

Interpolaspline

Rapport du stage applicatif

L3 MIN

Décembre 2019

CORBILLE Clément, DOUMBOUYA Mohamed, EL BOUCHOUARI
Zakaria, HEDDIA Bilel, PIASENTIN Béryl, RODET Amélys

Table des matières

Introduction	2
1 Organisation de la semaine	2
2 Contexte et objectifs	3
3 Définition des tâches	4
3.1 Création des tâches	4
3.2 Contenu des tâches	4
4 Cahier des charges	6
5 Planification des tâches	7
5.1 Organisation des tâches	7
5.1.1 Construction du diagramme de Gantt	7
5.1.2 Graphe des dépendances	8
5.2 Répartition des tâches	10
6 Risques	10
Conclusion	10
Annexes	10

Introduction

Dans le cadre de notre troisième année de licence mathématiques-informatique à l'Université Grenoble Alpes, nous devons effectuer un stage applicatif à l'UFR IM²AG. Ce stage, qui consiste en un projet, est une étape indispensable pour mettre en pratique les connaissances acquises durant cette formation et pour acquérir de nouvelles compétences de recherche. Les projets ont lieu durant quatre semaines, réparties en deux périodes de respectivement une puis trois semaines. Des groupes de six étudiants ont été construits en fonction des affinités de chacun pour les différents sujets proposés.

Le sujet de notre projet est l'interpolation avec données aberrantes. Il nous donne l'opportunité d'approfondir nos connaissances en Mathématiques appliquées, plus précisément d'approfondir l'enseignement d'algèbre linéaire pour le graphique et la CAO (Conception Assistée par Ordinateur) que nous avons reçu, ainsi que la programmation en langage python. De plus, ce stage va nous permettre d'acquérir des compétences en travail de groupe.

1 Organisation de la semaine

L'objectif de cette semaine de stage était de mettre en place l'organisation nécessaire à la réalisation du projet qui aura lieu en avril.

Nous avons commencé par désigner un chef de projet : Béryl a été choisie à cet effet. La chef de projet veille sur le bon déroulement de chaque tâche en plus de sa participation aux différentes tâches. Nous avons ensuite défini un nom pour le projet, en rapport avec le contenu mais également percutant. "Interpolaspline" a été le nom retenu, en tant que fusion du mot "interpolation" et du mot "spline", deux notions principales du sujet.

Durant cette semaine d'organisation, chaque membre du groupe a été responsable d'une ou de plusieurs tâches. Être responsable d'une tâche signifie s'assurer que celle-ci est correctement réalisée et terminée dans les délais fixés, mais pas obligatoirement qu'elle doit être réalisée seul.

Les premiers jours ont été mis à profit pour comprendre le sujet et trouver ce sur quoi nous allons travailler pendant trois semaines. Il a fallu ensuite établir un plan de travail. Cette réflexion nous a permis de construire un diagramme de Gantt et un graphe de dépendances des tâches, mais aussi de répartir les tâches entre les différents membres du groupe. Chacun sait donc de quelle tâche il est responsable et peut se projeter dans la période de réalisation.

Une fois tous ces points mis en place, nous avons établi le cahier de charges. Nous y avons décrit le problème, nos objectifs, le cadre du projet. La description des fonctionnalités livrées à la fin du projet y est également détaillée, ainsi que l'organisation temporelle du projet.

Tous les membres du groupes ont participé à la rédaction du rapport basé sur cette semaine de travail, incluant le cahier des charges, mais aussi à la préparation de la soutenance. Pour terminer cette semaine d'organisation, chacun a pris le temps de bien comprendre ce qu'il devrait faire durant les trois semaines de réalisation.

