MASTER EN SCIENCES MATHÉMATIQUES À FINALITÉ DIDACTIQUE

FIFARank

Classer les équipes nationales avec la théorie des réseaux

Brachotte, Loïc

Award date:
2021

Awarding institution:
Universite de Namur

UNIVERSITE DE NAMUR

Faculté des Sciences

FIFARank : Classer les équipes nationales avec la théorie des réseaux

Mémoire présenté pour l'obtention

du grade académique de master en « sciences mathématiques à finalité didactique »

Loïc BRACHOTTE

Juin 2021

Remerciements

Je tiens ici à remercier les différentes personnes qui m'ont aidé, soutenu ou ne serait-ce qu'encouragé tout au long de ce mémoire.

Je voudrais dans un premier temps remercier mon promoteur de mémoire, Monsieur Timoteo Carletti, qui a toujours été disponible pour répondre à mes questions malgré les conditions sanitaires que nous connaissons tous et les difficultés de communication qu'elles peuvent engendrer. J'ai toujours pu compter sur lui pour me lancer sur de nouvelles pistes ou m'aider lors de mes incompréhensions qui furent nombreuses.

J'aimerais remercier l'ensemble du corps professoral qui m'a accompagné dans mon développement sur de nombreux points, aussi bien cognitifs que psychologiques et même sportifs car il se trouve en effet que certaines personnes se débrouillent aussi bien en astronomie qu'en judo ! Je voudrais faire une mention spéciale pour Madame Pascale Hermans qui s'est toujours montrée plus que disponible pour les étudiants.

Je voudrais remercier mes proches, que je ne citerai pas pour ne pas risquer d'oublier quelqu'un mais qui se reconnaîtront. Ceux-ci m'ont supporté pendant ces deux ans et avaient déjà commencé ce dur travail durant mes trois années de bachelier. Je les remercie sincèrement pour leur présence et l'amour qu'ils m'ont apporté au quotidien.

Je voudrais terminer en remerciant les amis que j'ai pu oublier pendant certaines périodes de travail, nombreux sont ceux qui sont restés compréhensifs en permanence et à m'avoir soutenu peu importe mon humeur.

Résumé

FIFARank : Classer les équipes nationales avec la théorie des réseaux. Ce mémoire présente au travers de différentes analyses plusieurs classements qui pourraient espérer rivaliser avec le classement FIFA. Ces classements sont construits sur la base de la théorie des réseaux et plus particulièrement sur le Pagerank et le facteur h. Divers outils d'analyses sont disponibles à l'intérieur de ce mémoire afin de pouvoir comparer des classements entre eux et en estimer la fiabilité. Plusieurs classements seront donc ainsi générés avant de ne conserver que le meilleur. Combiner l'univers sportif et l'univers mathématique est une chose qu'il est possible de faire au moyen des réseaux. Prédire la victoire de la France à la coupe du monde 2018 aussi.

Mots-clefs

Théorie des réseaux, Centralité, Football, FIFA, Classement, Pagerank, Prédiction.

Abstract

FIFARank : Ranking national teams with network theory. This memory presents through different analyses severall rankings that could improve the official one provided by the FIFA. This classification is based on the theory of networks and more particularly on the Pagerank and the h-factor. Various analytical tools are available within this memory in order to compare rankings between them and estimate their reliability. Several rankings will thus be generated before retaining only the best. Combining the sports and mathematical worlds is something that can be done through networks. Predict France's victory at the 2018 World Cup as well.

Table des matières

Remerciements . . . 2
Résumé . . . 3
Mots-clefs . . . 3
Abstract . . . 3

Introduction **6**

1 Concepts clefs **8**
1.1 Réseaux . . . 8
1.2 Mesures de centralité . . . 11
1.2.1 Centralité de degré . . . 11
1.2.2 Centralité spectrale . . . 12
1.2.3 Centralité de proximité . . . 13
1.3 Pagerank . . . 15
1.4 H-facteur . . . 16
1.5 L'indice de Jaccard . . . 21
1.6 Modèle proposé . . . 23
1.6.1 Classement FIFA . . . 23
1.6.2 Notre idée de modèle . . . 25

2 Présentation des données **26**
2.1 Différents fichiers de données . . . 26
2.2 Préparer les analyses de données . . . 27
2.3 Méthodologie utilisée pour ce travail . . . 30
2.3.1 Paramètres préliminaires importants . . . 30
2.3.2 Trois types de classement . . . 30
2.3.3 Analyses inter-classements . . . 31
2.3.4 La fiabilité du classement FIFA contre la fiabilité de notre classement 31
2.3.5 Une évolution temporelle du classement . . . 31

3 Algorithmes et Méthodes utilisés **32**
3.1 Fichiers et éléments d'analyses reçus . . . 32
3.2 Fichiers implémentés . . . 32
3.2.1 Code_general.m . . . 33
3.2.2 GenDonnee.m . . . 34
3.2.3 GenScore.m . . . 34
3.2.4 GenGraph.m . . . 34
3.2.5 FacteurH.m . . . 36
3.2.6 Jaccard.m . . . 36
3.2.7 RechercheTop.m . . . 37

3.2.8 Fichiers restants 37

4 Résultats **38**
4.1 Résultats inter-classements 38
4.1.1 Classement réalisé sur une matrice aléatoire 39
4.1.2 Classement sur un an 40
4.1.3 Classement sur deux ans 41
4.1.4 Classement sur trois ans 42
4.1.5 Classement sur quatre ans 43
4.1.6 Analyse générale 44
4.2 Fiabilité de prédiction 45
4.3 Évolution temporelle des classements 51
4.3.1 Le classement Pagerank 52
4.3.2 Le classement utilisant le facteur h 57
4.3.3 Le classement FIFA 61
4.3.4 Conclusion relative à l'évolution temporelle 65

5 Perspectives **66**

Conclusion **69**

Annexes **74**

Introduction

Vous n'êtes pas sans savoir que toutes les applications et les logiciels qui nous entourent sont régis par des modèles prédictifs. Le but premier de ces applications est simple, plaire et convenir à un maximum d'utilisateurs. Les exemples de ces applications sont nombreux. Pensons à Netflix qui organise ses différentes propositions en fonction des films et séries ayant plu aux utilisateurs. Youtube qui nous propose automatiquement des vidéos dont le style correspond aux vidéos que nous avons préalablement visionnées. Nous pourrions également, pour n'en citer qu'un dernier, parler des moteurs de recherche qui tentent de nous proposer les pages les plus intéressantes possibles lors d'une recherche de l'utilisateur. Toutes ces applications sont régies par des modèles permettant de prédire les attentes des utilisateurs où comme en ce qui nous concerne, les résultats potentiels des futurs compétitions de football en utilisant les informations et les résultats des matchs passés.

Le football ainsi que ses différentes compétitions, que ce soit, la coupe du monde qui concerne les équipes nationales, la ligue des champions de l'UEFA (l'Union des Associations Européennes de Football) et ses équivalents sur les autres continents, représentent un impact beaucoup plus important que le simple attrait du sport. Un véritable aspect économique gravite autour du football. Les revenus générés par la vente d'objets dérivés, les recettes supplémentaires engrangées par certains cafés avoisinant les stades de football, l'engouement provoqué chez les jeunes voulant supporter leurs idoles, le recyclage d'argent [4] et les enjeux politiques, comme par exemple Berlusconi qui s'est servi de la réussite du Milan AC dont il était le président dans un but politique, sont tout autant de raisons qui poussent les nations et les différents clubs à investir dans la formation de leurs joueurs et leurs clubs de football. Il est donc important de pouvoir prévoir certains résultats afin de faire la promotion de son équipe et tenter d'augmenter l'engouement envers celle-ci. Il existe en plus de cela une autre raison financière très importante pour la Fédération Internationale de Football Association appelée FIFA, connaître le potentiel de chaque équipe lui permet de redistribuer correctement une partie des recettes à certains clubs tout comme organiser de manière idéale les différentes poules des compétitions. Ces prévisions, la FIFA tente de les effectuer grâce à son classement qu'elle met à jour de manière mensuelle et qui est disponible sur le site officiel de la FIFA [5]. Le problème que nous pouvons relever est que ce classement n'apporte pas une réelle prédiction des compétitions à venir. Nous avons vu par le passé que les équipes les mieux placées dans le classement n'étaient pas forcément celles qui finissaient aux meilleures places de la coupe du monde et des autres coupes internationales. Cette différence peut s'expliquer par le fait que le classement FIFA est construit suivant des coefficients arbitraires qui peuvent donc biaiser ce dernier.

C'est pourquoi nous avons décidé de travailler sur une nouvelle version possible du classement FIFA en construisant un ranking à l'aide du réseau généré grâce aux résultats des matchs en évitant au maximum possible l'utilisation de paramètres arbitraires. La motivation étant donc de travailler sur les réseaux en particularisant le sujet à un domaine sportif très important dans le milieu économique et social.

Nous aurons pour but au sein de ce mémoire, de prédire du mieux que possible les résultats des différentes rencontres de football. Pour ce faire, nous allons donc nous attarder sur la création de différents réseaux qu'il nous sera possible d'analyser et de comparer entre eux mais également avec le classement FIFA.

Ce mémoire permettra, nous l'espérons, à de nombreuses personnes, de comprendre l'utilité des réseaux en général. Nous focaliserons cet intérêt sur le milieu sportif avec le football afin de particulariser cette utilité à un cas concret. Cette lecture peut être divisée en trois parties différentes.
Une première partie théorique qui sera développée dans le chapitre 1 et qui introduira les différents concepts nécessaires à la lecture de ce mémoire. Nous y développerons également les outils d'analyses qui nous ont aidés tout au long de ces deux ans de travail. Nous essayerons autant que possible de nous ramener au football lorsque la théorie le permettra et nous en apprendrons ainsi d'avantages sur des grands concepts comme celui de centralité ou de Pagerank. Nous y aborderons comme outils d'analyses des facteurs tels que le facteur h ou encore l'indice de Jaccard. Nous clôturerons cette partie par une partie explicative concernant le classement FIFA et le modèle actuellement utilisé et nous introduirons l'idée que nous avons a proposé pour améliorer ce modèle.
Une deuxième partie plutôt descriptive qui aura pour but d'expliquer comment nous avons décidé de travailler les différentes données. C'est dans cette partie que nous abandonnerons un peu le côté théorique pour nous rapprocher du côté pratique et des informations que nous pouvons construire sur base des matchs internationaux de football. Les données principales que nous avons utilisées tout au long de ce travail seront présentées et décortiquées. Cette partie nous permettra également de vous présenter les différents codes que nous avons implémenté et qui sont primordiaux dans ce mémoire. Ils ont été utilisés pour générer nos différents résultats et réaliser l'ensemble des simulations, ceux-ci sont tous disponibles sur `https://gitlab.unamur.be/math/mac/memoires/potentielles-ameliorations-du-ranking-fifa`.
Cette partie sera constituée des chapitres 2 et 3.
La troisième et dernière partie, avant les perspectives d'avenir et la conclusion, sera constituée du chapitre 4 qui reprendra l'ensemble des résultats et des analyses que nous aurons pu faire pour ce mémoire. Les analyses se sont penchées sur plusieurs points particuliers que nous détaillerons dans le chapitre 2 mais reprendrons généralement l'ensemble des informations que nous pourrions essayer de trouver en travaillant avec des classements, c'est à dire comparer les différents classements entre eux, évaluer leur fiabilité et enfin observer leur comportement et leur évolution dans le temps.

Chapitre 1

Concepts clefs

Ce chapitre servira d'introduction théorique à la théorie des réseaux ainsi qu'aux différents outils et modèles dont nous aurons besoin pour la bonne compréhension de ce mémoire. Nous essayerons d'être le plus précis possible afin que n'importe qui n'ayant jamais travaillé précisémment avec des réseaux puisse malgré tout comprendre l'intérêt que nous leur portons.

1.1 Réseaux

« Les réseaux sont partout, de l'Internet aux réseaux sociaux et jusqu'aux réseaux génétiques qui déterminent notre existence biologique. » [1]. Cette phrase d'Albert-László Barabási énonce clairement l'importance des réseaux. Du secteur médical au secteur sportif en passant par les réseaux routiers et les différents mouvements bancaires, les réseaux se cachent derrière de nombreuses activités de la vie de tous les jours. En ce qui concerne la biologie par exemple, Fernando Vega-Redondo nous écrit, en se basant sur l'article [10], « Une première (et vaste) sous-unité de recherche a été la biologie moléculaire, le but étant de comprendre divers processus moléculaires tels que les réactions métaboliques. » [17]. [1]
L'importance des réseaux est capitale pour comprendre le fonctionnement de certains phénomènes et construire des solutions à de nombreux problèmes. Un problème qui peut être résolu par l'utilisation de réseaux est le problème de correspondance lors de voyages en train, Mark Newman reprend par exemple le travail de Sen et al. [16] et dit « Les gens ne se soucient pas tellement du nombre d'arrêts qui y sont en cours de route, tant qu'ils n'ont pas à changer de train. » [15]. Désengorger les réseaux routiers en fournissant un nouvel itinéraire lorsqu'une voie de circulation est bouchée, permettre la compréhension du fonctionnement du cerveau ou bien améliorer la qualité de nos moteurs de recherche font partie des nombreuses applications dans lesquelles les réseaux sont utiles.

Dans le cadre de ce mémoire, les réseaux représentent un outil indispensable pour modéliser notre problème. Nous allons de ce pas présenter les différents éléments qui constituent un réseau et identifier à quoi ils correspondent dans notre cas pour faciliter la compréhension de la suite. Les définitions que vous verrez dans cette partie sont pour la plupart tirées du syllabus de Théorie des graphes [12]. Commençons par définir la notion de graphe qui est la notion fondamentale de ce sujet.

1. Lorsque l'on considère l'ensemble de réactions métaboliques, on parle de réseau métabolique. Cette structure de réseau permet de modéliser et de mieux comprendre le métabolisme.

Définition 1 *Un graphe est un triplet (A,E,Φ) tel que :*

- *A est un ensemble dont les éléments sont appelés sommets ou noeuds,*
- *E est un sous-ensemble de A×A dont les éléments sont appelés arêtes,*
- *Φ est une fonction, dite fonction d'incidence, qui associe à chaque arête un sommet ou une paire de sommets.*

Définition 2 *Un graphe dirigé (A,E,Φ) est un graphe dans lequel la fonction d'incidence Φ associe à chaque arête une paire ordonnée de sommets.*

Chaque réseau peut-être représenté schématiquement par un graphe constitué des trois éléments précédents. Le jeu de données sur lequel nous travaillons vous sera présenté plus tard mais est constitué de l'intégralité des résultats de matchs internationaux de football entre le 30 novembre 1872 et le 1^{er} février 2020. Afin de faciliter la compréhension des différents graphes que nous utiliserons dans la suite, nous allons maintenant donner quelques définitions de base. Ainsi, chaque équipe sera représentée par un noeud et une arête dirigée existera du sommet a_i au sommet a_j si l'équipe a_i a gagné contre le sommet a_j. En cas de match nul, nous accorderons dans un premier temps la victoire à l'équipe ne jouant pas à domicile, nous verrons plus tard comment les traiter de manière plus intéressante. Dans la suite, on considérera aussi le cas des réseaux pondérés, c'est à dire que chaque arête a un poids positif ; nous pourrons ainsi dire que l'arête de a_i à a_j a un poids de 2 si l'équipe a_i a gagné deux fois contre l'équipe a_j. On pourrait aussi mettre un poids proportionnel au nombre de buts marqués.

