *Physics of Life Reviews*, 4 :128–156, June 2007.

[13] F. H. Cornelis, M. Martin, O. Saut, X. Buy, M. Kind, J. Palussiere, and T. Colin. Precision of manual two-dimensional segmentations of lung and liver metastases and its impact on tumour response assessment using RECIST 1.1. *European Radiology Experimental*, 1 :16, October 2017.

[14] P. Moireau, D. Chapelle, and P. Le Tallec. Joint state and parameter estimation for distributed mechanical systems. *Computer Methods in Applied Mechanics and Engineering*, 1987(6–8) :659–677, 2008.

[15] P. Moireau and D. Chapelle. Reduced-order Unscented Kalman Filtering with application to parameter identification in large-dimensional systems. *ESAIM : Control, Optimisation and Calculus of Variations*, 17(2) :380–405, 2011.

[16] P. Moireau and D. Chapelle. Erratum of article "Reduced-order Unscented Kalman Filtering with application to parameter identification in large-dimensional systems". *ESAIM : Control, Optimisation and Calculus of Variations*, 17(2) :406–409, 2011.

[17] D. Ambrosi and L. Preziosi. On the closure of mass balance models for tumor growth. *Math. Models Methods Appl. Sci.*, 12(05) :737–754, May 2002.

[18] Thierry Colin, François Cornelis, Julien Jouganous, Jean Palussière, and Olivier Saut. Patient specific simulation of tumor growth, response to the treatment and relapse of a lung metastasis : a clinical case. *Journal of Computational Surgery*, page 18, 2015.

[19] (1) Analysis of operator splitting for advection–diffusion–reaction problems in air pollution modelling.

[20] Guang-Shan Jiang and Cheng-chin Wu. A High-Order WENO Finite Difference Scheme for the Equations of Ideal Magnetohydrodynamics. *Journal of Computational Physics*, 150 :561–594, April 1999.

[21] (2) A fifth-order accurate weighted ENN difference scheme and its applications.

[22] Julien Jouganous. *Modélisation et simulation de la croissance de métastases pulmonaires.* Bordeaux, September 2015.

[23] Stanley Osher and Ronald Fedkiw. *Level Set Methods and Dynamic Implicit Surfaces.* Applied Mathematical Sciences. Springer-Verlag, New York, 2003.

[24] T.F. Chan and L.A. Vese. Active contours without edges. *IEEE Transactions on Image Processing*, 10(2) :266–277, 1991.

[25] Hong-Kai Zhao, T. Chan, B. Merriman, and S. Osher. A Variational Level Set Approach to Multiphase Motion. *J. Comput. Phys.*, 127(1) :179–195, August 1996.

[26] Annabelle Collin. *Analyse asymptotique en électrophysiologie cardiaque : applications à la modélisation et à l'assimilation de données.* Paris 6, October 2014.

[27] D.G. Luenberger. *Determining the State of a Linear with Observers of Low Dynamic Order.* PhD thesis, Stanford University, 1963.

[28] Thomas Michel. *Analyse mathématique et calibration de modèles de croissance tumorale.* Bordeaux, November 2016.

[29] R. Kalman and R. Bucy. New results in linear filtering and prediction theory. *Trans. ASME J. Basic. Eng.*, 83 :95—108, 1961.

[30] D. Simon. *Optimal State Estimation : Kalman, $H^\infty$, and Nonlinear Approaches.* Wiley-Interscience, 2006.

[31] Simon J. Julier and Jeffrey K. Uhlmann. A New Extension of the Kalman Filter to Nonlinear Systems. pages 182–193, 1997.

[32] Björn Engquist, Anna-Karin Tornberg, and Richard Tsai. Discretization of Dirac Delta Functions in Level Set Methods. *J. Comput. Phys.*, 207(1) :28–51, July 2005.

[33] Giovanni Russo and Peter Smereka. A Remark on Computing Distance Functions. *J. Comput. Phys.*, 163(1) :51–67, September 2000.

[34] Nicolae Cîndea, Alexandre Imperiale, and Philippe Moireau. Data assimilation of time under-sampled measurements using observers, the wave-like equation example. *ESAIM : COCV*, 21(3) :635–669, July 2015.