2 Contexte et objectifs

Ce stage applicatif est un court projet permettant de mettre en pratique des connaissances acquises durant la troisième année de licence en parcours mathématiques-informatique, et par ailleurs d'en acquérir sur des notions annexes au programme pédagogique de la licence, au cours de sa réalisation. Le sujet principal du projet sera l'interpolation (par spline de lissage) de données contenant des points aberrants, consistant en particulier en l'apprentissage de la gestion de ces derniers, en une et deux dimensions. Notre projet ne possède pas de client, mais le travail effectué sera évalué par une équipe d'enseignants en charge de superviser notre stage applicatif. Ces enseignants feront partie de notre jury lors de la soutenance finale. La seule hiérarchie apparente au sein de notre groupe sera notre chef de projet, Beryl PIASENTIN. Ci-dessous le reste de l'équipe :

- Clément CORBILLE
- Mohamed DOUMBOUYA
- Zakaria EL BOUCHOUARI
- Bilel HEDDIA
- Amelys RODET

Ce stage applicatif sera effectué dans son intégralité dans le bâtiment IM²AG à l'UFR. L'essentiel du travail sera réalisé sur les postes informatiques du bâtiment, aucune expérience particulière n'est prévue. Le sujet du projet ne nécessite pas circonstances temporelles autre que les périodes qui sont prévues : nous avons une semaine consacrée à la gestion de notre projet du 10 au 18 décembre puis 3 semaines en avril consacrées à sa réalisation (du 6 au 10, du 14 au 17 puis du 27 au 30).

Il n'est pas question de budget dans ce projet. La seule ressource (autre que humaine) sera matérielle. Le matériel nécessaire comprend les postes informatiques (un par personne) et les logiciels intégrés dans ceux-ci qui nous permettront de réaliser le travail demandé. Nous avons besoin de Python, dont les librairies usuelles de mathématiques (numpy), d'affichage (matplotlib), et d'interface graphique (tkinter).

Un objectif à court terme de ce stage applicatif est de le terminer dans les temps imposés par le corps enseignant. Un objectif à long terme est d'acquérir de la pratique concernant la gestion d'un projet similaire à celui d'un stage.

3 Définition des tâches

3.1 Création des tâches

Dans le but d'interpoler des données en utilisant des splines de lissage tenant compte des valeurs aberrantes, nous nous devons de savoir comment aborder le sujet. En discutant entre membres du groupe, nous avons convenu qu'il fallait dans un premier temps approfondir les notions générales sur les splines naturelles et de lissage en fonction de la répartition des points et trouver un moyen d'automatiser l'estimation du paramètre de lissage. Puis nous avons prévu d'étudier des méthodes d'identification des valeurs aberrantes que l'on souhaite à terme implémenter en deux dimensions, mais aussi des méthodes permettant d'interpoler des splines sans commencer par détecter les valeurs aberrantes.

3.2 Contenu des tâches

Les tâches sont les suivantes :

- Approfondissement des splines naturelles
- Approfondissement des splines de lissage uniformes
- Approfondissement des splines de lissages non-uniformes
- Estimation automatique du paramètre de lissage
- Identification des points aberrants
- Etude des splines de lissage avec points aberrants
- Etude de l'algorithme RANSAC
- Redéfinition d'autres fonctions d'erreur à minimiser
- Finalisation du projet

Voici quelques explications du contenu scientifique constituant la base des tâches listées précédemment.

Splines naturelles

L'interpolation est une opération mathématique consistant à déterminer une fonction passant par un nombre fini de points donnés.

Soient (x_i, y_i) les $n + 1$ données, $i \in \llbracket 0; n \rrbracket$. Une spline est une fonction définie par morceaux sur chaque intervalle $[x_i, x_{i+1}]$, $i \in \llbracket 0; n - 1 \rrbracket$. Le principe d'une spline est de combiner plusieurs morceaux de fonctions (en général on utilise des polynômes de petit degré) pour former une fonction d'une classe définie.

Les splines cubiques C^2 sont les splines définies par des polynômes cubiques dont la dérivée seconde est continue. On appelle spline naturelle la spline cubique C^2 dont les dérivées secondes aux extrémités de l'intervalle de définition sont nulles. La spline naturelle existe toujours pour des données distinctes, et est unique. Le principe de la spline cubique naturelle permet d'obtenir un système tridiagonal simple qui peut être résolu facilement, afin d'obtenir les coefficients des polynômes.

Splines de lissage

Les splines de lissage sont des splines cubiques naturelles dont chaque polynôme cubique est une approximation des données se trouvant sur son intervalle de définition. Cette spline ne passera dans la plupart des cas pas par tous les points, mais est la spline qui minimise une quantité liée à la distance entre les données et la spline. En général, la spline approximant les données cherche à minimiser l'erreur au carré : c'est l'approximation aux moindres carrés.