Vous pouvez voir sur la Figure 1.1 un exemple de graphe avec lequel on peut communément travailler. Prenons un exemple restreint d'équipes ainsi que des résultats factices de matchs les concernant. Nous représenterons ensuite le réseau que nous aurons construit. Prenons par exemple 5 équipes européennes, la Belgique, la France, l'Italie, l'Espagne et l'Allemagne, que nous représenterons sur le graphe par des sommets nommés respectivement a_1, a_2, a_3, a_4 et a_5. Imaginons ensuite les résultats factices suivants, la Belgique gagne contre toutes les équipes sauf l'Espagne, la France perd contre toutes les équipes sauf l'Allemagne, l'Italie ne dispute que deux matchs, gagne contre la France et perd contre la Belgique, l'Espagne gagne contre la Belgique et la France mais perd contre l'Allemagne et l'Allemagne perd contre la France et la Belgique mais gagne contre l'Espagne. Nous allons maintenant représenter l'entièreté de ces informations à la Figure 1.1. Appelons le graphe ainsi obtenu G_1, il est décrit par le triplet $(A_{G_1}, E_{G_1}, \Phi_{G_1})$, où l'ensemble A_{G_1} est l'ensemble des sommets suivants a_1, a_2, a_3, a_4, a_5. L'ensemble des arêtes E_{G_1} est quant à lui constitué des arêtes $e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8$ et la fonction d'incidence est définie de la façon suivante.

$$\begin{aligned}
\Phi_{G_1}(e_1) &= a_1a_2,\\
\Phi_{G_1}(e_2) &= a_1a_3,\\
\Phi_{G_1}(e_3) &= a_1a_5,\\
\Phi_{G_1}(e_4) &= a_4a_2,\\
\Phi_{G_1}(e_5) &= a_2a_5,\\
\Phi_{G_1}(e_6) &= a_4a_1,\\
\Phi_{G_1}(e_7) &= a_5a_4,\\
\Phi_{G_1}(e_8) &= a_3a_2.
\end{aligned}$$

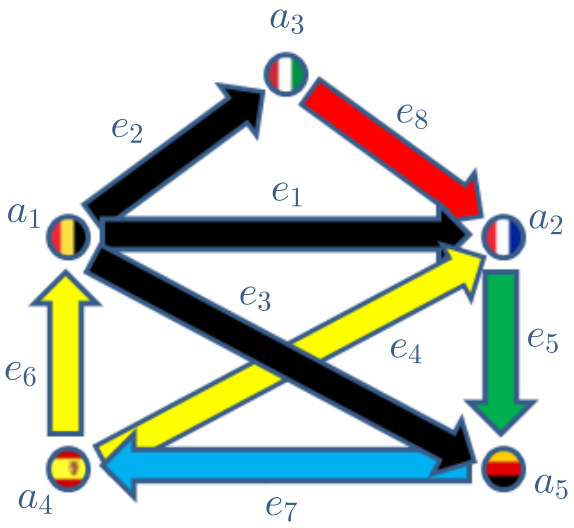

FIGURE 1.1 – Exemple de graphe

On voit clairement grâce aux explications précédentes qui a gagné contre qui, pour rappel, les arêtes dirigées partent de l'équipe victorieuse en direction de l'équipe perdante et dans notre exemple, elles ont un poids unitaire. Le graphe représenté à la Figure 1.1 ne nous apprend malheureusement rien d'autre en lui-même et ne nous permet pas d'avoir une idée d'un quelconque classement. Quelle serait l'équipe la plus forte dans ce tournoi ? Peut-être la Belgique car elle se détache des autres en ayant gagné trois matchs, mais comment classer par exemple l'Allemagne, la France et l'Italie, chacune victorieuse dans un seul match ? Observons aussi, que dans ce cas le réseau est assez petit (5 noeuds et 8 arêtes) et donc il peut être analysé à la main, mais comment faire pour des réseaux avec des centaines de noeuds et des milliers d'arêtes ? L'objectif de ce mémoire est d'utiliser la théorie des réseaux pour extraire un classement automatique une fois le réseau construit.

Nous allons avoir besoin pour en savoir plus de définir plusieurs notions. Ces différentes définitions nous permettront de poursuivre notre analyse des réseaux et d'introduire une notion qui nous sera fort utile dans la section suivante, la matrice d'adjacence du graphe (définition 8).

Définition 3 *Le nombre de sommets d'un graphe G est noté $\gamma(G)$, et celui d'arêtes $\epsilon(G)$*

Définition 4 *Un graphe complet est un graphe dans lequel chaque paire de sommets est reliée par au moins une arête.*

Définition 5 *Un parcours peut être vu de deux manières différentes avec la même idée sous-jacente, nous pouvons le considérer du point de vue des arêtes où un parcours serait une succession d'arêtes telle que deux arêtes consécutives possèdent un sommet en commun. La version équivalente en termes de sommets serait une suite de sommets dans laquelle deux sommets consécutifs sont connectés par une arête.*

Définition 6 *Un graphe est connexe si pour chaque paire de points il existe un parcours qui les relie. Les composantes connexes d'un graphe sont des sous-graphes maximaux.*

Définition 7 *Le degré d'un sommet est le nombre d'arêtes incidentes à celui-ci.*

Dans le cas d'un graphe dirigé, nous différencierons le degré entrant et le degré sortant. Plus le degré sortant d'un sommet sera élevé, plus l'équipe correspondante aura réalisé de victoires. Une approche pratique permettant une manipulation plus facile des graphes est une approche matricielle. La notion principale permettant cette approche matricielle est la matrice d'adjacence.

Définition 8 *La matrice d'adjacence de G est la matrice carrée $\gamma \times \gamma$ dont l'élément ij sera égal à 1 si il y a un lien entre le sommet a_i et le sommet a_j et 0 sinon. Nous la dénommerons communément M et utiliserons la notation m_{ij} pour désigner l'élément ij de la matrice.*

Dans le cadre de notre travail sur les différentes équipes de football, nous allons travailler avec une matrice d'adjacence qui ne contiendra pas uniquement que des 0 et des 1. Dans le cas d'un multigraphe, les éléments m_{ij} seront des entiers et représenteront exactement le nombre de liens entre a_i et a_j. Ce qui nous intéressera le plus dans notre cas, seront les réseaux pondérés pour lesquels les éléments m_{ij} seront des réels et correspondront au poids de l'arête.

Dans le cas de la matrice d'adjacence de la Figure 1.1, l'arête e_1 sera reprise dans l'élément m_{12} mais pas dans l'élément m_{21} étant donné qu'elle relie a_1 à a_2 et que le graphe est un graphe dirigé. Dans ce mémoire, l'ensemble de nos graphes le seront, nous comptabiliserons donc uniquement les arêtes sortantes dans notre matrice d'adjacence. Ainsi pour faire le lien avec notre jeu de données, si l'équipe a_i gagne un match contre l'équipe a_j, il existera une arête e_i sortant de a_i et se dirigeant vers a_j, l'élément m_{ij} sera égal à 1 et l'élément m_{ji} demeurera inchangé. Comme nous travaillerons avec des réseaux pondérés, nous incrémenterons nos arêtes avec un certain poids à chaque victoire supplémentaire.

1.2 Mesures de centralité

La centralité est une mesure qui permet de préciser l'importance de chaque noeud, l'importance n'est pas toujours accordée à un même critère pour des graphes différents, c'est pourquoi plusieurs mesures de centralité existent. Il est utile lorsque l'on analyse un graphe, de savoir quel sommet a le plus d'importance et lequel en a le moins. Cette recherche d'importance, nous pouvons le faire via les différentes mesures de centralité. En fonction du paramètre que nous voulons considérer et du critère auquel nous accordons de l'importance, il nous sera possible d'utiliser différentes mesures de centralité. Nous allons donc vous expliquer en quoi elles consistent et comment nous pouvons les calculer. Etant donné que chaque noeud possède une valeur de centralité, nous pouvons considérer cette centralité comme un vecteur c de taille γ, nous noterons la centralité du sommet a_i, c_i.

1.2.1 Centralité de degré

En considérant la matrice d'adjacence M symétrique d'un graphe G, la mesure de centralité la plus évidente que l'on puisse considérer est de trier les noeuds en fonction de leur degré. La centralité de chaque sommet est donc calculée de la manière suivante.

$$c_i = \sum_{j=1}^{\gamma} m_{ij},\ 1 \leq i \leq \gamma. \tag{1.1}$$

Il est également possible de calculer le vecteur de centralité en réalisant la sommation sur les lignes dans le cas d'un graphe non dirigé, le degré d'un noeud pouvant être calculé indépendamment en considérant la ligne ou la colonne de la matrice d'adjacence. Dans le cas d'un graphe dirigé, l'équation 1.1 correspondra au degré entrant d'un noeud tandis que l'expression similaire avec une sommation sur les colonnes permettra de calculer le degré sortant, ainsi le degré sortant du noeud a_i donnera le nombre de matchs que a_i a gagné et le degré entrant, le nombre de matchs perdus par a_i.

Une version normalisée de la définition est présentée dans les sources [3] et [8], celle-ci consiste, comme toute autre procédure de normalisation, à reprendre la définition de la centralité de degré exprimée en 1.1 et de diviser chaque élément du vecteur de centralité par $\gamma - 1$, qui correspond au degré maximal dans un réseau composé de γ noeuds. De la sorte, un sommet qui sera relié à tous les autres sommets du graphe aura une centralité de 1 et les autres valeurs de centralité lui seront inférieures ou égales.

Cette mesure de centralité, au détriment d'être simple, est intéressante pour autant que les arêtes aient la même importance dans le graphe ou que le caractère que nous jugeons important soit la quantité d'informations qui peut circuler par un élément, comme par exemple les plates-formes de correspondance aéroportuaire aussi appelé hub.

1.2.2 Centralité spectrale

Etant donné que notre problème sera de traiter les différents résultats des matchs de football internationaux, il est logique de ne pas pondérer de la même manière une victoire contre une équipe forte et une victoire contre une équipe faible, cette notion de pondération est connue sous le nom de centralité de vecteur propre où centralité spectrale.

Pour fixer les idées, un sommet sera important s'il est relié à plusieurs sommets qui le sont également. Gagner un match contre des équipes qui gagnent face à tous leurs opposants est largement plus valorisant que de ne gagner que contre des équipes qui accumulent des défaites. En terme mathématique, il s'agit de la continuité directe de la centralité de degré, nous allons pouvoir calculer la centralité d'un noeud en effectuant une combinaison linéaire sur la centralité de ses voisins.

$$c_i = \frac{1}{\alpha} \sum_{j=1}^{\gamma} m_{ji} c_j, \ 1 \leq i \leq \gamma, \ \alpha \geq 0. \tag{1.2}$$

Dans le cadre d'un graphe dirigé, il existera tout comme pour la centralité de degré, une centralité qui concernera les arêtes entrantes (équation 1.3) et une autre qui concernera les arêtes sortantes (équation 1.4).

$$c_i^{in} = \frac{1}{\alpha} \sum_{j=1}^{\gamma} m_{ji} c_j^{in}, \ 1 \leq i \leq \gamma, \ \alpha \geq 0, \tag{1.3}$$

$$c_i^{out} = \frac{1}{\beta} \sum_{j=1}^{\gamma} m_{ij} c_j^{out}, \ 1 \leq i \leq \gamma, \ \beta \geq 0. \tag{1.4}$$

Les équations 1.2 et 1.3 correspondent donc à un système d'équation faisant intervenir la matrice d'adjacence et le vecteur de centralité que nous pouvons écrire des deux façons suivantes pour la centralité entrante et la centralité sortante.

$$\left(c^{in}\right)^T = \frac{1}{\alpha} M^T \left(c^{in}\right)^T,\ \alpha \geq 0, \tag{1.5}$$

$$c^{out} = \frac{1}{\alpha} M c^{out},\ \alpha \geq 0, \tag{1.6}$$

où M sera la matrice d'adjacence du graphe dans le cas où nous voudrions calculer la centralité entrante, et sa transposée si nous voulons calculer la centralité sortante. Il est évident que pour un graphe non dirigé, le fait de transposer la matrice d'adjacence n'influence pas l'équation étant donné qu'elle est symétrique.

Par définition, nous pouvons apercevoir dans l'équation 1.4 que c est un vecteur propre de M avec α sa valeur propre. On peut montrer que le vecteur de centralité correspond au vecteur propre dominant de la matrice d'adjacence ; c'est-à-dire pour rappel, celui qui possède la plus grande valeur propre. Pour calculer ce vecteur de centralité, il est possible d'utiliser la méthode de la puissance qui est particulièrement pratique lorsque notre matrice d'adjacence est creuse.

1.2.3 Centralité de proximité

La dernière centralité que nous aimerions présenter dans cette section est la centralité de proximité, qui selon nous nous permet déjà d'avoir une idée plus visuelle et plus commune d'une ébauche de classement dans le cas de notre problème. Cette centralité utilise la notion de distance entre deux sommets. Cette notion de distance est variable, dans certains cas de réseaux, il sera utile de donner une taille proportionnelle à une distance aux différentes arêtes comme pour une carte routière où chaque route possède une taille prédéfinie contrairement au cas des réseaux sociaux par exemple où les arêtes symboliseraient uniquement le fait que deux personnes soient en contact ou non, dans ce cas-là, la notion de distance serait simplifée dans le sens où la distance entre deux noeuds connectés serait de 1 et la distance entre deux noeuds non connectés serait infinie. Cette centralité possède un aspect logique qui est qu'un sommet est important s'il est proche d'un grand nombre de sommets du graphe, un sommet isolé serait donc très peu important. La centralité d'un sommet est calculée en prenant sa proximité moyenne par rapport à l'ensemble des autres sommets du graphe.

$$c_i = \frac{\gamma - 1}{\sum_{j=1}^{\gamma} dist(a_i, a_j)},\ 1 \leq i \leq \gamma. \tag{1.7}$$

Dans le cadre d'un graphe dirigé, cette fonction de distance ne sera pas forcement symétrique, la distance entre a_i et a_j et celle entre a_j et a_i ne seront pas obligatoirement égales. Il nous faudra donc considérer deux centralités de proximité différentes tout comme pour les deux centralités précédentes.

Dans le cadre de notre problème de classement des équipes de football, le graphe étant dirigé et les arêtes symbolisant le résultat du match, il est plus intéressant de considérer la centralité de proximité avec les arêtes sortantes, qui pour rappel symbolisent les victoires. La distance entre deux noeuds adjacents (au sens du graphe dirigé) aura une valeur de 1 si

les équipes se sont affrontées une seule fois. Cette distance sera en réalité égale au nombre de d'arêtes sortantes existant entre les deux équipes mais il est plus simple pour l'expliquer de ne pas tenir compte des multiplicités. Ainsi les équipes qui auraient gagné un match contre chacune des autres équipes auraient une centralité de 1. Les équipes qui n'auraient gagné aucun match auraient une valeur de centralité infinie.

Ce qui est selon nous très visuel dans cette centralité en ce qui concerne les résultats de matchs de foot est que la centralité de chaque sommet nous donne une information directe sur le nombre d'équipes contre laquelle elle a été victorieuse. Plus la valeur sera proche de 1 et plus nous aurons l'information que l'équipe concernée a gagné contre de nombreuses équipes, or les résultats des matchs ne sont généralement pas liés au hasard. A moins que deux équipes possèdent un niveau de jeu équivalent, il est fréquent d'observer à de nombreuses reprises une victoire de la même équipe. Cette manière de penser n'est cependant pas fiable, s'il était possible de déterminer à l'avance le vainqueur d'un match de manière sûre, le classement resterait en permanence inchangé.

Afin de terminer cette section, nous allons maintenant analyser la centralité de chaque sommet de la Figure 1.1 pour chacune des méthodes expliquées ci-dessus. Les centralités qui nous intéressent sont bien entendu la version qui correspond aux arêtes sortantes. Voici, à la Figure 1.2 un tableau des différentes valeurs de centralité. Nous pouvons constater que peu importe la méthode de calcul utilisée pour obtenir le vecteur de centralité, les sommets a_1 et a_4 ont en permanence une valeur de centralité plus élevée que les 3 autres sommets qui s'explique par leur nombre de victoires plus élevé.

Nous pouvons observer que l'équipe a_1 est toujours la première car elle a gagné trois matchs, l'équipe a_4 suit avec ses deux matchs gagnés. Observons que la centralité de degré n'est pas capable de départager les équipes a_2,a_3 et a_5 (chacune avec une seule victoire), par contre cela est possible avec les deux autres centralités.
Le détail important à relever est la valeur de centralité plus élevée du sommet a_5 pour la centralité de vecteurs propres et la centralité de proximité alors qu'il comptabilise le même nombre de victoires que les sommets a_2 et a_3. Cette valorisation est dûe à "l'importance" de sa victoire, le fait de gagner contre l'équipe a_4, qui est une équipe plus importante, lui accorde une plus grande importance que les deux autres.

Récapitulatif des différentes centralités "sortantes" des sommets de la Figure 1.1.

Type de centralité et sommet concerné	Centralité de degré	Centralité de vecteurs propres	Centralité de proximité
Sommet a_1	0,75	0,60	0,8
Sommet a_2	0,25	0,28	0,4
Sommet a_3	0,25	0,19	0,4
Sommet a_4	0,5	0,60	0,67
Sommet a_5	0,25	0,4	0,5

FIGURE 1.2 – Tableau comparatif des résultats

1.3 Pagerank

Après vous avoir présenté trois types de centralité différents, il nous semble utile de vous présenter une autre valeur de centralité initialement conçue pour les pages Internet, le Pagerank. Nous avons séparé cette centralité des autres car nous l'utilisons particulièrement au sein de nos codes. Les centralités présentées précédemment seront pour certaines utilisées également dans nos implémentations mais sont des centralités « connues » qu'il nous semblait intéressant de présenter peu importe leur utilité dans le cadre de ce mémoire.

Le Pagerank est un algorithme qui permet de déterminer l'intêret d'un site web. Dans une implémentation initiale, plus le Pagerank d'un site était élevé, plus la page avait de chances d'être proposée par un moteur de recherche lors d'une recherche concernant le sujet demandé par l'internaute. A l'heure actuelle le Pagerank bien que toujours important, tend à devenir moins intéressant car comme le souligne l'équipe de Search Engine Optimization (SEO) [14], il n'est qu'un des différents critères utilisés pour le référencement d'un site. De nombreuses méthodes existent pour augmenter le Pagerank de son propre site, c'est pourquoi lui accorder une importance fondamentale est risquée. Avant d'aller plus loin dans son explication, regardons pourquoi et comment il a été créé.