[35] K. R. Swanson, E. C. Alvord, and J. D. Murray. A quantitative model for differential motility of gliomas in grey and white matter. *Cell Prolif.*, 33(5) :317–329, October 2000.

[36] Hana L. P. Harpold, Ellsworth C. Alvord, and Kristin R. Swanson. The Evolution of Mathematical Modeling of Glioma Proliferation and Invasion. *J Neuropathol Exp Neurol*, 66(1) :1–9, January 2007.

[37] Matthieu Lê, Jan Unkelbach, Nicholas Ayache, and Hervé Delingette. GPSSI : Gaussian Process for Sampling Segmentations of Images. In Alejandro F. Frangi, Joachim Hornegger, Nassir Navab, and William M. Wells, editors, *MICCAI - Medical Image Computing and Computer Assisted Intervention - 2015*, volume 9351 of *Lecture Notes in Computer Science - LNCS*, pages 38–46, Munich, Germany, October 2015. Springer.

# Conclusion et Perspectives

Dans cette thèse, nous avons présenté deux projets visant à utiliser au mieux les données cliniques collectées afin de prédire l'évolution tumorale. Lorsque les données sont nombreuses et acquises à un seul instant, il n'est pas possible d'utiliser un modèle mécanistique qui décrit l'évolution en temps. Les méthodes d'apprentissage automatique sont alors bien adaptées. Elles permettent de prendre en compte une grosse quantité d'informations qui peuvent être variées. Dans notre exemple, des informations d'imagerie médicales ont été combinées à des données issues de la génomique de la tumeur. **L'indicateur d'hétérogénéité tumorale que l'on a construit** permet de prendre en compte la structure spatiale de la tumeur et apparaît corrélé au PFS. **Le prétraitement des données par double réduction de l'espace des caractéristiques** permet de réduire le bruit des données, et de comparer plusieurs algorithmes de classification. **Plusieurs tests de comparaison de classification ont ensuite été appliqués** afin de confirmer la **meilleure prédiction de notre méthode par rapport à ce qui est actuellement utilisé en clinique.** La force principale de cette méthode est aussi sa limite : les méthodes algorithmiques utilisées sont agnostiques et ne prennent pas en compte la spécificité de la biologie du cancer. En particulier, cela ne permet pas de tirer plus d'informations que celles prédites par apprentissage.

Dans le second projet, les données sont cette fois longitudinales et un modèle est construit pour prédire la croissance tumorale de manière spécifique pour chaque patient. Toute l'information contenue dans les examens disponibles du patient est utilisée pour corriger et estimer les paramètres du modèle afin d'expliquer les changements de forme observés. **On a montré théoriquement que lorsque la tumeur initiale est suffisamment régulière, le terme de correction d'état corrige effectivement le contour**. **La correction état-paramètres est validée sur des cas synthétiques**, puisque l'évolution cible est retrouvée même lorsque l'état initial ou les paramètres initiaux sont décalés. La résolution numérique a nécessité **l'interpolation des données**, et on a montré **les avantages et inconvénients des différents types de *sigma points***. L'estimation de la vascularisation initiale sur une grille permet de **prendre en compte l'hétérogénéité**

**spatiale du milieu dans le modèle**. **L'amélioration de la prédiction sur des cas réels** confirme l'intérêt de la méthode globale développée. La limite ici est de ne pouvoir prendre en compte que les données d'imagerie, qui ne permettent pas d'extraire d'informations sur l'hétérogénéité de la tumeur ou de l'environnement. En particulier, cela nous oblige à tenter de reconstruire la vascularisation tumorale.

Ces deux exemples montrent bien les problèmes auxquels on fait face lorsque l'on souhaite prédire un comportement à partir de données cliniques : comment construire un modèle mécanistique qui a du sens biologiquement, qui est calibrable et qui utilise le maximum de données acquises en routine clinique. L'objectif des perspectives est de voir comment l'apprentissage automatique et la modélisation mathématique peuvent être combinés. Nous distinguons plusieurs méthodes qui peuvent être envisagées.