Les splines de lissage permettent d'éviter les oscillations qui seraient présentes avec une spline naturelle passant par tous les points, provoquées avec un nombre de données très grand.

Paramètre de lissage

La quantité à minimiser pour construire la spline approximant des données est composée de deux parties dont la seconde est pondérée par un paramètre de lissage : la première assure le passage de la spline par nos données tandis que la seconde assure le lissage de la courbe. Cela signifie que plus on augmente la valeur du paramètre de lissage, plus le lissage sera fort ce qui donne par suite une spline passant de moins en moins proche des données.

Données aberrantes

En statistiques, une donnée aberrante est une valeur ou une observation qui est « distante » des autres observations effectuées sur le même phénomène, c'est-à-dire que c'est une valeur qui s'écarte fortement des valeurs des autres observations : elle est anormalement faible ou élevée par rapport à celles qui lui sont proches. Sur un graphe on peut les visualiser comme des points qui se situent loin du nuage de points principal.

Les méthodes pour créer une spline de lissage malgré les données aberrantes sont variées : certaines déterminent les points aberrants pour ensuite les supprimer ou leur associer un poids faible tandis que d'autres créent la spline tout en gérant les points aberrants, sans les détecter auparavant.

RANSAC

RANSAC est l'abréviation de RANdom SAmple Consensus, c'est à dire consensus d'échantillons aléatoires. C'est le nom d'un algorithme qui gère les points aberrants sans prendre le temps de les détecter dans un premier temps. Cet algorithme est itératif, et non-déterministe. Un algorithme non déterministe est un algorithme qui ne produit pas toujours la même sortie pour une même entrée.

Interpolation paramétrique

L'objectif de l'interpolation paramétrique est de trouver un polynôme paramétrique qui interpole les données. Pour ce faire, deux méthodes sont possibles :

- L'interpolation séparée des n données (t_i, x_i) et des n données (t_i, y_i) , $i \in \llbracket 0; n \rrbracket$, avec t un paramètre bien choisi. On obtient deux splines P et Q interpolant respectivement les données (t_i, x_i) et (t_i, y_i) , $i \in \llbracket 0; n \rrbracket$. En associant les deux splines obtenues grâce au paramètre t , on obtient une spline paramétrique S interpolant les données de départ, $S(t) = (Q(t), R(t))$. En particulier, $P(t_i) = (x_i, y_i)$, $i \in \llbracket 0; n \rrbracket$.
- L'interpolation en une fois des données, sans appliquer la méthode à une dimension sur chaque dimension.

4 Cahier des charges

Le cahier des charges est l'annexe n° 4. Nous y avons détaillé le projet, du sujet à l'archive rendue, en passant par sa réalisation. Ce cahier des charges représente notre compréhension globale du sujet ainsi que les principales idées d'organisation et de réalisation.

Ce document est censé être le lien entre notre équipe et le client, mais il est ici le lien entre notre équipe et l'équipe d'enseignants responsable du stage applicatif. Nous allons l'utiliser durant la seconde fin du projet. Notre objectif est de réaliser tous les objectifs qui y sont définis, dans les délais impartis.

5 Planification des tâches

5.1 Organisation des tâches

5.1.1 Construction du diagramme de Gantt

Il nous faut donc planifier de façon la plus optimale possible nos tâches à accomplir sur les dates qui nous sont allouées pour le projet. Le diagramme de Gantt sera notre principal outil.

La construction du diagramme fut basée sur un développement par prototypes successifs, c'est-à-dire de façon itérative, nous sommes “revenus dessus” plusieurs fois pour l'affiner tout en déterminant et analysant les dépendances des tâches.

Ci-joint en annexe 1 notre diagramme de Gantt terminé.

Nous l'avons représenté sous la forme d'un tableau à double entrée comprenant le détail des tâches (et leur numéro de tâche respectif) en ordonnée et la chronologie représentée par les jours de projet en abscisse.