Inventé vers la fin des années 1900 par Larry Page et Sergey Brin, fondateurs de la société Google, le Pagerank classe les différentes pages Internet en fonction d'un critère assez simple, le nombre et la qualité des autres pages Internet qui redirigent les internautes vers la page concernée. Il est en quelque sorte assez proche de la centralité spectrale dans le principe mais diffère au niveau de son implémentation. Il prend en compte le Pagerank des pages qui pointent vers la page considérée mais pondère chaque contribution en divisant le Pagerank d'une page par le nombre de liens qu'elle génère. Cette idée reste très logique, être la seule page à être ciblée par une page importante permettra d'augmenter son Pagerank de manière considérable, ce qui ne serait pas le cas en étant l'une des 1000 pages ciblées par une autre importante. Une expression basique du calcul du Pagerank peut s'exprimer comme ceci,

$$c_i = \sum_{j=1}^{\gamma} \frac{m_{ji}}{k_j^{out}} c_j, \ 1 \leq i \leq \gamma, \tag{1.8}$$

avec k_j^{out} qui représente le degré sortant du sommet j et est donc calculé grâce à la matrice d'adjacence en sommant les éléments m_{ji} avec i qui parcourt les différentes colonnes de la matrice. Il est évident que dans le cas d'Internet, le γ de l'expression 1.8 sera démesurément grand étant donné la quantité de pages Internet existantes, la matrice d'adjacence sera cependant creuse car il est fort probable que de nombreuses pages ne se pointent pas les unes les autres et donc la somme sera restreinte à un nombre relativement petit de pages.

Une autre façon de comprendre le Pagerank est de comptabiliser le nombre de fois qu'un internaute arriverait sur chacune des pages web en se déplaçant de manière aléatoire vers un des liens existants sur la page sur laquelle il se trouve. Afin d'éviter de se retrouver enfermé dans un cycle ou dans un sous-groupe du Web dans lequel les pages seraient connectées uniquement entre elles, une faible probabilité de se déplacer ailleurs que sur les liens présents est également ajouté au processus du calcul du Pagerank. Si l'on considère que l'internaute se déplace comme expliqué précédemment sans tenir compte des pages qu'il a déjà visitées, nous nous retrouvons alors dans un contexte spécifique connu sous le nom

de processus de Markov. Une introduction aux chaînes de Markov est réalisée d'une manière plus qu'abordable dans le Syllabus de l'Université de Paris-Saclay qui est disponible en format PDF [13].

Définition 9 *Une chaîne de Markov est l'appelation donnée à une suite $(X_n)_n$ lorsque l'élément X_n dépend uniquement de l'élément X_{n-1} pour tout n appartenant à l'ensemble des naturels.*

Le Web étant un graphe dirigé d'une taille extraordinairement grande avec en plus la caractéristique d'être en perpétuelle évolution, le calcul du Pagerank, qui revient au calcul d'un vecteur propre de la matrice d'adjacence du web, est impossible directement, c'est pourquoi de nombreux algorithmes d'approximation sont utilisés.

L'idée générale du Pagerank étant assez simple à comprendre, de nombreux propriétaires de pages Internet réussissaient à augmenter le Pagerank de leurs pages en ajoutant des liens pointant vers leurs pages sur des pages subsidiaires ou des forums. Suite à cette manipulation anormale du Pagerank, Google a modifié son algorithme, il n'est maintenant plus possible de connaître facilement le Pagerank de sa propre page et la formule exacte du calcul est maintenant conservée par Google.

Nous aimerions cependant pour terminer cette section parler brièvement d'une modification du caractère aléatoire établi en 2010 par Google, en effet, leurs techniciens ont modifié la "marche aléatoire" qui sert à calculer le Pagerank de chaque page. La probabilité de se déplacer sur chaque lien présent sur une page n'étant plus identique, le marcheur conserve une part d'aléatoire dans son déplacement mais la position, la pertinence ainsi que la taille des différents liens sont pris en compte pour calculer le poids de chacun et ainsi la probabilité de s'y déplacer.

Notre sujet se rapportant au football, il est peut-être étrange d'avoir présenté le phénomène du Pagerank, cependant certaines similarités existent entre le graphe du Web et le graphe général de nos matchs de football. Il sera donc intéressant de considérer, dans les sections suivantes, l'utilisation du Pagerank afin de juger de son utilité dans le cas de notre problème. Nous l'utiliserons en particulier dans l'un des trois classements que nous construirons, nous calculerons le vecteur de centralité en employant la méthode du Pagerank de Matlab.

1.4 H-facteur

Tout comme pour le Pagerank, nous allons maintenant vous présenter un coefficient qui a été initialement introduit pour évaluer l'impact scientifique des différents auteurs d'articles mais qui peut tout aussi bien nous être utile dans notre réflexion sur le classement des différentes équipes.

Définition 10 *L'indice h, suggéré par Jorge Hirsch en 2005, est un indice permettant de donner une valeur d'importance à un scientifique en fonction de ses articles. Il dépend du nombre de publications de l'auteur en question mais également du nombre de citations de ses publications les plus citées.*

Un nombre similaire au facteur h est le nombre d'Eddington qui est utilisé par les cyclistes. Le nombre d'Eddington prendra la valeur d si une distance d a été réalisée au moins d jours.

Les différents cyclistes pourront ainsi comparer leur performance et il est impossible, tout comme pour l'utilisation que nous ferons du facteur h, d'augmenter son nombre d'Eddington autrement quand parcourant des kilomètres à vélo. Contrairement au facteur h qui n'est pas forcément représentatif du travail accompli par un scientifique ; à cause des « auto-citations » par exemple ; le nombre d'Eddington est un bon élément indiquant les prestations physiques fournies par un cycliste.

L'indice h est calculé de la manière suivante, le facteur h d'un scientifique est égal au plus grand nombre [2] n de ses articles qui possèdent au moins n citations. Illustrons cette explication sur un exemple très simple afin de fixer la compréhension.

Dans l'exemple 1.3, le point critique est relevé en rouge dans le tableau. Nous ne pouvons plus prendre en compte le huitième article [3] de l'auteur car il ne possède que 5 citations. Ce point décisif est également visible sur la Figure 1.4, il se situe à l'intersection de la courbe représentant le nombre de citations par article et la première bissectrice. Le facteur h de ce scientifique serait donc de 7, si l'auteur arrive cependant à recevoir trois citations supplémentaires sur son huitième article, son facteur h augmentera de 1 et passera donc à 8.

Le Centre universitaire de la santé McGill propose dans l'un de ses articles des logiciels pour calculer son facteur lorsque l'on fait partie du cercle de leur Université et explique très brièvement le fonctionnement du facteur en donnant un exemple ; dont nous nous sommes inspirés pour créer le notre ; qui nous a permis de facilement comprendre son utilisation [2].

Hirsch, qui pour rappel est le créateur du facteur, aurait apparemment proposé un système de gradations dans lequel le facteur h déterminait différents postes pouvant être

Exemple d'utilisation du facteur h

Numéro de l'article	Nombre de citations de l'article
1	70
2	66
3	59
4	42
5	27
6	15
7	7
8	5
9	2
10	1

FIGURE 1.3 – Tableau recensant des données factices représentant le nombre de citations de chaque article.

2. Cette précision est apportée afin d'appuyer le fait évident que nous désirons obtenir un facteur h le plus élevé possible.

3. Les articles sont pris en compte par ordre décroissant de citations reçues et non par ordre de publication. Le huitième article pourrait tout aussi bien être la première publication de l'auteur ou bien sa dernière.

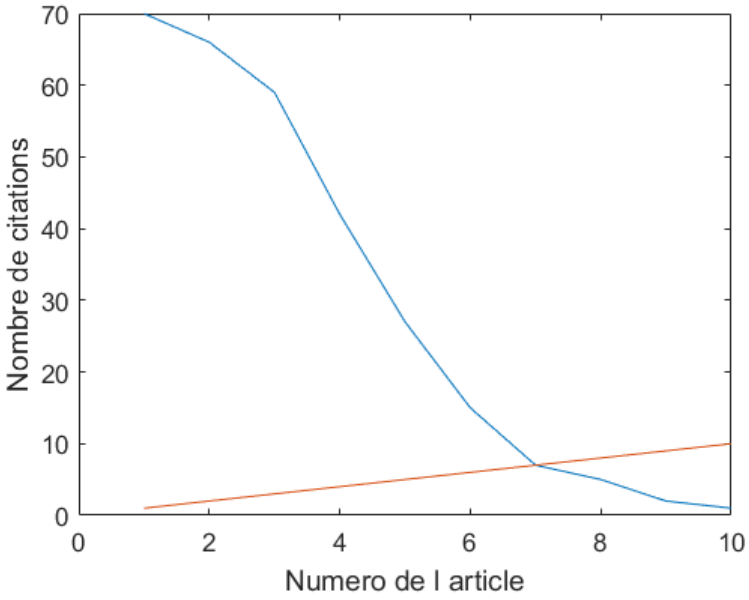

FIGURE 1.4 – Graphe représentant l'évolution du nombre de citations par article.

pourvus par les scientifiques, nous n'avons cependant trouvé qu'une seule source parlant de cette information [18] et nous n'avons donc pas pu la vérifier mais nous la trouvons personnellement plausible car cette proposition permettrait de ne fournir les différents postes de travail ou de recherches qu'à des scientifiques ayant un nombre suffisant de publications fortement citées. Cette technique est très intéressante pour quantifier l'intérêt des articles d'un scientifique mais peut également posséder de nombreux inconvénients. Une modification du h facteur a notamment été nécessaire afin de limiter « l'auto-citation » qui permettait à un auteur d'augmenter son facteur sans pour autant que ses articles ne soient cités par d'autres personnes que lui-même. Un autre problème que nous pouvons relever avec ce facteur est qu'il prend uniquement en compte le nombre de liens qui cite les articles du scientifique, une citation négative permet donc également d'augmenter le facteur. Il est également important de savoir à qui le facteur h accorde une citation lorsque l'article est co-signé, le facteur accorderait une citation à chaque auteur de l'article indépendamment de sa position et de son investissement. Etant donné ces défauts assez importants et d'autres que nous n'avons pas cités, nombreux s'accordent pour dire que le facteur h initial n'est pas plus précis qu'une moyenne du nombre de citations de chaque article d'un scientifique.

Notre intérêt pour cette mesure n'est cependant pas négligeable car nous pouvons l'appliquer d'une certaine manière à nos résultats de football. Les inconvénients n'auront plus lieu d'être et nous ne conserverons que le positif du concept initial. Nous calculerons le facteur h de chaque équipe en suivant la procédure suivante, premièrement nous allons sélectionner l'équipe dont nous voulons calculer le facteur h, appelons la l'équipe A. Ensuite, nous regarderons l'ensemble des équipes contre lesquelles l'équipe A a remporté au moins un match, nous appellerons cet ensemble E. Imaginons pour l'exemple que l'ensemble E soit constitué de n équipes, nous comptabiliserons pour chacune de ses n équipes leur nombre de victoires respectif au total [4].

4. Les victoires contre toutes les équipes, y compris celles en dehors de l'ensemble E.

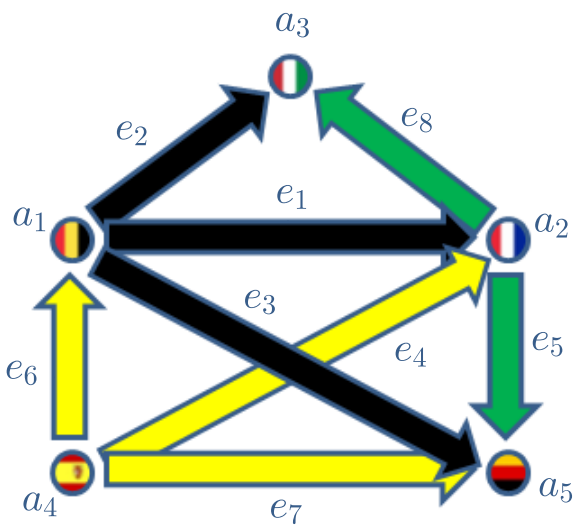

FIGURE 1.5 – Graphe de la Figure 1 modifié

Une fois tous ces calculs effectués, le facteur h de l'équipe A sera égal à h s'il y a au moins h équipes de l'ensemble E qui auront réalisé au moins h victoires, autrement dit l'équipe A aura gagné h matchs contre des équipes ayant gagné au moins h matchs, au plus h sera grand au plus on pourra considérer l'équipe A comme étant forte.

Pour illustrer cette procédure, nous allons détailler le calcul du facteur h d'une des équipes du graphe de la Figure 1.1 et donner le facteur de toutes les équipes. Détaillons le calcul du facteur h de la Belgique représenté par le noeud a_1.

1. La Belgique a gagné contre l'Italie, la France et l'Allemagne, l'ensemble E sera donc constitué de ces 3 équipes.
2. Dans l'ensemble E,
   - L'Italie a gagné au total 1 match.
   - La France a gagné au total 1 match.
   - L'Allemagne a gagné au total 1 match.
3. La Belgique a beau avoir gagné contre 3 équipes différentes, son facteur h vaut 1 car il n'y a pas 2 de ces équipes qui ont gagné au moins 2 matchs pour augmenter son facteur.

Dans ce graphique, chaque équipe possède un facteur équivalent qui vaut 1 du à la grande simplicité du graphe. Afin de rendre l'exemple intéressant, modifions légèrement le graphe en inversant le sens de la flèche reliant l'Allemagne à l'Espagne et le sens de la flèche reliant l'Italie à la France.

Sur la Figure 1.5, nous pouvons voir que l'Espagne et la Belgique ont toutes les deux gagné 3 matchs, la France 2 matchs tandis que l'Italie et l'Allemagne ont tout perdu. Calculons le facteur h des différents pays dans cet exemple afin d'apercevoir les modifications. On voit rapidement que ce facteur est nul pour l'Italie et l'Allemagne car elles ne possèdent

aucune victoire, le facteur de la France est également nul bien qu'elle ait gagné deux matchs car c'est deux opposants eux, n'en ont gagné aucun. Nous allons comparer le facteur de la Belgique et de l'Espagne, qui au vu de leur nombre de victoires identique pourraient être considérées comme égales en terme d'importance. Commençons par l'équipe Belge.

1. L'ensemble E relié à la Belgique n'a pas été modifié et est donc constitué de l'Italie, la France et l'Allemagne.
2. Dans l'ensemble E,
   - L'Italie n'a pas gagné match.
   - La France a gagné au total 2 matchs.
   - L'Allemagne n'a pas gagné de match.
3. La Belgique a beau avoir gagné contre 3 équipes différentes, son facteur h vaut 1 car une seule des équipes contre lesquelles elle s'est imposée à réussi à gagner au moins un match.

Les modifications apportées entre la Figure 1.1 et la Figure 1.5 ne changent rien dans le calcul du facteur h de la Belgique. En ce qui concerne l'Espagne par contre, nos modifications permettent un changement dans le calcul du facteur h.

1. L'ensemble E relié à l'Espagne est constitué de la Belgique, la France et l'Allemagne.
2. Dans l'ensemble E,
   - La Belgique a gagné 3 matchs.
   - La France a gagné au total 2 matchs.
   - L'Allemagne n'a pas gagné de match.
3. Le fait de gagner contre la Belgique qui a elle aussi gagné 3 matchs permet à l'Espagne d'avoir un facteur h de valeur 2[5].

On remarque donc une modification importante lorsque l'on renforce l'Espagne mais cela est également du au fait d'accorder une victoire supplémentaire à la France. Dans le cas où la France, l'Italie et l'Allemagne ne gagnent chacune qu'un seul match, le facteur h est tiré vers le bas, même si l'Espagne et la Belgique gagnent contre 3 équipes différentes. En modifiant le lien entre la France et l'Allemagne, nous avons ainsi permis à l'Espagne, victorieuse face à 3 équipes, de gagner contre 2 équipes plus "fortes" tandis que la Belgique elle n'est pas récompensée car sa seule victoire valorisante est celle contre la France.
Vous pouvez voir sur la Figure 1.6 la représentation du facteur h de l'Espagne en fonction du nombre de matchs remportés par les équipes contre lesquelles elle s'est imposée.

5. Elle est limitée à un indice de valeur 2 car la France n'a gagné que 2 matchs et que même dans le cas où la France aurait eu une victoire supplémentaire, il aurait fallu une troisième équipe possédant au moins 3 victoires.