La première consiste à ajouter de nouvelles caractéristiques issues d'un modèle. Si l'on possède des données longitudinales, il est en effet possible de calibrer un modèle simple de croissance sur ces données et d'ajouter les paramètres régissant l'évolution de chaque patient (taux de prolifération des cellules cancéreuses, taux de consommation des nutriments, taux de dissémination, etc...) à la liste des caractéristiques. Ces informations ne sont pas mesurables en pratique et l'utilisation de modèles permet de rendre compte de cette spécificité pour chaque patient.

Dans la deuxième partie, la réduction du modèle 3D à un modèle OD s'est fait par intégration de l'équation aux dérivées partielles. Lorsque le modèle n'est pas intégrable, il est possible d'utiliser des méthodes d'apprentissage afin d'estimer les paramètres du modèle. La base d'apprentissage est construite en lançant un grand nombre de simulations du modèle avec différents jeux de paramètres. A partir d'une évolution tumorale, les algorithmes d'apprentissage permettent alors d'estimer le jeu de paramètres adapté au patient.

Une autre possibilité lorsque l'on possède une grosse base de données est d'utiliser des algorithmes d'apprentissage afin de trouver une corrélation entre ces paramètres et les caractéristiques d'imagerie ou génomique des patients. L'intérêt de cette méthode est qu'elle permet de donner un sens biologique aux paramètres puisqu'ils seront obtenus par extraction et combinaison des données cliniques.

Dans les deux méthodes de calibrage précédentes, la sensibilité des paramètres peut rendre difficile l'estimation par apprentissage automatique. Une autre idée consiste donc à utiliser l'apprentissage pour simplement réduire l'espace de recherche des paramètres afin de fa-

ciliter le calibrage. On peut par exemple détecter deux types généraux de comportements tumoraux et essayer de classer chaque patient dans une de ces deux classes, afin d'avoir un *a priori* sur les paramètres à estimer.

Enfin, il est possible de remplacer la boite noire agnostique de l'apprentissage automatique par un algorithme qui dépend d'un modèle mécanistique. La prédiction est alors obtenue à partir d'un modèle qui est spécifique au problème considéré, et donc *a priori* plus pertinent qu'une approche généraliste.

# Annexes

Dans cette annexe, nous allons démontrer les lemmes et le théorème non prouvés dans le chapitre 3.

**Rappel : Lemme 1.** *Supposons que $P_0 \in H^s(\mathcal{B})$ où $\mathcal{B} \subset \mathbb{R}^n$, avec $s > \frac{n}{2}$ et soit $C = 2||P_0||_{H^s}$. Alors il existe $t_0$ tel que le compact $K_C = \{P(t,x), ||P||_{L^\infty_{t_0},H^s} \leqslant C\}$ soit stable par $\Phi \circ \Psi$.*

*Démonstration.* Cette démonstration est inspirée de [1]. Soit $Q \in K_C$ et notons $P = \Phi \circ \Psi(Q)$. Notons $\mathbf{v} = \Psi(Q)$. Soit $m \leqslant s$. Notons $\partial_x^m$ une dérivée partielle quelconque d'ordre $m$. En dérivant l'équation sur $P$ définie dans (3.61), on obtient :

$$\partial_x^m \partial_t P + \partial_x^m \nabla \cdot (\mathbf{v}\, P) = M \partial_x^m P. \tag{93}$$

En multipliant cette équation par $\partial_x^m P$ et en l'intégrant sur $\mathcal{B}$, cela donne :

$$\frac{1}{2}\frac{d}{dt}\left(||\partial_x^m P||_2^2\right) + \int_{\mathcal{B}} \partial_x^m \nabla \cdot (\mathbf{v}\, P) \partial_x^m P dx = M||\partial_x^m P||_2^2. \tag{94}$$

Evaluons le second terme de cette égalité en utilisant la formule de Leibniz :

$$\begin{aligned} \int_{\mathcal{B}} \partial_x^m \nabla \cdot (\mathbf{v}\, P) \partial_x^m P dx &= \int_{\mathcal{B}} \nabla \cdot (\partial_x^m (\mathbf{v}\, P)) \partial_x^m P dx, \\ &= \sum_{k=0}^{m} \int_{\mathcal{B}} \nabla \cdot (\partial_x^k \mathbf{v}\, \partial_x^{m-k} P) \partial_x^m P dx, \\ &=: \sum_{k=0}^{m} I_m^k. \end{aligned} \tag{95}$$