La hiérarchie des tâches est montrée. Ainsi la tâche 1 regroupe 1a et 1b et 5a contenue dans la 5 regroupe 5aa, 5ab, 5ac et 5ad. La durée de ces tâches de haut niveau sont représentées par des bandelettes jaunes sur le diagramme. Les tâches de niveau inférieur sont planifiées par des cases grises contenant un chiffre désignant le nombre de personne nous paraissant indispensable pour sa réalisation.

Pour gérer les ressources humaines, nous avons ajouté une ligne à la fin du diagramme pour contrôler les effectifs affectés et aider à la prise de décision pour la répartition du travail.

Un diagramme de Gantt “réduit” comprenant uniquement les tâches de plus haut niveau est également fourni en annexe 2.

5.1.2 Graphe des dépendances

Le graphe des dépendances a pour objectif de représenter ces dépendances de la façon la plus claire possible. Malgré les nombreuses dépendances, nous avons trouvé un moyen pour rendre ce graphe lisible et facilement compréhensible.

Tâches "parents" :

Pour trouver les tâches nécessaires au début d'une, il suffit de remonter dans l'arborescence. Par exemple dans l'exemple ci-dessous, la tâche 6d a besoin de 6bb, 6c et 6aa.

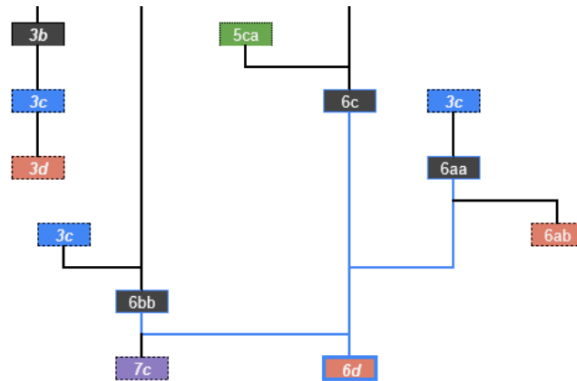


FIGURE 1 – Exemple de tâches "parents"

Tâches "fils" :

De même, pour trouver les tâches découlant de la réalisation d'une précédente, il suffit de descendre dans le graphe.

Par exemple dans l'exemple ci-après, la réalisation de 4b permet de commencer 4c et 5bb (même si 5bb aura également besoin de la réalisation de 5ba).

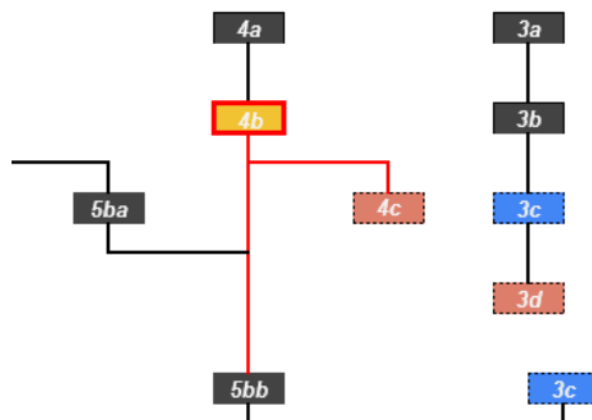


FIGURE 2 – Exemple de tâches "fils"

Tâches "à-multiples-apparitions" :

Le meilleur moyen que nous avons trouvé pour éviter le surplus de connexions dans le graphe dû aux multiples dépendances, est de faire apparaître une tâche plusieurs fois. Voilà pourquoi un jeu de couleur à été employé pour permettre une meilleure visibilité de ces tâches "complexes".

Par exemple dans le cas ci-dessous, la tâche 3c dépend de la fin des tâches 0a,5ac,5bc et 3b et sa réalisation est nécessaire pour commencer 6bb et 6aa.