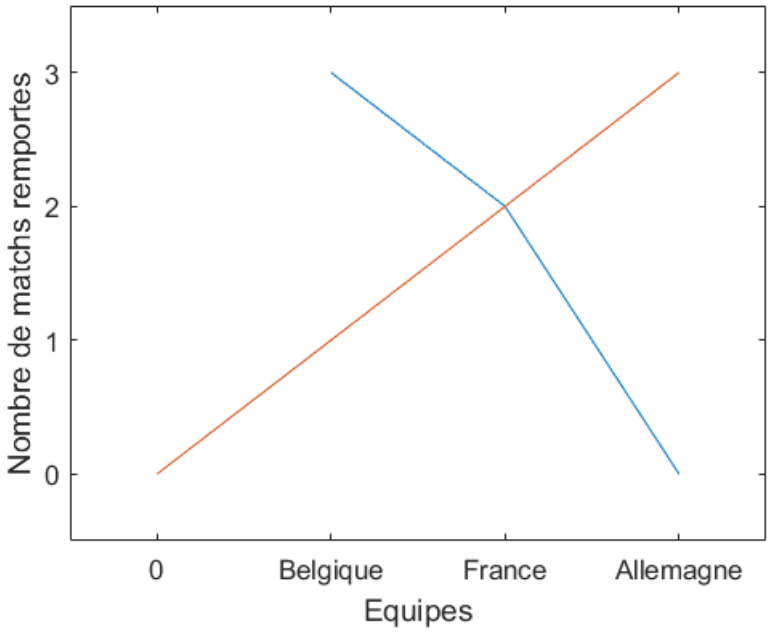

FIGURE 1.6 – Graphe de la Figure 1 modifié

Il est évident, dans le cas d'un graphe si peu fourni et si peu complexe, que le facteur h n'a que peu d'intérêt étant donné que le nombre de victoires possibles est limité. Il sera intéressant de voir dans la suite du travail, ce que permet ce facteur lorsque nous l'utilisons sur des graphes beaucoup plus grands tant par le nombre de noeuds que le nombre d'arêtes qui les relient.

1.5 L'indice de Jaccard

Étant donné que nous allons travailler avec de nombreux classements reprenant des équipes de football sur plusieurs années différentes, il est également intéressant de pouvoir les comparer. Nous aurons, pour certaines informations, recours à une classification manuelle, c'est à dire que nous regarderons simplement si les classements « ont l'air » similaires. [6] Pour d'autres informations nous aurons cependant besoin d'une expertise plus critique qui nous permettra de comparer deux classements dans leur globalité.

Cette expertise, nous l'obtiendrons en utilisant l'indice de Jaccard. L'indice de Jaccard est un coefficient qui permet de comparer deux ensembles, plus les ensembles seront similaires, plus cet indice sera élevé. Cet indice se calcule en prenant le rapport entre le cardinal de l'intersection des deux ensembles et le cardinal de leur union, mathématiquement, nous l'écrivons comme ceci,

$$J(A,B) = \frac{\#(A \cap B)}{\#(A \cup B)}.$$

Avec un tel calcul, on se rend vite compte que si nos deux ensembles, que nous allons par la suite appeler liste, sont identiques, nous aurons un indice de Jaccard qui vaudra 1 car

6. Il est aisé de vérifier si les trois premières équipes d'un classement se retrouvent dans le top 10 d'un autre classement.

l'union et l'intersection seront confondues et auront donc le même cardinal.

Un « problème » subsiste encore dans l'utilisation de notre indice, si nous souhaitons par exemple comparer le top 10 de deux classements, cet indice nous permet de déterminer si oui ou non les listes d'équipes sont les mêmes. Il ne permet cependant pas de déterminer, en l'état, si un quelconque ordre est préservé. En effet, si les ensembles de nos différents tops 10 contiennent les mêmes 10 équipes, l'indice de Jaccard nous permettrait de conclure que les listes sont similaires même si l'ordre est totalement inversé. Pour palier à ce problème d'ordre, nous allons modifier cet indice en utilisant la méthode décrite dans un document réalisé par le département de mathématique de l'Unamur en collaboration avec l'équipe de recherche DICE [9].
Nous allons utiliser la même appellation que le document et appeler l'indice modifié, l'indice de Jaccard étendu. Cet indice permettra de tenir compte des listes ordonnées d'éléments. Afin d'en expliquer le fonctionnement, nous allons directement passer à un exemple utilisant les équipes de football.
Considérons les équipes suivantes, la Belgique, l'Italie et la France et imaginons le classement suivant, la Belgique première, l'Italie deuxième et la France dernière. Nous avons initialement une liste ordonnée de longueur 3 (généralisée à N), nous allons maintenant créer une liste dans laquelle certains éléments seront répétés mais pour laquelle l'ordre n'aura pas d'importance. Chaque élément de la liste initiale sera répété proportionnellement à sa position, que nous allons noter P, dans la nouvelle liste. Nous répéterons chaque élément (N−P+1) fois. Ainsi, la Belgique, qui se trouve en position P= 1, sera répétée $3 - 1 + 1 = 3$ fois. L'Italie sera quant à elle répétée 2 fois et la France ne sera présente qu'une seule fois. La nouvelle liste non ordonnée sera donc la suivante,

$$\{\text{Belgique, Belgique, Belgique, Italie, Italie, France}\}.$$

Considérons maintenant un deuxième classement dans lequel nous inversons les deux premières places, la liste non ordonnée sera la suivante,

$$\{\text{Italie, Italie, Italie, Belgique, Belgique, France}\}.$$

Il nous suffira ensuite pour comparer ces deux listes, de calculer leur indice de Jaccard classique, l'indice de Jaccard étendu de deux listes est donc simplement l'indice de Jaccard de deux listes remaniées. Dans notre exemple, l'intersection des deux listes est

$$\{\text{Belgique, Belgique, Italie, Italie, France}\},$$

tandis que leur union est

$$\{\text{Belgique, Belgique, Belgique, Italie, Italie, Italie, France}\}.$$

En effectuant le quotient entre les deux cardinaux, nous obtenons un indice de Jaccard qui vaut $0,7143$ $\left(\frac{5}{7}\right)$. Pour plus de faciliter, nous allons multiplier cette indice par 100 afin d'obtenir un pourcentage de similitude entre les deux listes. Dans le cas de l'exemple, les deux listes sont similaires à $71,43\%$. La motivation derrière cette procédure est que si une équipe perd beaucoup de positions, alors elle sera présente en petit nombre dans sa liste élargie, ce qui pourra engendrer un gros changement de l'indice. Si par contre une équipe perd peu de positions alors le nombre de fois qu'elle sera reprise dans la liste élargie ne changera pas beaucoup ainsi que l'indice de Jaccard.

1.6 Modèle proposé

Il reste deux dernières choses dont nous voudrions discuter avant de passer à la présentation des résultats et le traitement de ceux-ci. Nous aimerions vous expliquer le type de modèle que nous souhaiterions utiliser pour la création d'un nouveau classement mais afin de comprendre cette partie nous avons d'abord besoin d'introduire le fonctionnement actuel du classement pour pouvoir comparer le fonctionnement des deux modèles et espérer faire mieux.

1.6.1 Classement FIFA

Un document expliquant le fonctionnement du classement actuellement utilisé est donné sur le site officiel de la FIFA [6]. Une nouvelle méthode est entrée en vigueur en août 2018, elle est basée sur un système d'addition de points dépendants du résultat de chaque match ainsi que de son importance. L'ancienne version se basait sur une moyenne temporelle mais n'était pas concluante et a donc nécessité une modification.

Analysons donc les détails de la procédure, il faut avant tout savoir que cette façon de calculer les points a été pensée dans l'optique de rester intuitive, de manière à ce que chacun puisse comprendre le fonctionnement du classement et en diminuant l'impact qu'avait les différences de confédération [7] de manière à ce que n'importe quelle équipe puisse remonter dans le classement si ses performances sont remarquables et ce, même si ses scores passés étaient médiocres. Les points ajoutés ou retirés sont calculés en fonction de plusieurs facteurs et la formule est la suivante,

$$P = P_{\text{précédent}} + I * (R - R_a). \tag{1.9}$$

Les différents paramètres utilisés dans cette formule représentent les informations suivantes, P représente le nombre de points dans le classement FIFA après le match, $P_{\text{précédent}}$ le nombre de points de l'équipe avant le match, R le résultat du match qui est défini de la manière suivante.

$$\begin{aligned} &R = 1 \text{ en cas de victoire,} \\ &R = 0 \text{ en cas de défaite,} \\ &R = 0.5 \text{ en cas de match nul.} \end{aligned}$$

Le facteur R_a quant à lui correspond à une estimation du résultat attendu qui se calcule de la manière suivante,

$$R_a = \frac{1}{10^{\text{différence}/600} + 1}, \tag{1.10}$$

avec « différence » qui est égal à la différence entre les points des deux équipes avant le match. Nous n'avons pas trouvé d'explication justifiant le nombre 600 qui permet de diviser cette « différence » dans l'exposant. Si ce nombre est effectivement choisi de manière arbitraire, cela renforce notre idée qu'il est nécessaire d'utiliser un classement différent. Le dernier facteur, le facteur I tient compte de l'importance du match. Cette importance recense 9 niveaux qui sont les suivants.

7. Les différentes équipes nationales ne jouent pas toutes dans les même "groupes", généralement les différentes confédérations représentent les différents continents.

1. Pour les matchs amicaux disputés en dehors des fenêtres du calendrier international des matchs, I sera égal à 5.
2. Pour les matchs amicaux disputés dans une fenêtre du calendrier international des matchs, I sera égal à 10.
3. Pour les matchs de groupe de Ligue des Nations, I sera égal à 15.
4. Pour les matchs de classement et finale de Ligue des Nations, I sera égal à 25.
5. Pour les matchs de qualification pour la compétition continentale d'une confédération ou la compétition finale de la Coupe du Monde de la FIFA, I sera égal à 25.
6. Pour les matchs de compétition continentale d'une confédération jusqu'aux huitièmes de finale (inclus), I sera égal à 35.
7. Pour les matchs de compétition continentale d'une confédération à compter des quarts de finale ; tous les matchs de la Coupe des Confédérations de la FIFA, I sera égal à 40.
8. Pour les matchs de compétition finale de la Coupe du Monde de la FIFA jusqu'aux huitièmes de finale (inclus), I sera égal à 50.
9. Pour les matchs de compétition finale de la Coupe du Monde de la FIFA à compter des quarts de finale, I sera égal à 60.

Les appellations précédentes ainsi que leurs valeurs associées sont reprises du document décrivant la procédure de calcul des points [7]. Certaines modifications sont cependant apportées lors de circonstances exceptionnelles, c'est pourquoi un match se décidant par une séance de tir au but verra le facteur R modifié de manière à fournir une victoire et une défaite modérée pour les deux équipes [8] ce qui permet ainsi de tenir compte de l'équité du niveau des deux équipes. La deuxième modification sera "d'annuler" la perte de points des équipes perdantes de phases à élimination directe lors d'une compétition finale, ce qui a pour but de ne pas pénaliser une équipe ayant réussi à se hisser au sommet d'une compétition avant d'échouer contre une autre équipe ayant réussi la même performance.

Une autre précision qu'il est important de mentionner, est comment le passage d'une méthode à l'autre a elle été effectuée. Le but premier était de conserver l'ordre des différentes équipes dans le classement afin de ne pas tout déstructurer et ne pas totalement repartir de zéro. La décision a donc été prise de répartir uniformément les équipes en fonction de leur position dans le classement sur échelle de points. La valeur par défaut de la première équipe du classement a été fixée à 1600 points et les équipes étaient ensuite placées de manière à n'avoir que 4 points d'écart par rapport à l'équipe juste devant elle. La formule de transition suivante,

$$P = 1600 - (Pos - 1) * 4,$$

dans laquelle Pos représente la position de l'équipe concernée dans le classement initial, a facilement permis une transition assez fluide d'une méthode de classement à l'autre.

Voyons maintenant comment nous pourrions, si cela est possible, améliorer cette formule de manière à ce que le classement permette de prédire les équipes victorieuses lors de grandes compétitions ce qui n'est pas le cas actuellement.

8. Le facteur R de l'équipe gagnante vaudra 0.75 au lieu de 1 et celui de l'équipe perdante sera égal à 0.5 et non pas 0.

1.6.2 Notre idée de modèle

N'ayant pas la prétention de critiquer tout le travail réalisé jusqu'ici par les créateurs du classement FIFA actuel, nous voulons avant tout proposer des idées de classements « simples ». C'est justement cette simplicité qui pourrait mener, à la suite des analyses, à un échec. Nous trouvons simplement qu'il est primordial de limiter au maximum l'utilisation de paramètres, cette restriction nous empêchera certainement de faire des choix qui auraient semblé opportuns. Nous allons donc dans ce mémoire proposer 3 classements différents que nous présenterons plus tard dans les sections adéquates.

Il ne sera cependant pas possible de n'utiliser aucun paramètre, la quantité de points donnée pour une victoire en est un qui est quasiment indispensable. D'autres encore seront utilisés comme le critère dépendant du type de match dont nous détaillerons le fonctionnement plus tard dans la section 2.3.1.

Les codes seront fournis dans un projet Gitlab dont voici le lien :
`https://gitlab.unamur.be/math/mac/memoires/potentielles-ameliorations-du-ranking-fifa`.
Ils pourront donc ainsi être réutilisés et les paramètres modifiés. Au terme de nos différentes analyses qui se trouvent dans les chapitres suivants, nous conserveront l'un des trois classements, le meilleur, et nous en discuterons les avantages et les inconvénients.

Ce chapitre aura permis a chacun de se familiariser avec la manipulation de réseaux, les différentes centralités qu'il serait adéquat d'utiliser et l'idée générale qui a guidé nos analyses tout au long de notre travail. Nous allons maintenant passer au deuxième Chapitre dans lequel nous allons vous présenter les différentes données que nous avons utilisées et la façon dont nous les avons analysées dans le but d'obtenir les résultats.

Chapitre 2

Présentation des données

Dans ce Chapitre, nous allons donc vous présenter brièvement les différents fichiers qui contiennent les données que nous utiliserons dans la suite et qui permettront l'ensemble des analyses ainsi que la façon dont nous avons « trié » ces différentes données dans le but d'obtenir des jeux de données plus petits qui nous permettront une analyse plus précise.

2.1 Différents fichiers de données

Lors du choix du mémoire, nous avons récupéré un ensemble de données sur lequel nous avons travaillé tout au long de ce mémoire ainsi que différents papiers d'introduction au classement FIFA. Les fichiers de données sont au nombre de 3 et sont tous enregistrés dans un format CSV [1]. L'un des 3 fichiers est primordial pour l'entièreté des analyses car il reprend toutes les données des différents matchs internationaux de 1872 jusqu'au mois de février 2020, une ancienne version de cette base de données m'avait initialement été donnée, il est possible de trouver les mises à jour de cette ressource sur le site Kaggle [11]. Au total, ce sont près de 40000 matchs qui sont recensés dans ce fichier et pour lesquels nous possédons, la date à laquelle a eu lieu le match, les deux équipes concernées, leur score respectif, le type de tournoi dans lequel était joué le match et enfin le lieu de la rencontre. Une information qui pourra éventuellement nous être utile et qui se déduit de la précédente est la question de la neutralité du lieu qui peut avoir un impact sur la performance des joueurs. Nous devrons donc analyser ces différentes données et voir quelles informations nous seront potentiellement inutiles car pour rappel, nous désirons éviter au maximum le nombre de paramètres à introduire. Les deux autres fichiers qui m'ont été fournis contiennent premièrement les informations concernant les différentes coupes du monde ayant déjà eu lieu et deuxièmement, les différents classements FIFA par ordre chronologique avec le nombre de points de chaque équipe et différentes informations qui ne seront énoncées que plus tard si elles nous sont utiles. Ces deux fichiers seront donc utilisés, au moment de la comparaison de notre classement avec celui de la FIFA.

1. Le sigle CSV signifie Comma-Separated Values, ce format consiste en un format de fichier texte qui reprend les données d'un quelconque tableau en séparant les différents éléments du tableau par une virgule.

2.2 Préparer les analyses de données

Une première étape à franchir lorsque nous avons obtenu le jeu de données était de comprendre comment nous allions devoir les manipuler. Travaillant avec le logiciel Matlab pour construire les graphes et effectuer les analyses, nous devions dans un premier temps fournir à Matlab tout ce dont il avait besoin pour construire un graphe cohérent. Les fonctions de graphes, qu'ils soient dirigés ou non, sont toutes deux construites en fournissant la matrice d'adjacence du graphe que nous avons définie dans le premier Chapitre (Définition 8). La première chose à faire était donc de construire cette fameuse matrice d'adjacence, la représentation la plus évidente était de relier par une arête dirigée le gagnant et le perdant du match. Notre arête pointerait naturellement comme nous l'avions déjà introduit précédemment vers l'équipe perdante. Nous avons dans un premier temps tenté naïvement de construire la matrice d'adjacence générale du graphe en nous basant sur les 40000 matchs, le résultat n'était pas très concluant comme vous pouvez le voir sur la Figure 2.1. Il est impossible de constater quelque chose de précis sur ce graphe tant il est surchargé, ce que nous pouvons cependant apercevoir, ce sont les différents amas de noeuds, faiblement interconnectés, qui représentent les différentes confédérations. Ce graphe reprend l'information de plus de 100 ans de matchs de football. Nous pourrons à défaut de s'en servir de manière précise, nous servir de ce graphe pour tirer des informations générales en ce qui concerne le réseau de football international. Nous voyons que les matchs ont surtout lieu à "l'intérieur" des confédérations ce qui correspond à la logique des matchs que nous voyons généralement à la télévision.

FIGURE 2.1 – Première tentative de création du réseau.