— Cas $k = 0$

Quand $k = 0$, le terme se réécrit :

$$\begin{aligned} I_m^0 &= \int_{\mathcal{B}} \nabla \cdot (\mathbf{v}\, \partial_x^m P) \partial_x^m P dx, \\ &= \int_{\mathcal{B}} (\nabla \cdot \mathbf{v})(\partial_x^m P)^2 dx + \int_{\mathcal{B}} \mathbf{v} \cdot \frac{1}{2} \nabla((\partial_x^m P)^2), \\ &= \int_{\mathcal{B}} (\nabla \cdot \mathbf{v})(\partial_x^m P)^2 dx - \frac{1}{2} \int_{\mathcal{B}} (\nabla \cdot \mathbf{v})(\partial_x^m P)^2 dx, \\ &= \frac{1}{2} \int_{\mathcal{B}} (\nabla \cdot \mathbf{v})(\partial_x^m P)^2 dx, \end{aligned} \tag{96}$$

en utilisant une intégration par partie. On obtient alors :

$$|I_m^0| \leqslant \frac{1}{2} ||v||_{H^{s+1}} ||P||_{H^m}^2. \tag{97}$$

— Cas $k \geqslant 1$

Dans le cas général, nous utilisons l'inégalité d'interpolation de Gagliardo-Nirenberg [2] : si $u \in C_c^\infty(\mathbb{R}^n)$ et si $p, q, m, j$ sont des entiers vérifiant $\frac{1}{p} = \frac{j}{n} + \frac{1}{q} - \frac{r}{n}$ et avec $j \leqslant r$, alors il existe $C$ tel que :

$$||\partial_x^j u||_{L^p} \leqslant C ||\partial_x^r u||_{L^q}. \tag{98}$$

Si $k \geqslant 1$, on obtient :

$$\begin{aligned} I_m^k &= \int_{\mathcal{B}} \nabla \cdot (\partial_x^k \mathbf{v}\, \partial_x^{m-k} P) \partial_x^m P dx, \\ &= \int_{\mathcal{B}} (\nabla \cdot \partial_x^k \mathbf{v})\, \partial_x^{m-k} P\, \partial_x^m P dx + \int_{\mathcal{B}} \partial_x^k \mathbf{v} \cdot \nabla(\partial_x^{m-k} P) \partial_x^m P dx, \\ &=: I_{1m}^k + I_{2m}^k, \end{aligned} \tag{99}$$

— si $k = m$ et $m \geqslant 2$ :

$$|I_{1m}^m| \leqslant ||\nabla \cdot \partial_x^m \mathbf{v}||_{L^2} ||P||_{L^\infty} ||\partial_x^m P||_{L^2} \leqslant ||v||_{H^{s+1}} ||P||_{H^m}^2, \tag{100}$$

en utilisant le plongement continu $H^m \hookrightarrow L^\infty$, puisque $m \geqslant 2$.

— sinon :

$$|I_{1m}^k| \leqslant ||\nabla \cdot \partial_x^k \mathbf{v}||_{L^4} ||\partial_x^{m-k} P||_{L^4} ||\partial_x^m P||_{L^2} \leqslant C_1 ||v||_{H^{s+1}} ||P||_{H^m}^2, \tag{101}$$

en utilisant l'inégalité d'interpolation de Gagliargo-Nirenberg avec $j = k + 1$, $p = 4$, $q = 2$ et $r = k + 1 + \frac{n}{4} \leqslant s + 1$ pour le premier terme, et avec $j = m - k$, $p = 4$, $q = 2$ et $r = m - k + \frac{n}{4} \leqslant s$ pour le second terme.