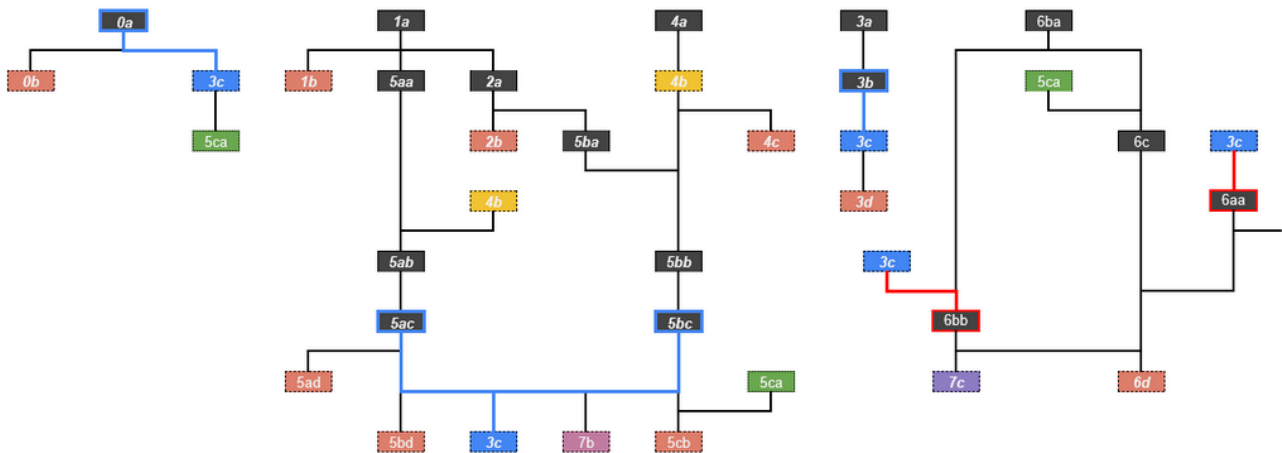


FIGURE 3 – Exemple de tâches "à-multiples-apparitions"

Le graphe des dépendances complet est visible en annexe 4 .

Moyens de communication :

Même si nous avons sérieusement planifié notre projet, nous avons conscience qu'il est rare que tout se passe comme prévu. Pour éviter l'effet tunnel, autrement dit anticiper les retards et réagir au plus vite sur les ressources nécessaires sur une même tâche, nous mettrons en place des bilans sur le travail de chacun lors que chaque demi-journée et ainsi leur permettre de partager au groupe leur avancée. Des réunions plus complète sur celle du projet seront également mises en place, cette fois-ci tous les deux jours.

Nous utiliserons Discord comme moyen de communication inter-groupe. Pour ce qui est des fichiers du projet, nous utiliserons le service web d'hébergement Github (et également les applis de Google drive pour la mise en commun et la modification en temps réel sur les fichiers destinés au rapport final).

5.2 Répartition des tâches

Pour éviter les risques, la répartition du travail s'est effectuée dans l'optique d'avoir un minimum de demi-journées où la totalité des membres de groupe est responsable d'une tâche, de sorte à pouvoir répondre aux risques de retard sur l'avancée d'une tâche.

Le groupe entier a pu participer à la réunion visant à répartir le travail, pour que les membres puissent choisir de travailler sur les tâches où ils sentiront les plus efficaces, pour également limiter les risques de retard.

Le tableau légendé fourni en annexe 3 est le résultat de cette réunion.

6 Risques

Le risque majeur du projet est de ne pas avoir estimé correctement le temps nécessaire pour réaliser certaines tâches. En effet, si une fonctionnalité met trop de temps à être implémentée, certaines tâches qui en dépendent risquent de prendre du retard, cela est l'effet boule de neige. Pour éviter cela, le diagramme de Gantt prévoit des membres "flottants", qui se greffent sur les tâches posant des difficultés.

Conclusion

Nous avons confiance en l'aboutissement du projet. En effet, nous avons construit une bonne cohésion de groupe, malgré le fait que l'on n'ait pas l'habitude de travailler ensemble. De plus, tous les membres du groupes ont bien compris la notion de base du sujet (les splines), développée durant l'enseignement d'algèbre linéaire pour le graphique et la CAO. Certains d'entre nous ont également déjà reçu des enseignements de statistiques approfondis. Cela nous permettra de mieux appréhender l'identification des données aberrantes, la répartition des tâches étant faite en fonction des capacités de chacun.

Dans l'idéal, nous souhaiterions réussir à développer le projet d'une manière qui permette de pouvoir ensuite facilement traiter des données dans n'importe quelle dimension.