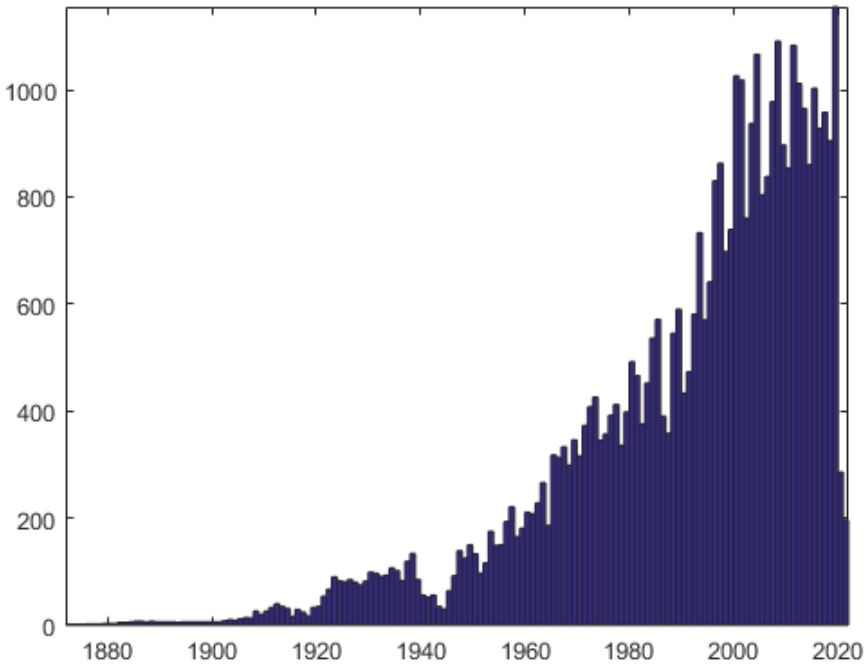

FIGURE 2.2 – Nombres de matchs de football joués en fonction de l'année.

Il nous a donc semblé évident de faire en sorte d'obtenir des informations liées à un moment précis et non pas sur une trop vaste étendue temporelle. Nous avons alors décidé de considérer les différents réseaux en fonction d'un laps de temps que nous pouvions faire varier en fonction de la période voulue. Au début du recensement, la quantité de matchs joués au niveau international était infime et ne considérer qu'une seule année revenait parfois à considérer un seul match, élément fort peu intéressant pour la création objective d'un classement. Cette quantité n'est cependant plus du tout infime pour des années du $21^{ème}$ siècle. Malgré la crise du Covid19, la quantité de matchs joués en 2021 n'a rien à envier aux années du $19^{ème}$ siècle, en effet on dénombre 196 matchs rien que jusqu'au 31 mars 2021 [2]. Vous pouvez d'ailleurs voir sur la Figure 2.2 l'évolution du nombre de matchs joués par années Nous avons donc obtenu des graphes comme ceux de la Figure 2.3 et de la Figure 2.4. La Figure 2.3 représente le graphe constitué des matchs depuis le premier que nous possédons dans notre base de données jusqu'au dernier de l'année 1880, nous pouvons voir que seulement 3 pays ont participé à des matchs internationaux, il n'y a cependant pas eu que 4 matchs (4 arêtes sont visibles sur le graphe). Lorsque nous fournissons une matrice d'adjacence à Matlab, celui-ci crée un graphe en regroupant toutes les arêtes "identiques" [3] en une seule arête en lui donnant un poids équivalent au nombre d'arêtes normalement présentes. Nous devrons tenir donc de ce poids dans nos futurs calculs de centralité, nous verrons plus tard dans ces calculs que le poids de l'arête ne dépendra pas uniquement du nombre d'arêtes. Par souci de clarté et de lisibilité, la Figure 2.3 ne comporte pas d'informations concernant les poids des arrêtes. Nous utiliserons ces données dans nos algorithmes.

2. Date de dernière mise à jour de la base de données.
3. C'est à dire des arêtes qui possèdent la même origine et la même destination.

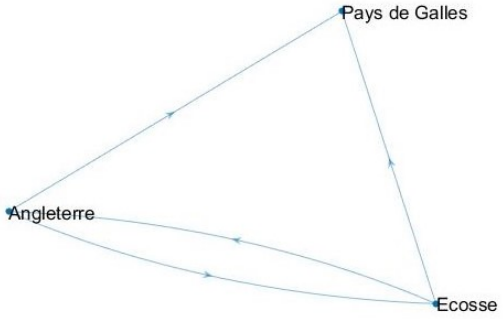

FIGURE 2.3 – Réseau international de football de 1870 à 1880.

Le graphe de la Figure 2.4 par contre présente le même problème que le graphe présenté à la Figure 2.1, il est beaucoup trop surchargé et aucune déduction n'est possible en utilisant un graphe pareil. Nous voyons donc que grâce à ces deux exemples que l'idée d'utiliser des fenêtres de match est très intéressante mais il est inutile de subdiviser ces fenêtres en suivant un laps de temps prédéfini car pour les premières années où le football n'était pas encore trop développé, le graphe ne serait pas pertinent et de la même manière, vu le grand nombre de matchs qui a lieu chaque année depuis que le foot se développe, nous nous retrouverions avec un graphe incompréhensible pour les années les plus récentes. L'information visuelle n'est cependant pas la plus importante, malgré les quelques informations que nous pouvons obtenir sur des graphes peu fournis, il est très compliqué pour ne pas dire impossible d'en déduire un classement objectif car de trop nombreuses informations sont manquantes. Nous nous intéresserons donc uniquement par la suite à des résultats numériques qui seront fiables et lisibles malgré de grands nombres d'informations. L'idée étant de toute façon de fournir un classement et non pas un graphe, peu importe l'allure de celui-ci nous nous contenterons que ses données soient intéressantes. Nous nous intéresserons donc uniquement par la suite à des résultats numériques qui seront fiables et lisibles malgré de grands nombres d'informations. L'idée étant de toute façon de fournir un classement et non pas un graphe, peu importe l'allure de celui-ci nous nous contenterons que ses données soient intéressantes.

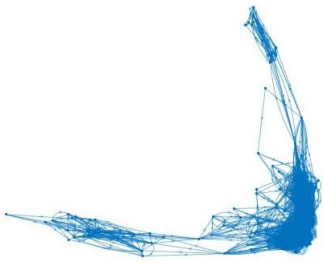

FIGURE 2.4 – Réseau international de football de 2010 à aujourd'hui.

2.3 Méthodologie utilisée pour ce travail

Nos analyses ont consisté en plusieurs points assez différents que nous allons développer ici, les résultats relatifs à ces différentes analyses seront présentés dans le Chapitre 4 qui sera entièrement consacré aux résultats.

2.3.1 Paramètres préliminaires importants

Avant de présenter les différentes analyses réalisées, il est primordial de parler des différents paramètres qui régissent nos implémentations. Ceux-ci sont des modifications des paramètres utilisés par la FIFA et sont purement liés à des choix personnels. Ils pourraient donc parfaitement être modifiés et il est possible que d'autres jeux de valeurs améliorent les résultats de ce mémoire.

1. Le nombre de points gagnés en fonction du résultat du match,
   - 3 points en cas de victoire,
   - 0 en cas de défaite,
   - 1.1 en cas de match nul à l'extérieur,
   - 0.9 en cas de match nul à domicile.

   Cette légère différence de points accordés lors d'un match nul permet de favoriser l'équipe qui ne jouait pas sur son terrain avec potentiellement moins de public.
2. Le facteur utilisé pour les différents types de matchs. Afin de respecter notre volonté qui était d'utiliser le moins de paramètres possibles, nous avons laissé un coefficient neutre pour tous les types de matchs exceptés pour les matchs amicaux. Un facteur de 0.2 a été attribué à ces matchs afin de valoriser beaucoup plus les matchs officiels.

Dans le classement officiel de la FIFA, un des paramètres importants est celui qui rend compte de la puissance des équipes. Dans la majorité de nos analyses, il a été nullifié. Ce choix peut paraître étrange mais nous avons tenu compte de la « puissance » d'une équipe par la suite en utilisant l'algorithme du Pagerank pour certaines analyses. Ne pas inclure ce facteur dans nos implémentations dès le départ permet encore une fois d'utiliser un minimum de paramètres et conserver un classement le plus « brut »possible. Nous n'avons également pas mis de facteur influençant la puissance des équipes par rapport à la confédération concernée.

2.3.2 Trois types de classement

Au terme de nos implémentations, nous avons décidé de construire et analyser 3 classements différents basés sur des principes assez distincts de ceux utilisés par la FIFA.

1. Premièrement, un classement assez simple dans lequel le seul élément influençant la position d'une équipe dans le classement est son nombre de points final, en d'autres termes, le degré sortant du noeud. Dans ce type de classement, plus une équipe à gagné de matchs plus son nombre de points est important.
2. Deuxièmement, un classement assez similaire qui se base cependant l'algorithme du Pagerank. Ce classement permet en quelque sorte de compenser la puissance des différentes équipes que nous n'avons pas prise en compte.
3. Troisièmement, un classement basé sur le facteur h de chaque équipe (voir la section1.4). Ce classement est censé être assez différent des autres étant donné que sa génération est liée à des éléments totalement différents.

Pour ces trois types de classements, une modification commune a été utilisée en cas d'égalité. Lorsque deux équipes obtiennent le même nombre de points, nous effectuons une vérification supplémentaire basée simplement sur le nombre de buts marqués. L'équipe ayant marqué le plus de buts est alors mise en avant par l'algorithme.

2.3.3 Analyses inter-classements

Une des premières étapes de l'analyse consiste à comparer chacun de nos classements entre eux ainsi qu'avec le classement FIFA. L'indice de Jaccard permettra de voir le pourcentage de similitude entre les différents classements, il sera ainsi possible de savoir si certaines implémentations se recoupent et si ce que nous avons choisi d'utiliser diffère fortement du classement FIFA ou non. Il nous sera alors possible d'éliminer un classement si la similitude entre ce classement et un autre est proche des 100%. Cette étape sera la première dont nous analyserons les résultats et son but sera de comparer les différents classements mais n'apportera aucune information capitale lié aux véritables résultats de matchs de football.

2.3.4 La fiabilité du classement FIFA contre la fiabilité de notre classement

Cette étape permettra de générer la partie la plus importante des résultats. En effet l'objectif premier de ce mémoire est avant tout que l'un de nos classements se rapproche le plus possible des résultats des différentes coupes du monde, contrairement au classement FIFA actuel qui ne prédit pas de manière satisfaisante les résultats des différents matchs. Si l'on compare les résultats annoncés par le classement FIFA et les résultats effectifs, il y a souvent peu de similitudes. Que nos classements soient forts différents de celui de la FIFA n'est donc, en soit, pas un problème.
Afin de pouvoir dire explicitement que nos classements prédisent les résultats des coupes avec un plus grand taux de fiabilité que le classement FIFA, nous allons vérifier les classements afin de voir si le trio gagnant des différentes coupes du monde se retrouve dans un top particulier. Cette partie de l'analyse nous permettra également de vérifier quel type de classement est le plus performant afin de nous concentrer sur celui-ci pour la suite des analyses.

2.3.5 Une évolution temporelle du classement

Une fois un classement ou plusieurs classements choisis, il nous a alors paru intéressant d'observer comment ces derniers évoluent au fur et à mesure des différentes années. Nous vous présenterons un graphique représentant l'évolution d'un certain nombre d'équipes en fonction des années. Un tel graphique peut potentiellement nous amener des informations sur des éventuelles prévisions. Nous serons à même de déterminer à quel moment une équipe commence à remonter dans le classement et à partir de quel moment une équipe en chute. Peut-être sera-t-il alors possible de voir qu'une équipe montante à toutes ses chances pour les compétitions futures, nous répondrons à cette interrogation à la fin des analyses.

Chapitre 3

Algorithmes et Méthodes utilisés

Dans ce Chapitre, nous allons présenter les différents codes qui ont été utilisés pour générer les différents fichiers et résultats. Nous présenterons quelques captures d'écran afin de montrer comment ils fonctionnent. Nous avons implémenté la totalité des codes Matlab de notre propre chef mais nous avons cependant utilisé certains fichiers qui existaient avant le début ce mémoire. Ce chapitre peut paraître anodin car il ne fait que présenter les différents fichiers utilisés mais nous trouvons primordial de permettre à tous de comprendre leur fonctionnement. Nous avons effectivement travaillé sur l'amélioration du classement des équipes internationales de football, mais diverses améliorations seraient encore possibles, celles-ci seront d'ailleurs décrites dans le chapitre 5 mais il est capital de comprendre l'utilisation des différents codes et leur fonctionnement pour pouvoir comprendre les analyses suivantes et pouvoir éventuellement les retravailler dans un potentiel travail futur.

3.1 Fichiers et éléments d'analyses reçus

Nous avons, dans un premier temps, utilisé un fichier python qui permettait de ressortir les différents classements FIFA directement du site officiel de la FIFA. Le code n'est malheureusement plus opérationnel, le site a été remanié et n'est plus du tout présenté de la même façon. Nous avions heureusement en notre possession les fichiers contenant les classements FIFA de Juillet 1993 à mai 2018. Il sera donc aisé de réaliser les tests concernant les coupes du monde ayant lieu entre ces deux dates. En ce qui concerne les dates antérieures, le classement FIFA n'ayant pas encore été créé, nous pourrons uniquement comparer notre classement avec les résultats des coupes du monde.

3.2 Fichiers implémentés

Cette partie sera bien plus fournie étant donné que nous allons expliquer à quoi ont servi les différents codes et à quel point ils sont modulables. Nous fournirons également un lien sur lequel les différents fichiers seront disponibles. Ces explications permettront donc premièrement que les lecteurs comprennent comment les différents résultats ont été générés et deuxièmement que les personnes intéressées, s'il y en a, reproduisent les tests.

Pour générer les différents résultats, nous avons utilisé le logiciel Matlab [1]. Onze fichiers différents ont été implémentés pour pouvoir mener à bien les analyses, mais nous n'en

1. La version précise est la suivante, « Matlab R2016a ».

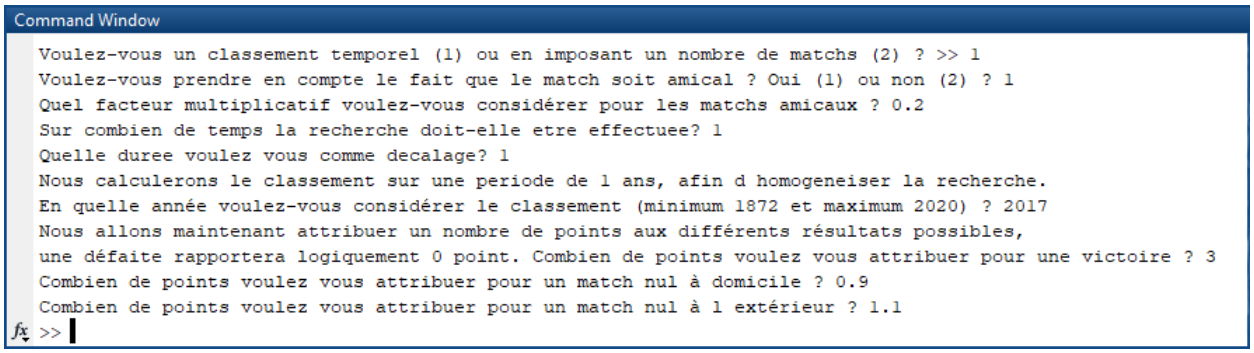

FIGURE 3.1 – Aperçu de la boîte de commande du Code_general.m

présenterons que sept, nous expliquerons en quelques lignes le fonctionnement des quatre derniers à la fin de ce chapitre. Deux de ces fichiers sont des parties nécessaires au bon fonctionnement du code général.

3.2.1 Code_general.m

Ce fichier est la partie principale des implémentations, grâce à lui, il est possible de générer les premiers fichiers permettant la création des classements. La première phase de l'exécution est une étape qui permet de générer la liste principale sur laquelle nous allons travailler, il s'agit de la liste contenant l'ensemble de toutes les équipes présentes dans la base de données des matchs. L'ordre d'apparition des équipes dans la liste va dépendre de la date d'apparition de l'équipe dans la base de données.
Vient ensuite une étape de sauvegarde qui mène à la partie essentielle de ce fichier, l'interaction avec l'utilisateur. Étant donné que les paramètres d'exécution doivent être modulables, il est impératif de tenir compte des valeurs choisies pour les différents paramètres. Concrètement, le code pose les questions visibles sur la Figure 3.1 à l'utilisateur, ce dernier peut alors décider d'encoder les paramètres comme bon lui semble, attention toutefois que ces différents paramètres impacteront les classements qui seront générés par la suite.

Ce code permet globalement de différencier les classements dès la première question pour organiser un classement qui soit défini par une durée fixe ou par un nombre fixe de matchs. Dans le cas d'un classement temporel, l'entièreté des matchs d'une période donnée sont pris en compte, une option a été implémentée dans le but de ne pas devoir faire le code plusieurs fois pour des périodes consécutives. Ainsi si nous souhaitons générer des classements consécutifs de 3 ans, et avoir l'information sur les 10 années qui précèdent, le code générera 4 classements successifs de 3 ans.
Dans le cas d'un classement dépendant d'un nombre de matchs N, le code commencera simplement à une date donnée et prendre en compte les N premiers match, si nous souhaitons lui faire réaliser plusieurs classement à la suite, il recommencera le deuxième classement au $N + 1$ ème match et considéra les N suivants.