De même, pour le second membre :

— si $k = 1$ :

$$|I_{2m}^1| \leqslant ||\partial_x \mathbf{v}||_{L^\infty} ||\nabla(\partial_x^{m-1} P)||_{L^2} ||\partial_x^m P||_{L^2} \leqslant ||v||_{H^{s+1}} ||P||_{H^m}^2, \tag{102}$$

en utilisant le plongement continu $H^s \hookrightarrow L^\infty$.

— sinon, quand $k \geqslant 2$ :

$$|I_{2m}^k| \leqslant ||\partial_x^k \mathbf{v}||_{L^4} ||\nabla(\partial_x^{m-k} P)||_{L^4} ||\partial_x^m P||_{L^2} \leqslant C_2 ||v||_{H^{s+1}} ||P||_{H^m}^2, \tag{103}$$

en utilisant l'inégalité d'interpolation de Gagliargo-Nirenberg avec $j = k$, $p = 4$, $q = 2$ et $r = k + \frac{n}{4} \leqslant s + 1$ pour le premier terme, et avec $j = m - k + 1$, $p = 4$, $q = 2$ et $r = m - k + 1 + \frac{n}{4} \leqslant s$ pour le second terme.

En additionnant les inégalités ci-dessus, on obtient l'existence d'une constante $C_0 \in \mathbb{R}$ telle que $\forall t \in [0, t_0]$ :

$$\partial_t ||P||_{H^s}^2 \leqslant ||P||_{H^s}^2 (C_0 ||v||_{H^{s+1}} + 2m_0 \exp(-\alpha t)). \tag{104}$$

Or, puisque $\mathbf{v} = \Psi(Q)$, on a :

$$\nabla \cdot \mathbf{v} = MQ, \tag{105}$$

et donc

$$||\mathbf{v}||_{H^{s+1}} \leqslant M ||Q||_{H^s}. \tag{106}$$

Puisque $Q \in K_C$ on obtient alors :

$$\begin{aligned} \partial_t ||P||_{H^s}^2 &\leqslant ||P||_{H^s}^2 m_0 (C_0 C + 2), \\ \partial_t ||P||_{H^s}^2 &\leqslant C' ||P||_{H^s}^2. \end{aligned} \tag{107}$$

Finalement, on a donc $\forall t \in \mathbb{R}$ :

$$||P||_{L^\infty_{[0,t]}, H^s} \leqslant ||P_0||_{H^s} \exp\left(\frac{C't}{2}\right) \leqslant \frac{C}{2} \exp\left(\frac{C't}{2}\right), \tag{108}$$

donc pour $t = t_0 \leqslant \frac{2\ln(2)}{C'}$, la propriété est vérifiée. □

**Rappel : Lemme 2.** *Soit $t_0$ le temps fixé par le lemme précédent. Il existe $t_1 \in [0, t_0]$ tel que l'application $\Phi \circ \Psi$ soit contractante sur $K_C = \{P(t,x), ||P||_{L^\infty_{[0,t_1]},H^s} \leqslant C\}$.*

*Démonstration.* Soient $P_1, P_2 \in K_C$ et notons $Q_1 = \Phi \circ \Psi(P_1)$ et $Q_2 = \Phi \circ \Psi(P_2)$. D'après le lemme précédent, pour $t \in [0, t_0]$, on a : $Q_1, Q_2 \in K_C$. Notons $u_1 = \Psi(P_1)$ et $u_2 = \Psi(P_2)$. Enfin, notons $Q = Q_1 - Q_2$. On a :

$$\begin{aligned} \partial_t Q_1 + \nabla \cdot (u_1 Q_1) &= M Q_1, \\ \partial_t Q_2 + \nabla \cdot (u_2 Q_2) &= M Q_2, \end{aligned} \tag{109}$$

d'où

$$\partial_t Q + \nabla \cdot (u Q_1) \nabla \cdot (u_2 Q) = M Q_1, \tag{110}$$

avec $Q(0,x) = 0$. En multipliant l'équation par $Q$ et en l'intégrant sur $\Omega$, on obtient, en utilisant le même raisonnement qu'au lemme précédent :

$$\partial_t ||Q||^2_{H^s} \leqslant ||Q||^2_{H^s}(m_0(C_0||u_2||_{H^{s+1}} + 2)) + C_0||Q||_{H^s}||u||_{H^{s+1}}||Q_1||^s_H, \tag{111}$$