Annexes

1. Diagramme de Gantt (en 2 parties)
2. Diagramme de Gantt simplifié et répartition des tâches
3. Graphe de dépendance des tâches
4. Cahier des charges

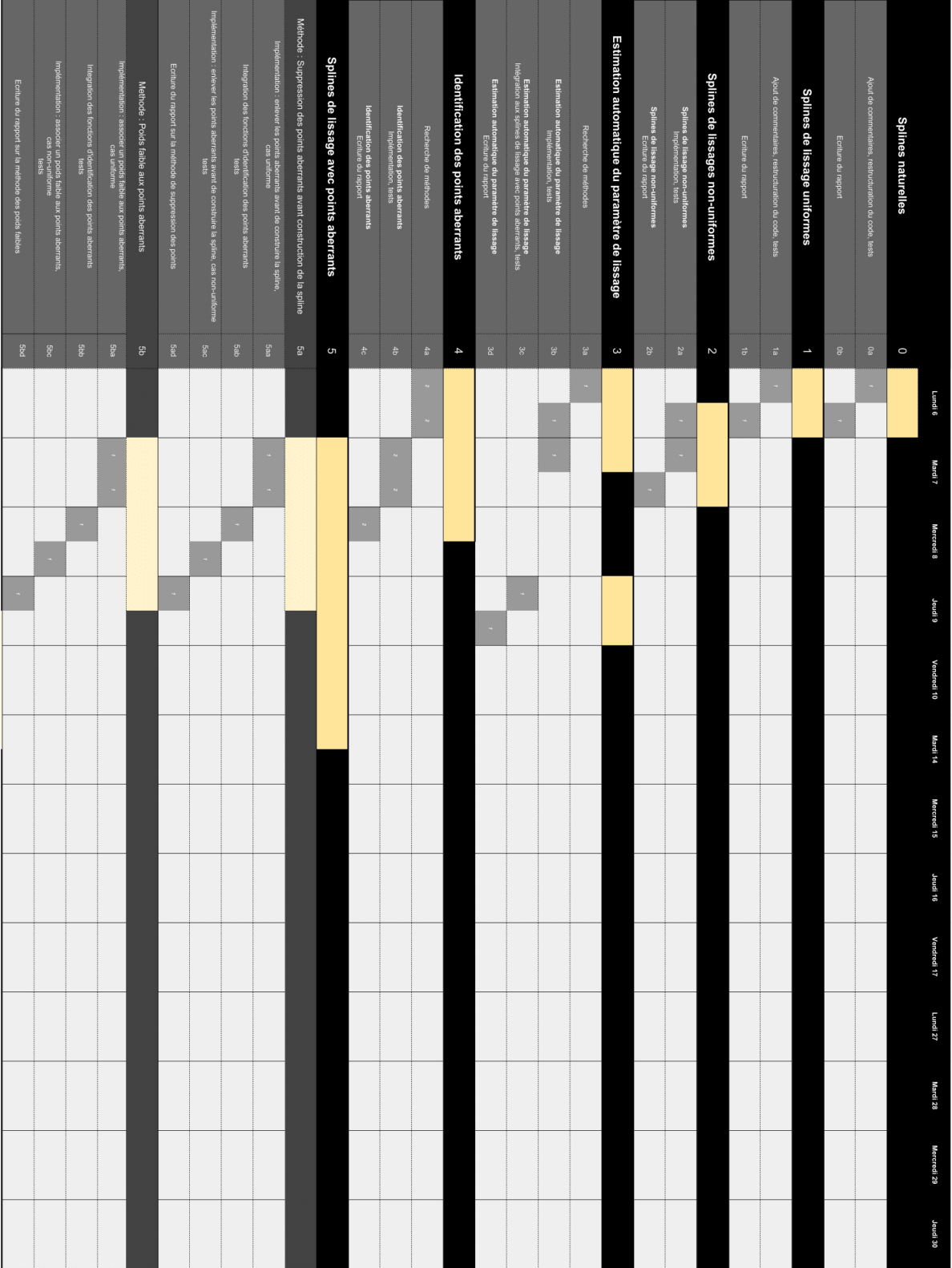


FIGURE 4 – Diagramme de Gantt (partie 1)