3.2.2 GenDonnee.m

Ce petit fichier sert uniquement à générer les « débuts »de classements. Il sert à alléger donc dans le cas d'un classement temporel, à générer les différentes dates à partir desquelles les classements démarreront [2]. Dans le cas de classements régis par un nombre de matchs fixe, cette fonction permet de fournir la date de début du premier classement ainsi que le nombre de matchs nécessaires. Cette structure sera utilisée de toute façon dans les codes suivants qui varieront en fonction de l'option choisie, classement temporel ou classement lié au nombre de matchs.

3.2.3 GenScore.m

Ce fichier, à défaut d'être le fichier principal, est celui qui contient la plus grosse partie des implémentations. En fonction de l'option choisie au préalable, le code va se diriger vers la partie qui lui correspond afin de pouvoir générer ce qui nous intéresse. Dans les deux cas, le code va parcourir le fichier results.csv afin d'obtenir toutes les informations nécessaires dont nous avons besoin. Au terme de son exécution, 3 matrices seront générées.
La première contient l'ensemble des goals marqués par les différentes équipes, l'élément m_{ij} est par exemple égal au nombre de goals marqués par l'équipe i contre l'équipe j.
La deuxième contient les « scores »que nous pourrions donner aux différentes équipes en fonction de leurs victoires. Prenons par exemple le cas d'un réseau dans lequel chaque victoire rapporte 3 points et chaque match nul 1 point, si l'équipe i gagne 4 matchs contre l'équipe j, réalise 2 matchs nuls et perd 3 matchs, elle aura au total 14 points qui seront stockés dans l'élément m_{ij}. L'élément m_{ji} sera quant à lui égal à 11.
La troisième matrice sera celle relative au facteur h. Etant donné la présentation de ce facteur qui a été faite au point 1.4, la seule information dont nous aurons besoin, est le nombre de matchs gagnés par chaque équipe et ce contre chaque équipe. Nous avons donc simplement incrémenté l'élément m_{ij} de 1 pour chaque nouvelle victoire de l'équipe i contre l'équipe j.

Grâce à ces trois différentes matrices, il nous a été possible de générer l'entièreté des résultats et par conséquent de réaliser l'ensemble des analyses. Ces analyses ont été réalisées à l'aide de fichiers qui suivront ceux permettant la génération des classements.

3.2.4 GenGraph.m

Ce fichier utilise principalement la deuxième matrice générée par le fichier précédent. Il sert à générer nos deux premiers classements, c'est à dire le classement se basant uniquement sur le score des équipes et le classement utilisant l'algorithme du Pagerank. Dans les deux cas de figure, nous avons décidé de « diviser » le classement afin de considérer les composantes connexes séparément, en effet il est totalement illogique de placer deux composantes non connexes dans un même classement car nous ne saurions pas comment les comparer entre-elles. Pour la suite des analyses, nous considérerons uniquement la composante principale. Les composantes restantes sont généralement constituées de petites équipes isolées qui jouent uniquement entre-elles et n'influencent en rien les autres équipes.

2. Les dates représentent le début de chaque classement mais elles servent respectivement de date butoir pour le classement précédent. Cela facilite donc l'organisation du code.

Le premier classement se constitue en comptabilisant les scores accumulés par une équipe contre ses adversaires. Pour rappeler ce que nous avons dit dans le point 2.3.2, ce classement consiste donc a classer les équipes en fonction de leur degré sortant.

Le deuxième classement quant à lui se construit sur la base de la même matrice de score que pour le classement précédent, à la différence que dans nos implémentations, nous calculons la centralité de chacune des équipes en utilisant l'algorithme du Pagerank. Cet algorithme est déjà pré-implémenté dans Matlab au sein de la fonction *centrality*. Nous l'avons utilisé de la manière traditionnelle en laissant la variable *FollowProbability* à sa valeur par défaut qui est de 0.85. Une fois le vecteur de centralité calculé, nous ordonnons les équipes en suivant un ordre croissant.

Une fois l'exécution du code finie, une sauvegarde automatique de ces différents classements a lieu de manière à pouvoir les utiliser facilement lors des analyses suivantes. Il est également le seul à fournir une représentation visuelle des graphes. Comme nous l'avons mentionné plus haut, cette représentation visuelle est beaucoup moins intéressante que les résultats numériques. Nous trouvons cependant important de vous montrer les graphes obtenus lors d'un exemple d'exécution afin de justifier notre choix de ne considérer que la composante connexe principale.

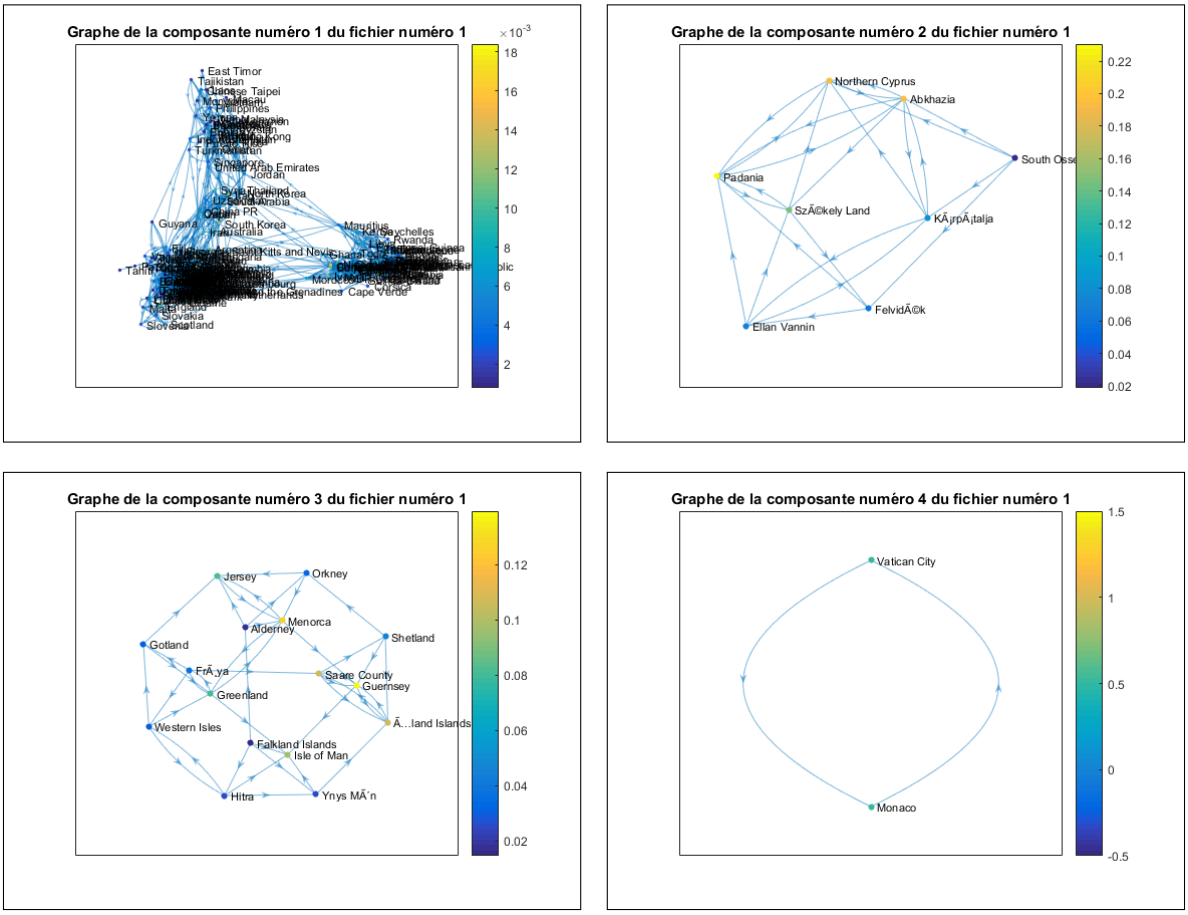

FIGURE 3.2 – Aperçu des graphes générés par la fonction GenGraph.m

Sur la Figure 3.2, nous pouvons voir les 4 composantes connexes d'un graphe dont le réseau se constitue de l'ensemble des matchs de l'année 2017 avec le reste des paramètres classiquement utilisé, 0.2 pour le facteur concernant les matchs amicaux et respectivement 3, 0.9 et 1.1 points pour chaque victoire, nul à domicile et nul à l'extérieur. Vous pouvez voir sur ces figures que certains noms d'équipes sont mal orthographiés, cela concerne les équipes possédant des caractères spéciaux dans leurs noms qui n'ont malheureusement pas été accepté par le fichier contenant les données. Il semble par conséquent évident de ne travailler qu'avec la composante connexe numéro 1 qui est celle dans laquelle se passe la majorité des activités.

3.2.5 FacteurH.m

Ce fichier permet de générer le $3^{ème}$ et dernier classement que nous avons utilisé, celui-ci correspond, comme le nom du fichier l'indique, à la création du dernier classement qui utilise le critère du facteur h. Le code a donc été effectué dans l'idée de calculer le facteur h de chacune des équipes avant de classer celles-ci en respectant la croissance de leur facteur respectif.

3.2.6 Jaccard.m

Ce fichier est probablement celui qui va le plus nous servir lors des analyses. En effet, nous avons implémenté à l'intérieur de celui-ci, l'indice de Jaccard qui a été présenté dans la section 1.5. Nous pouvons grâce à ce fichier, fournir deux classements, et obtenir la similitude entre les deux. Nous avons donc effectué toute une série de tests afin de comparer les classements entre eux et avec le classement de la FIFA. Il nous sera alors possible de voir si certains classements sont redondants ou si tous les classements sont intéressants en eux-mêmes. L'analyse concernant ces similitudes seront faites dans les sections ultérieures.

Sur la Figure 3.3, nous pouvons voir un exemple d'exécution du code. Celui-ci demande dans un premier temps à l'utilisateur de choisir jusqu'à quelle place il doit vérifier la similitude entre les deux classements. Dans l'exemple considéré, nous avons choisir de tester la similitude entre le classement FIFA du 17 mai 2018 et notre classement généré à l'aide du facteur h sur l'année 2017. Le nombre de combinaison possible étant très élevé, nous avons préféré ne pas automatiser le choix des fichiers, ces choix sont facilement modifiables et ne sont donc pas contraignants à gérer. Nous voyons sur cette figure que le taux de similitude entre les deux classements est de 15.7%, nous analyserons ces similitudes en détail dans le chapitre suivant et grâce à elles, nous pourrons comparer les différents classements entre eux.

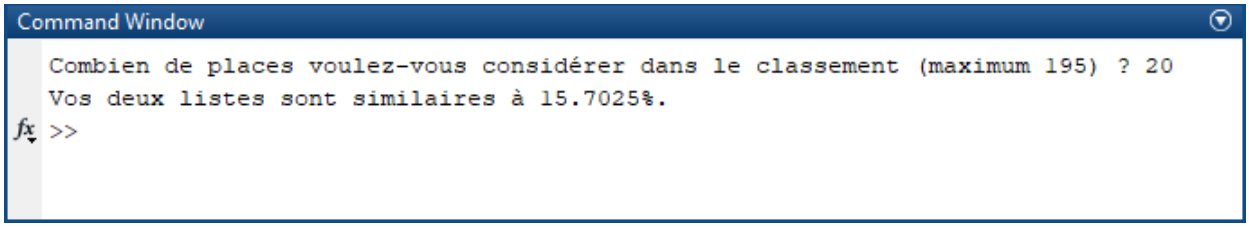

FIGURE 3.3 – Aperçu du test de similitude par l'indice de Jaccard.

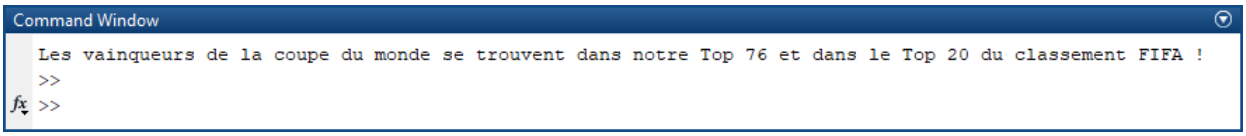

FIGURE 3.4 – Aperçu du code cherchant ou se situe le top 3.

3.2.7 RechercheTop.m

Ce dernier fichier permet également de comparer deux classements mais pas sur leurs similitudes. Le but de cette partie de code est de reprendre le trio de tête d'une des coupes du monde et de les situer dans les deux classements comme expliqué dans la section 2.3.4. Les résultats obtenus dans cette section permettront ensuite de déterminer le classement le plus efficient.

Sur la Figure 3.4, nous pouvons voir un exemple du fonctionnement du dernier code implémenté. Les 3 équipes gagnantes de la coupe du monde 2018, qui pour rappel, sont la France, la Croatie et la Belgique, se retrouvent dans les 76 premières équipes de notre classement et dans les 20 premières équipes du classement FIFA. L'inconvénient avec ce « Top », est qu'il est totalement influencé par la troisième place. Nous pourrions tout aussi bien trouver le Top 2 dans les 5 premières équipes et retrouver la troisième équipe à la $76^{ème}$ place que de trouver tout le Top 3 dans les places proches de la $76^{ème}$. Nous devrons donc faire attention à cela dans le cadre de nos analyses.

3.2.8 Fichiers restants

Les cinq derniers fichiers relèvent plus de « l'utilitaire », en effet, deux d'entre sont des fichiers servant uniquement à générer des graphes, respectivement la répartition du nombre de matchs par année et l'évolution des classements sur une période de dix ans. Le troisième est un simple code permettant de supprimer des lignes et des colonnes de nos matrices d'adjacence lorsque certaines conditions sont remplies. Les deux derniers sont des redites des fichiers GenGraph.m et FacteurH.m qui ont pour simple différence de ne pas générer des classements réels mais l'ensemble de nos matrices et classements aléatoires.

Nous avons jugé bon de présenter les différents codes ainsi que leurs utilités car ils sont essentiels à la construction de ce mémoire, de la manipulation initiale des données à l'analyse des classements en passant par la génération de ces derniers, tout ce qui est commenté ici est le résultat de ces différents codes. Pour rappel, les différents codes sont disponibles via le lien suivant :
`https://gitlab.unamur.be/math/mac/memoires/potentielles-ameliorations-du-ranking-fifa`
Dans le chapitre suivant nous allons présenter les différents résultats obtenus et les commenter au mieux.

Chapitre 4

Résultats

Dans ce chapitre nous allons reprendre l'ensemble des résultats fournis par nos différentes implémentations et les analyser. Le but ici sera dans un premier temps de se rendre compte si certains classements nous permettent d'obtenir des résultats similaires. Dans un deuxième temps nous analyserons la fiabilité des classements afin de voir si certains sont meilleurs que les autres pour le but que nous nous sommes proposé et nous finirons par une analyse spéculative basée sur une évolution de la temporalité.

Il est important de préciser avant de démarrer ces analyses qu'elles ont toutes été réalisées sur la même date précise. Étant donné que nous disposons des résultats de la coupe du monde 2018 ainsi que du classement FIFA de mai 2018, nous avons généré de plusieurs manières les classements de l'année 2017 entière. Afin de ne pas perdre cinq mois d'informations, ce qui risquerait de fausser les résultats, nous avons comparé nos classements avec celui du 18 Janvier 2018 de la FIFA. Ce dernier servira de référence lorsque nous ne préciserons pas avec quel classement FIFA nous travaillons.

Afin de passer aux différentes analyses, nous avons donc réalisé nos classements sur quatre intervalles de temps différents, ceux-ci tiennent donc compte dans leurs constructions des matchs des années précédentes et ce jusqu'à 4 ans d'ancienneté. Nous avons donc simplement considéré pour le premier classement l'ensemble des matchs de l'année 2017. Le second est quant à lui constitué des matchs des années 2016 et 2017 et ainsi de suite. Cela nous fait donc au total douze [1] classements sur lesquels nous allons porter nos analyses. En faisant varier le laps de temps nous aurons accès à une information qui est liée à la réinitialisation du classement après un certain laps de temps. Intuitivement parlant, une plus grande continuité de nos classements devrait nous permettre de nous rapprocher de celui de la FIFA. Un classement année après année devrait cependant être plus fiable étant donné que les résultats, d'une coupe du monde par exemple, ne sont pas fondamentalement impactés par le passé mais plutôt par la force actuelle des différentes équipes en présence.

4.1 Résultats inter-classements

La première phase d'analyse concerne donc les comparaisons entre nos différents classements relatifs à un même laps de temps ainsi qu'avec le classement FIFA de janvier 2018. Nous allons pour simplifier la lecture, diviser cette analyse en fonction du nombre d'années

1. Les classements sont construits de trois façons différentes pour chaque période utilisée. Vu que nous utilisons quatre périodes, cela nous fait bien douze classements.

considérées. Ces similitudes seront calculées et affichées pour les 10 premières équipes de chaque classement. Afin de compléter nos analyses et ne pas nous arrêter aux dix premières équipes, nous avons également calculé ces similitudes dans le cas où nous nous intéressions aux 20 premières équipes. Les résultats de ces analyses serviront de compléments aux différentes figures et analyses. Nous y ferons donc mention lorsque cela sera judicieux.