et en utilisant l'inégalité arithmético-géométrique $||Q||_{H^s}||u||_{H^{s+1}} \leqslant \frac{||Q||^2_{H^s}}{4} + ||u||^2_{H^{s+1}}$, on obtient :

$$\partial_t ||Q||^2_{H^s} \leqslant ||Q||^2_{H^s} \left( m_0(C_0||u_2||_{H^{s+1}} + 2) + \frac{C_0 C^2}{2} \right) + C C_0 ||u||_{L^\infty_{[0,t_0]} H^{s+1}}, \tag{112}$$

puisque $Q_1$ et $Q_2$ sont dans $K_C$. On obtient alors en intégrant l'existence de constantes $k_1$ et $k_2$ telles que, $\forall t \in [0, t_0]$ :

$$||Q||^2_{L^\infty_{[0,t]},H^s} \leqslant k_1 ||u||_{L^\infty_{[0,t]},H^{s+1}} \left(\exp(k_2 t) - 1\right). \tag{113}$$

Or, par définition, $||u||_{L^\infty_{[0,t]},H^{s+1}} \leqslant m_0 ||P_1 - P_2||_{L^\infty_{[0,t]},H^{s+1}}$, d'où finalement :

$$||Q_1 - Q_2||_{L^\infty_{[0,t]},H^s} \leqslant k_1' ||P_1 - P_2||_{L^\infty_{[0,t]},H^s} \left(\exp(k_2 t) - 1\right), \tag{114}$$

donc pour $t = t_1 = \min \left( t_0, \frac{\ln\left(1 + \frac{1}{k_1'}\right)}{k_2} \right)$, l'application est contractante. □

**Rappel : Théorème 1.** *Supposons que $P_0 \in H^s$ avec $s > \frac{n}{2}$. Alors il existe un temps $t_2 \in \mathbb{R}$ tel que la solution au système (3.48), notée $P$ satisfasse : $||P||_{L^\infty_{[0,t_2]},H^s}$ bornée.*

*Démonstration.* D'après les lemmes 1 et 2, il existe $t_1$ tel que $K_C$ soit stable par l'application $\Phi \circ \Psi$ et tel que cette dernière soit contractante sur $K_C$. Par le théorème du point fixe, on en déduit que la solution au système (3.48), notée $P$ satisfait :

$$||P||_{L^\infty_{[0,t_1]},H^s} \leqslant C.$$

D'après l'équation (104), on a $\forall t \in \mathbb{R}$ :

$$\partial_t ||P||^2_{H^s} \leqslant ||P||^2_{H^s}(C_0||v||_{H^{s+1}} + 2m_0), \tag{115}$$

avec cette fois :

$$||v||_{H^{s+1}} \leqslant m_0||P||_{H^s}, \tag{116}$$

donc :

$$\partial_t ||P||^2_{H^s} \leqslant ||P||^2_{H^s}(m_0(C_0||P||_{H^s} + 2). \tag{117}$$

Lorsque $||P||_{H^s} \leqslant 1$, on a bien $||P||_{H^s}$ bornée. Lorsque $||P||_{H^s} \geqslant 1$, on a la majoration :

$$\partial_t ||P||^2_{H^s} \leqslant C'||P||^3_{H^s}. \tag{118}$$

En intégrant cette inégalité, on obtient :

$$||P||^2_{H^s} \leqslant \frac{||P_0||^2_{H^s}}{(1 - ||P_0||^2_{H^s}C't)^2}, \tag{119}$$

donc lorsque $t = t_2 < \frac{1}{||P_0||^2_{H^s}C'}$, la solution $P$ est prolongeable sur $[0, t_2]$ en une fonction telle que $||P||_{L^\infty_{t_2},H^s}$ soit borné. □

## Bibliographie

[1] Thomas Michel. *Analyse mathématique et calibration de modèles de croissance tumorale*. Bordeaux, November 2016.

[2] Serge Alinhac, Patrick Gérard, and Stephen S Wilson. *Pseudo-differential operators and the Nash-Moser theorem*. 2007. OCLC : 493662128.