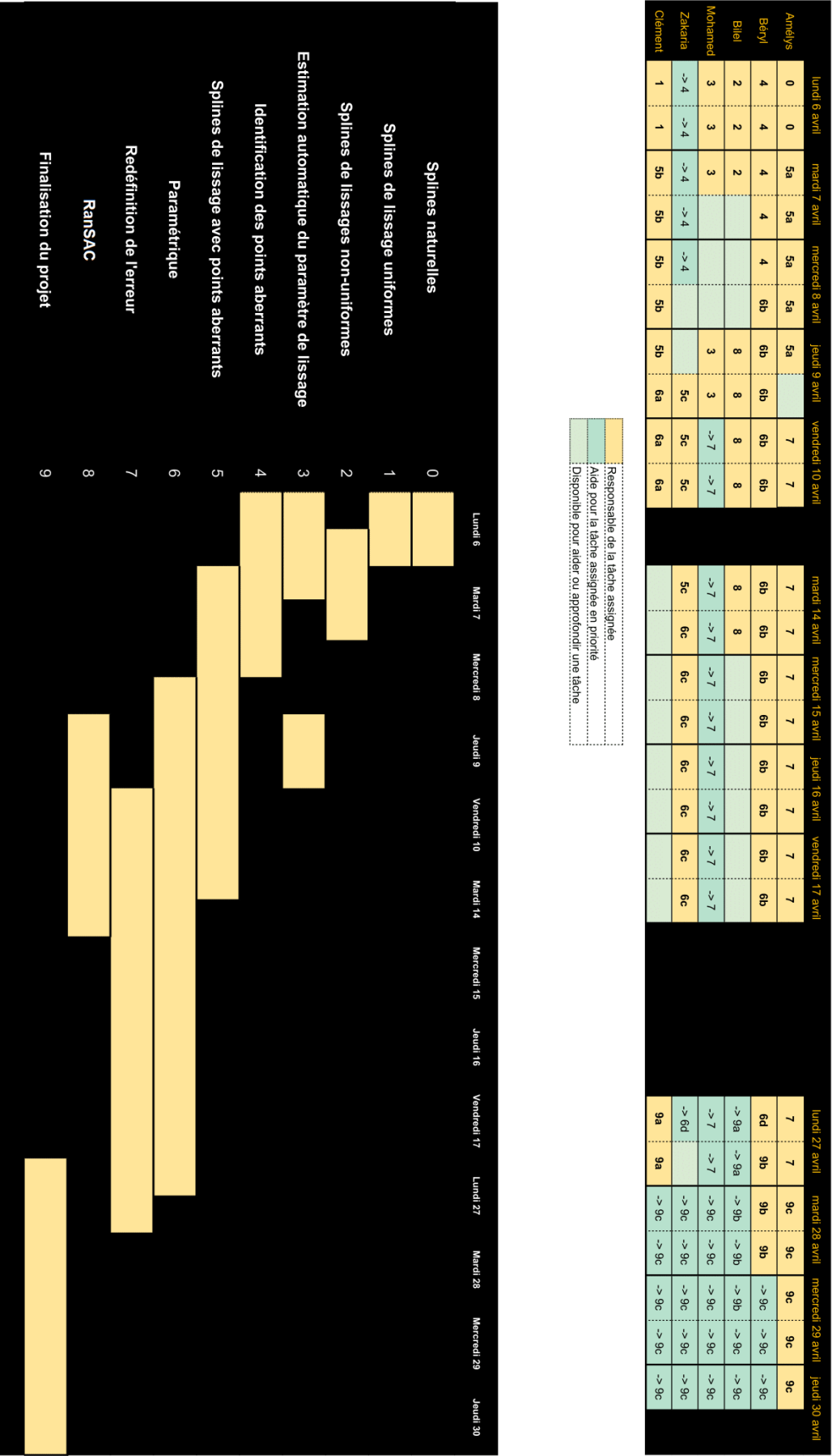


FIGURE 6 – Diagramme de Gantt simplifié et répartition des tâches