Par souci de simplicité, nous allons parfois remplacer le nom des différents classements par un nom plus générique. Ainsi, « classement numéro x » dénotera le classement x en respectant l'ordre présenté dans la section 2.3.2. Dans les différentes analyses qui vont suivre, nous avons décidé de fixer une période temporelle et de comparer les méthodes. Nous ne réaliserons pas l'analyse inverse qui consisterait à fixer une méthode et comparer les différentes périodes utilisées. Nous avons en effet jugé que les changements seraient trop importants et inintéressants car certains classements auraient tenu compte uniquement des résultats récents et d'autres également des résultats passés. Cela aurait entraîné une potentielle désinformation due à de possibles changements au sein de l'équipe.

4.1.1 Classement réalisé sur une matrice aléatoire

Nous allons, avant toutes autres choses, réaliser nos tests de similitudes sur des classements issus d'une génération aléatoire. Cela nous permettra de savoir si les similitudes observées plus tard, sont dues uniquement à la manière dont sont construits les classements ou si ces similitudes tiennent compte des différentes constructions de réseaux préalables. Afin de réaliser ces tests, nous avons généré 1000 matrices basées sur de l'aléatoire et nous

Synthèse des différentes similitudes entre les classements.

	Classement numéro 1	Classement numéro 2	Classement numéro 3	Classement FIFA
Classement numéro 1	/	10.65%	2.865%	1.9528%
Classement numéro 2	10.65%	/	2.1351%	1.931%
Classement numéro 3	2.865%	2.1351%	/	2.7822%
Classement FIFA	1.9528%	1.931%	2.7822%	/

FIGURE 4.1 – Tableau comparatif des résultats pour une matrice aléatoire et 10 équipes.

avons ensuite calculé la moyenne des similitudes. Par exemple, si nous prenons deux matrices M1 et M2, nous pouvons construire, grâce à elles, 6 classements dont 3 sont relatifs à M1 et 3 sont relatifs à M2. Pour calculer la similitude entre le classement 1 et le classement 2, nous avons calculé la similitude entre ces deux classements pour M1, réalisé le même calcul pour M2 et ensuite pris la moyenne des deux similitudes. Notre génération de matrice aléatoire est cependant très simple et respecte très peu d'éléments relatifs à l'étude de matchs de football, nous détaillerons dans les perspectives une meilleure manière pour implémenter ces matrices et pouvoir faire de meilleures analyses.

Nous pouvons observer sur la Figure 4.1, que hormis le taux similitude entre le classement numéro 1 et le classement numéro 2, tous nos taux sont largement inférieurs à 10%. Il n'y a donc à priori pas de lien flagrant entre nos différents classements lorsque nous considérons une matrice aléatoire. Cette constatation est également vérifiée lorsque nous nous intéressons aux 20 premières équipes des classements. Le tableau relatif est disponible dans les annexes sur la Figure 1.

4.1.2 Classement sur un an

On remarque que dans ce cas, il y a très peu de similitudes entre les classements. Les deux seuls que nous pourrions à priori comparer sur la Figure 4.2, sont les numéros 1 et 2, rappelons que ceux-ci sont à la base générés par une même matrice d'adjacence. Ces taux de similitudes sont, malgré tout, en permanence bien plus élevés que dans le cas de

Synthèse des différentes similitudes entre les classements.

	Classement numéro 1	Classement numéro 2	Classement numéro 3	Classement FIFA
Classement numéro 1	/	39.2405%	19.5652%	18.2796%
Classement numéro 2	39.2405%	/	19.5652%	14.5833%
Classement numéro 3	19.5652%	19.5652%	/	14.5833%
Classement FIFA	18.2796%	14.5833%	14.5833%	/

FIGURE 4.2 – Tableau comparatif des résultats pour un laps de temps de un an.

notre matrice généré aléatoirement. Nous pouvons remarquer que les résultats respectent une certaine structure dans le sens où le taux de similitude le plus élevé implique les deux premiers classements dans les deux cas de figures là où le reste des taux est assez constant. Nous verrons pour les autres durées d'analyse si cette ressemblance persiste. Le manque d'informations dû au fait qu'une seule année de match ait été prise en compte pourrait expliquer certains résultats.

Cette similitude entre les deux chutes cependant dans le cas où nous considérons les 20 premières équipes, cela est visible dans l'annexe 2. En élargissant le spectre des équipes considérées, le classement numéro 1 voit l'ensemble de ses similitudes avec les autres classements (excepté FIFA converger vers les 30%. Le reste des combinaisons varie de manière moins compréhensible avec une diminution globale des similitudes impliquant le classement FIFA et une augmentation de la similitude entre les classements numéro 2 et 3.

4.1.3 Classement sur deux ans

Dans ce deuxième cas de figure, on remarque une hausse générale des similitudes à l'exception de deux qui baissent légèrement. Le fait que ces taux de similitudes augmentent laisserait donc penser que tenir compte de deux années consécutives est un facteur intéressant dans la création des classements. La similitude plus importante entre les deux premiers classements semble se confirmer et on peut déjà remarquer que le classement numéro 3 semble être assez différent des autres étant donné ses similitudes plus faibles qui sont visibles sur la Figure 4.3.

Synthèse des différentes similitudes entre les classements.

	Classement numéro 1	Classement numéro 2	Classement numéro 3	Classement FIFA
Classement numéro 1	/	37.5%	14.5833%	22.2222%
Classement numéro 2	37.5%	/	23.5955%	26.4368%
Classement numéro 3	14.5833%	23.5955%	/	12.2449%
Classement FIFA	22.2222%	26.4368%	12.2449%	/

FIGURE 4.3 – Tableau comparatif des résultats pour un laps de temps de deux ans.

En augmentant le nombre d'équipes considérées, on constate cependant une augmentation globale de toutes les similitudes sauf celle entre les deux premiers classements. A l'exception des similitudes entre le classement 1 et 3 et le classement 3 et celui de la FIFA, toutes les similitudes tournent aux alentours des 33% comme cela est visible sur l'annexe 3. Le classement 3 reste donc encore une fois celui qui est le moins similaire aux autres. Il faudra donc vérifier que cette différence se retrouve dans les deux derniers cas de figure. Si cette différence est effective, il sera intéressant de conserver ce classement pour les analyses suivantes afin d'obtenir des résultats contrastés.

On remarque aussi sur les deux premiers cas de figure que le fait d'augmenter le spectre des équipes nous permet d'avoir un taux de similitude plus ou moins constant et moins éparpillé.

4.1.4 Classement sur trois ans

Pour cet avant dernier cas de figure, on se retrouve avec des taux de similitudes beaucoup plus faibles, même la ressemblance entre le classement numéro 1 et le classement numéro 2 semble être minime alors qu'elle est la seule valeur de la Figure 4.4 à dépasser les 20%. En considérant trois années de matchs consécutives dans le classement, nous n'obtenons étrangement pas d'amélioration par rapport à la Figure 4.3. Augmenter naïvement le nombre d'informations n'est donc pas forcément intéressant.

Synthèse des différentes similitudes entre les classements.

	Classement numéro 1	Classement numéro 2	Classement numéro 3	Classement FIFA
Classement numéro 1	/	20.8791%	17.0213%	15.7895%
Classement numéro 2	20.8791%	/	17.0213%	12.2449%
Classement numéro 3	17.0213%	17.0213%	/	19.5652%
Classement FIFA	15.7895%	12.2449%	19.5652%	/

FIGURE 4.4 – Tableau comparatif des résultats pour un laps de temps de trois ans.

Cette chute des taux de similitudes entre les différents classements et plus particulièrement entre les deux premiers, nous permet cependant de remarquer une évolution nette en élargissant le spectre des équipes (Annexe 4). En effet, chaque taux de similitude augmente lorsque nous considérons une vingtaine d'équipes.

Les taux de similitudes observés dans les combinaisons impliquant le classement numéro 3 semblent cependant augmenter de manière plus importante que dans les cas de figure précédent. Cela contredit les hypothèses réalisées précédemment mais pourrait suggérer une explication plausible ...

4.1.5 Classement sur quatre ans

Si nous comparons ce dernier cas de figure avec le premier par exemple, on remarque une augmentation générale de toutes les similitudes à l'exception de celle liant les deux premiers classements. Cette augmentation est d'ailleurs plus importante en ce qui concerne les combinaisons impliquant le classement FIFA et en particulier lorsque l'on compare ce dernier avec le classement numéro 3. Il est possible que cette ressemblance soit due au fait de considérer quatre années de matchs consécutives, de cette manière, nous nous rapprochons du classement FIFA qui ne se réinitialise pas.

La similitude remarquable sur la Figure 4.5 entre le classement FIFA et le classement utilisant le facteur h n'est cependant plus aussi visible lorsque nous considérons une vingtaine d'équipe. En effet, le taux passe de 55% à presque 39% sur l'annexe 5, il reste cependant

Synthèse des différentes similitudes entre les classements.

	Classement numéro 1	Classement numéro 2	Classement numéro 3	Classement FIFA
Classement numéro 1	/	26.4368%	26.4368%	26.4368%
Classement numéro 2	26.4368%	/	29.4118%	29.4118%
Classement numéro 3	26.4368%	29.4118%	/	54.9296%
Classement FIFA	26.4368%	29.4118%	54.9296%	/

FIGURE 4.5 – Tableau comparatif des résultats pour un laps de temps de quatre ans.

conséquent comparé aux autres similitudes qui se situent entre 27% et 35%. Le reste des similitudes augmente cependant toutes lorsque nous considérons un nombre d'équipes plus important.

On remarque, pour ce dernier tableau, des taux de similitudes beaucoup plus élevé en ce qui concerne les combinaisons impliquant le classement FIFA. L'hypothèse de l'augmentation du taux de similitudes dépendant du classement numéro 3 semble également se confirmer.

Nous allons maintenant tenter d'exprimer un avis général concernant ces quatre cas de figure avant de passer à la partie suivante.

4.1.6 Analyse générale

En discutant ces résultats, nous pouvons déjà souligner quelques hypothèses intéressantes. Nous pourrons à la suite de celles-ci, juger de l'utilité de certains de nos classements.

Une des premières choses qu'il est important de signifier, est le taux de similitude entre le classement basé sur la centralité de degré et celui basé sur le Pagerank qui est plus élevé que les autres. Cela peut paraître logique étant donné que ces deux classements sont construits à la base sur la même matrice d'adjacence. La seule chose qui diffère entre les deux est la façon dont nous l'avons manipulée. Dans le premier classement nous nous sommes « bêtement » contentés de sommer les scores obtenus par une équipe contre les différentes équipes. Le second classement quant à lui, accorde plus d'importance à certains résultats qu'à d'autres mais se base fondamentalement sur la même matrice. Une autre différence est également à rappeler, la centralité de degré utilisée dans le classement numéro 1 est locale, elle tient seulement en compte le degré du noeud en question, le Pagerank, qui est utilisé dans le classement numéro 2, est quant à lui moins local, car il met en valeur les équipes importantes qui sont elles-mêmes liées à d'autres équipes importantes. Il sera donc intéressant après avoir vu leurs performances respectives de se demander s'il est intéressant de considérer ces deux classements dans la conclusion ou pas.

Un deuxième élément qu'il nous semble important de préciser concerne le classement utilisant le facteur h, en effet nous avons d'abord pu voir dans les premiers cas de figure que ce facteur ne semblait ressembler à aucun autre mais rapidement les derniers cas de figure nous montrent que c'est ce classement qui affiche, en moyenne, les plus grands taux de similarité par rapport aux autres classements. Cela est loin d'être négatif en soit, le classement FIFA, bien que n'étant pas toujours utile pour prévoir l'issue des matchs, présente malgré tout une certaine cohérence, obtenir des résultats totalement différents n'apporterait par conséquent pas grand chose d'objectif. Il faudra par la suite voir si les différences encore présentes entre le classement utilisant le facteur h et celui utilisé actuellement par la FIFA sont des différences qui permettent à notre classement de prédire de manière plus fiable les résultats. Cette augmentation du taux de similitudes entre nos classements et celui de la FIFA au fil des cas de figure n'est cependant par réellement une surprise. En effet, le classement FIFA est un classement évolutif qui prend sa source il y a de nombreuses années et se base sur un très grand nombre de données, il semble donc logique que plus nos analyses portent sur une durée importante, plus elles se rapprochent du classement FIFA [2]. Rappelons également

2. En supposant que les créateurs du classement FIFA, tout comme nous ont tenté de suivre les mêmes

pour valoriser notre classement, que celui-ci est d'une simplicité déroutante étant donné qu'il se base simplement sur le facteur d'importance qu'est le facteur h, il serait donc satisfaisant d'obtenir des résultats qui collent avec les véritables résultats footballistiques en utilisant uniquement cette notion.

Nous voyons cependant que tous nos taux de similitudes sont beaucoup plus élevés lorsque nous utilisons une matrice implémentée par nos soins que lorsque nous utilisons nos classements sur une matrice aléatoire. Cela nous informe donc que nos matrices d'adjacence sont intrinsèquement liées entre elles, cela est rassurant étant donné qu'elles sont toutes censées refléter la même information au final. Nos classements sont donc assez différents les uns des autres, car nous n'obtenons malgré tout pas de taux supérieur à 60% lorsque nous considérons les 10 premières équipes, mais possèdent un seuil d'informations commun qui nous rassure sur le fait qu'aucune matrice de base n'est totalement erronée.

Un dernier point mérite d'être énoncé en ce qui concerne la variation du taux de similitudes en fonction de la variation de la quantité d'équipes considérées. En général, on remarque qu'en considérant des classements de vingt équipes, on obtient un taux supérieur à celui observé pour des classements de 10 équipes. Nous avons effectivement essayé de comparer certains classements sur l'ensemble de toutes les équipes disponibles et nous obtenons des taux de similitudes bien supérieurs à ceux obtenus avec 10 ou 20 équipes. Nous pouvons déduire de cela que les différents classements sont globalement semblables, lorsque l'on considère un grand nombre d'équipes, mais sont fortement soumis à de fréquents changements plus ou moins minimes qui « pénalisent » ce taux de similitude lorsque l'on considère un petit nombre d'équipes.

Il est difficile de tirer d'autres informations de ces taux de similitudes. Ce taux pourrait valoir la valeur que l'on souhaite, il est malheureusement possible que l'entièreté des classements soient malgré tout erronés. Il nous faudra donc passer à la phase d'analyse suivante pour pouvoir estimer de manière plus adéquate si les différents classements prédisent avec précision les résultats des différents matchs. Nous allons donc maintenant nous attaquer à la partie que nous avons jugé bon d'appeler « Fiabilité de prédiction » qui nous permettra de savoir à quel point le trio gagnant des différentes compétitions se retrouve à des positions intéressantes au sein de nos classements et de celui de la FIFA.

4.2 Fiabilité de prédiction

Dans cette partie nous allons maintenant regarder avec précision quelles sont les aptitudes de prédiction de nos différents classements. Nous allons réaliser la même analyse pour le classement FIFA afin de se rendre compte de la puissance de ce classement. Nous réaliserons également une « critique » concernant la liste d'équipes présente dans le top de notre classement afin d'avoir un élément de réflexion supplémentaire en ce qui concerne nos classements.

Avant de pouvoir discuter les résultats obtenus grâce à nos classements, nous allons réaliser nos tests sur les mêmes matrices aléatoires que celles utilisées pour calculer nos taux de similitudes. Nous allons pouvoir déterminer, de cette manière, où se situent nos résultats

lignes directrices pour donner de l'importance aux équipes qui en ont vraiment.

Tableau comparatif basé sur de l'aléatoire.

	Classement numéro 1	Classement numéro 2	Classement numéro 3
Matrice générée aléatoirement	148	146	147

FIGURE 4.6 – Tableau recensant la pire place occupée par le trio gagnant de la coupe du monde 2018 lorsque nous utilisons une matrice initiale aléatoire.

par rapport à ceux de FIFA mais également par rapport à un classement uniquement basé sur de l'aléatoire. Ces résultats issus de l'aléatoire sont disponibles sur la Figure 4.6 et seront utilisés à la fin de cette section pour critiquer l'ensemble de nos résultats.

Nous nous sommes toujours basé sur le classement FIFA de janvier 2018 et nos classements réalisés sur l'année 2017 entière. Les résultats ne sont malheureusement pas très satisfaisants. En effet, pour retrouver le trio de tête de la coupe du monde 2018 dans le classement FIFA, nous devons considérer les 15 premières équipes. La France et la Belgique étaient toutes deux dans le top 10 mais la Croatie se trouvait cependant à la $15^{ème}$ place.

En ce qui concerne nos différents classements cependant, nous pouvons voir sur la Figure 4.7 que les résultats sont beaucoup moins bons. En effet sur 12 tests, le meilleur trouve le trio gagnant dans les 19 premières places du classement. Nous ne nous attarderons évidemment pas sur la performance du classement numéro 3 lorsque nous avons considéré une seule année, l'équipe se situant à la $76^{ème}$ place aurait vraisemblablement réalisé une avancée extraordinaire pendant les 6 mois précédant la coupe du monde.

Intéressons nous cependant à ce qui se passe à l'intérieur même des classements afin de comprendre ce qui pose « problème ». Pour ce faire, observons les Figures 4.8 et 4.9. Afin de suivre l'ordre du classement, il est nécessaire de lire ces figures ligne par ligne et non pas colonne par colonne, dans le classement de la FIFA (Figure 4.8 par exemple, le Brésil occupe donc la $2^{ème}$ place et non pas la $9^{ème}$ qui est quant à elle accordée au Pérou. On remarque dans le classement FIFA une forte présence que nous connaissons tous comme étant de « grosses » équipes dont chacun a déjà entendu parler, l'Allemagne, le Brésil, l'Espagne, ... On se rend cependant compte qu'aucune équipe africaine ne se trouve dans le top 20 du classement FIFA et que l'on y retrouve très peu d'équipes méconnues. Ce constat ne se répète cependant pas dans notre classement à la Figure 4.9, on y retrouve toujours des pays dont on entend souvent parler tels que l'Allemagne, la France ou l'Espagne, mais l'on retrouve également des équipes dont on entend très peu parler au niveau international telles que la Russie, la Corée du Sud ou encore le Burkina Faso qui lui se retrouve à la $17^{ème}$ place de notre classement. Cela entraîne des placements assez illogiques tels que l'Italie et la Croatie respectivement aux places 33 et 34 du classement. Cette constatation nous encouragerait en quelque sorte à imposer un critère qui dépendrait des confédérations.

Tableau comparatif réel.

	Classement numéro 1	Classement numéro 2	Classement numéro 3
1 an de données	39	34	76
2 ans de données	30	19	29
3 ans de données	38	26	34
4 ans de données	32	24	38

FIGURE 4.7 – Tableau recensant la pire place occupée par le trio gagnant de la coupe du monde 2018.

	1	2	3	4	5	6	7	8	9	10
1	Germany	Brazil	Portugal	Argentina	Belgium	Spain	Poland	Switzerland	France	Chile
2	Peru	Denmark	Colombia	Italy	Croatia	England	Mexico	Iceland	Sweden	Wales
3	Netherlands	Uruguay	Tunisia	USA	Costa Rica	Northern Ir...	Senegal	Slovakia	Austria	Paraguay
4	Republic of ...	Scotland	IR Iran	Serbia	Ukraine	Australia	Romania	Turkey	Congo DR	Bulgaria
5	Bosnia and ...	Morocco	Egypt	Montenegro	Greece	Czech Repu...	Bolivia	Venezuela	Hungary	Jamaica
6	Cameroon	Nigeria	Panama	Ghana	Japan	Norway	Burkina Faso	Korea Repu...	Albania	Algeria
7	Russia	Cape Verde ...	Slovenia	Saudi Arabia	Honduras	Finland	Ecuador	China PR	Mali	Guinea
8	CÃƒÂ´te d'I...	Uzbekistan	Palestine	FYR Maced...	Syria	Zambia	South Africa	Uganda	United Arab...	Trinidad an...

FIGURE 4.8 – 80 premières équipes du classement FIFA de janvier 2018.

	1	2	3	4	5	6	7	8	9	10
1	Mexico	Germany	Colombia	Senegal	Cameroon	Russia	Chile	France	Spain	United States
2	South Korea	Iraq	Brazil	Portugal	Sweden	Belgium	Burkina Faso	Ivory Coast	Argentina	Japan
3	Netherlands	Venezuela	England	Honduras	Syria	Peru	Zimbabwe	Iran	Uganda	Australia
4	Denmark	Nigeria	Italy	Croatia	Costa Rica	Algeria	Libya	Kenya	Ghana	Greece
5	Bulgaria	Republic of ...	Zambia	Estonia	Bahrain	Finland	Burma	Morocco	Egypt	Uzbekistan
6	India	Uruguay	Gabon	Luxembourg	Benin	Saudi Arabia	Panama	Bolivia	Tanzania	Canada
7	China PR	Tajikistan	Wales	Austria	Romania	Qatar	Tunisia	DR Congo	Jamaica	Scotland
8	Kyrgyzstan	Madagascar	Mali	Iceland	Mozambique	North Korea	Fiji	Jordan	South Africa	Saint Kitts a...

FIGURE 4.9 – 80 premières équipes du classement utilisant le Pagerank sur une année de matchs.

Nous nous sommes également posé la question de savoir si cette différence ne résultait pas du coefficient impactant les matchs amicaux. En effet, celui-ci aurait pu être trop faible et impliquer une perte d'information due aux matchs amicaux réalisés par certaines équipes européennes. Afin de caricaturer cette pensée, imaginons que ce facteur soit carrément égal à 0, cela reviendrait à ne tout simplement pas tenir compte des matchs amicaux, une équipe qui en ferait donc un grand nombre serait considérablement désavantagée. Dans un sens cela ne changerait rien car toutes les équipes seraient sur un même pied d'égalité si le nombre de matchs officiels est identique pour chaque équipe. Il ne faut cependant pas oublier que les joueurs prennent en permanence le risque de se blesser et seraient, avec le rythme des entraînements et des matchs amicaux, beaucoup plus fatigués physiquement que des joueurs n'ayant participé à aucun match amical.

Nous avons donc testé de générer nos classements de plusieurs façons différentes afin de vérifier si ce facteur pouvait influencer ou pas le classement. Nous pouvons d'ores et déjà vous affirmer que ce facteur ne change fondamentalement rien peu importe le type de classement que nous utilisons. En effet, nous avons donc premièrement essayé de nullifier l'intérêt de ce facteur en l'égalant à 1 [3] et deuxièmement de le rendre totalement influant en l'égalant à 0.01 [4]. De cette manière, nous pouvions réaliser les classements comme si tous les matchs étaient des matchs officiels mais également en réaliser comme si seuls les matchs officiels avaient une importance.

Nous avons réalisé ces tests uniquement sur des durées de 1 et 4 ans (nous avons jugé qu'il était inutile de réaliser ces vérifications pour des laps de temps intermédiaires) et il s'avère que les résultats ne varient presque pas. Les taux de similitude entre nos classements et celui de la FIFA sont quasiment identiques. Cela est déjà en soit une information qui nous laisse penser que les changements seraient minimes, nous en avons eu la confirmation en vérifiant les places occupées par le trio gagnant de la coupe du monde. Les places obtenues étaient à quelques places près les mêmes que celles obtenues en laissant le facteur égal à 0.2. Fort de cette information, nous allons donc continuer nos analyses.

Afin de poursuivre cette analyse et pouvoir développer nos hypothèses antérieures, nous allons dans un premier temps essayer de comprendre ce qu'il s'est passé pour le classement numéro 3 après une année de matchs et nous discuterons ensuite les résultats du classement numéro 2 après deux années de matchs. En s'attardant donc un instant sur la Figure 4.10, et le résultat décevant que nous y voyons, il est possible d'insister sur l'argument que nous avons avancé précédemment qui est que de nombreuses équipes « méconnues » se trouvent à une place intéressante du classement. Certains indémodables restent cependant dans la tête du classement, dû certainement à leurs performances, comme l'Allemagne, le Brésil ou l'Espagne. Le « problème » réside cependant dans les équipes inédites qui se retrouvent placées là où elles ne devraient pas. Pour attester de la renommée de certaines équipes, il est nécessaire de préciser que de nombreuses recherches internet ont été nécessaires pour savoir à quoi correspondaient certains noms tels que « Comoros » [5] visibles sur la Figure

3. Ce facteur est un facteur multiplicatif, multiplier le nombre de points par 1 revient donc bien à ne rien faire

4. L'égaler à zéro n'était pas possible car ce faisant, le nombre de composantes connexes augmentait considérablement et ne permettait plus de créer de classements constructifs, nous pourrions toujours travailler avec la plus grande composante connexe mais nous risquerions de perdre certaines équipes importantes en route en fonction de la période considérée.

5. Les Comores en français.

	1	2	3	4	5	6	7	8	9	10
1	Cameroon	Germany	Tunisia	Egypt	Iraq	Brazil	Netherlands	Morocco	Senegal	United States
2	Spain	Mexico	Comoros	Paraguay	Bulgaria	Romania	Finland	Peru	Jamaica	Ecuador
3	Iran	Nigeria	Argentina	Libya	Iceland	South Korea	Costa Rica	Chile	Trinidad an...	Colombia
4	Australia	Poland	China PR	Ivory Coast	Denmark	Japan	Honduras	Zambia	Saudi Arabia	Czech Repu...
5	DR Congo	Guinea	France	Ghana	Algeria	Sweden	Belgium	Portugal	Brunei	Andorra
6	Niger	Saint Kitts a...	Maldives	Kosovo	Wales	French Guia...	Luxembourg	Belarus	CuraÃ§ao	Zanzibar
7	Northern Ir...	Uruguay	Bolivia	Thailand	Republic of ...	Azerbaijan	Austria	Scotland	Slovakia	Barbados
8	Montenegro	South Sudan	Norway	Kenya	El Salvador	Croatia	Serbia	Turkey	England	New Zealand

FIGURE 4.10 – 80 premières équipes du classement utilisant le facteur h sur une année de matchs.

4.10. Ce pays, qui est en réalité une île, est un état d'Afrique Australe qui se retrouve à la $13^{ème}$ place de notre classement tout en étant en $130^{ème}$ position dans le classement FIFA. Cette observation n'est pas la seule à être aussi spécifique à notre classement, on remarque effectivement pour n'en citer que quelques unes dans notre top 20, l'Egypte, l'Irak, le Maroc et l'Equateur qui sont respectivement placés aux positions 43,42,85 et 67. Attention que les équipes citées juste avant ne sont pas toutes aussi méconnues que les Comores, au contraire, l'Egypte et le Maroc font partie des trois seules équipes africaines à être un jour rentrées dans le top 10 du classement FIFA. Celles-ci ne sont cependant pas souvent constantes au sein de ce dernier. Un élément qui pourrait être « rassurant » avec l'Egypte et le Maroc, est qu'elles ont toutes les deux réussi à se qualifier pour la coupe du monde 2018 et donc faire partie des 32 équipes retenues. Nous pourrions donc, dans une idéalisation totale des résultats, nous dire que notre classement avait « prédit » cette qualification. Cela ne compense malheureusement pas le fait que la Croatie, classée $15^{ème}$ dans le classement FIFA et finaliste de cette même coupe du monde, soit classée $76^{ème}$ dans notre classement.

L'ensemble de ces réflexions renforce d'autant plus l'idée que ne pas différencier les confédérations et donner la même importance à tous les matchs [6] n'est pas une excellente façon de procéder. Nous allons maintenant analyser pour terminer cette partie, le classement qui obtient les meilleurs résultats. En analysant la Figure 4.11, nous serions presque soulagés de voir des résultats pareils. Nous retrouvons toujours quelques équipes auxquelles on ne

	1	2	3	4	5	6	7	8	9	10
1	France	Portugal	Chile	Germany	Mexico	Spain	Cameroon	Ivory Coast	United States	Russia
2	Poland	Sweden	South Korea	England	Belgium	Syria	Iraq	Senegal	Croatia	Republic of ...
3	Iran	Colombia	Brazil	Italy	Netherlands	Nigeria	Iceland	Australia	Argentina	Costa Rica
4	Venezuela	Denmark	Slovakia	Qatar	Burkina Faso	Romania	Japan	Uganda	Zambia	Tonga
5	Switzerland	Uruguay	DR Congo	Jersey	China PR	Morocco	Panama	Tunisia	Peru	Ukraine
6	South Africa	Kyrgyzstan	Jordan	Guinea	Uzbekistan	Bahrain	Turkey	Zimbabwe	Ghana	Algeria
7	Wales	Gabon	Egypt	Libya	Belarus	Honduras	Estonia	Ecuador	Austria	Canada
8	Thailand	Greece	Kenya	Northern Ir...	Hungary	North Korea	Paraguay	Mali	Bulgaria	Finland

FIGURE 4.11 – 80 premières équipes du classement utilisant le Pagerank sur deux années de matchs.

6. Nous parlons bien sur ici de la même importance pour tous les matchs amicaux et la même importance pour tous les matchs officiels. Il ne faut pas ici comprendre que les deux types de matchs ont la même importance.

s'attendait pas comme l'Irak ou le Sénégal mais nous obtenons aussi des informations très intéressantes. La France notamment qui finit première du classement, rassurant lorsque l'on sait qu'elle finit par gagner la coupe quelques mois plus tard. Si l'on considère le top 32 de notre classement, on retrouve 21 équipes qualifiées pour la coupe du monde. Nous trouvons que ce résultat est plus que correct, d'autant plus que parmi les 11 équipes présentes dans notre classement et non qualifiées pour la coupe du monde se trouvent, la Côte d'Ivoire, l'Italie, les Pays-Bas, les Etats-Unis, le Cameroun et le Chili, qui sont des équipes « phares ». En effet, certaines d'entre elles sont des équipes habituées de cette compétition et d'autres sont des équipes dont l'ensemble du milieu footballistique attendait la qualification à la suite de leurs prestations récentes.

Nous avons pris la liberté d'analyser plus en détails le classement utilisant le Pagerank afin de voir combien d'équipes étaient présentes sur les 32 qualifiées en fonction du laps de temps que nous considérions. On remarque que nous avons généralement une vingtaine en commun entre notre top 32 et les équipes qualifiées pour la coupe du monde 2018. Ce résultat pour le classement numéro 2 est malgré tout plus que satisfaisant surtout lorsque l'on se rend compte sur la Figure 4.8 vue précédemment que le résultat est similaire pour le classement FIFA, en effet sur les 32 premières équipes du classement FIFA, 21 sont reprises dans les équipes qualifiées pour la coupe du monde.

Possédant l'ensemble de ces informations, nous nous demandons alors ce qui est le mieux. Effectivement le classement FIFA retrouve le trio gagnant dans ses 15 premières équipes et pas 19, mais si l'on suit les informations qu'il nous communique, la France n'est que neuvième et la Belgique est classée devant elle en cinquième position. Nous voudrions par conséquent vous montrer quelque chose de très intéressant. Sur la Figure 4.12, vous pouvez voir les 20 premières places de chacun de nos 3 classements pour un laps de temps de deux ans.

Liste3

8x10 cell

	1	2	3	4	5	6	7	8	9	10
1	Germany	Mexico	DR Congo	Portugal	United States	Cameroon	France	Senegal	Costa Rica	Argentina
2	Brazil	Belgium	Chile	Burkina Faso	Switzerland	Spain	Panama	Morocco	Italy	Japan

Liste1

8x10 cell

	1	2	3	4	5	6	7	8	9	10
1	France	Portugal	Chile	Germany	Mexico	Spain	Cameroon	Ivory Coast	United States	Russia
2	Poland	Sweden	South Korea	England	Belgium	Syria	Iraq	Senegal	Croatia	Republic of ...

Liste4

8x10 cell

	1	2	3	4	5	6	7	8	9	10
1	France	Germany	Brunei	Iraq	Serbia	South Korea	Iran	Chile	Japan	Mexico
2	Spain	Belgium	Portugal	Syria	Republic of ...	Egypt	England	Ivory Coast	Costa Rica	Qatar

FIGURE 4.12 – Observation concernant le classement de la France par rapport à celui de la Belgique.

On remarque que dans chacun de nos classements, la France est mieux placée que la Belgique et il est également facile de voir qu'à l'exception du classement basique utilisant simplement la matrice d'adjacence du graphe, la France finit carrément première du classement.

Cela n'est vrai que dans le cas où nous considérons les matchs sur deux ans, en effet lorsque nous considérons un laps de temps de 4 ans, c'est l'Allemagne qui est première dans les classements numéro 2 et 3. Rappelons que l'Allemagne est l'équipe championne du monde en 2014, or les matchs de cette coupe ayant eu lieu endéans les 4 ans, nous les considérons également lorsque nous considérons un tel laps de temps. Si nous prenons par contre en considération les matchs sur les trois années précédant l'année 2018, les classements nous permettent au mieux de voir la France à la deuxième place.

Avant de conclure cette section, nous pouvons également nous rassurer avec un point très intéressant qui se base sur nos différentes analyses et sur la Figure 4.6. En effet, sur cette figure qui a été présentée au début de la section, nous voyons que lorsque les matrices initiales sont aléatoires, peu importe le classement utilisé, on retrouve le trio gagnant entre le top 145 et le top 150. Si nous comparons ces résultats aux nôtres, nous pourrions presque nous satisfaire du classement numéro 3 lorsque celui-ci était utilisé sur une année de données. En effet, en ayant le trio gagnant dans son top 76, le classement numéro 3 est bien plus efficace qu'un classement aléatoire.

Nous pouvons donc clôturer cette section sur plusieurs informations positives, nos classements, bien que n'étant pas parfaits semblent constater pour deux d'entre eux (les deux derniers) des résultats satisfaisants en adéquation avec la « réalité ». Nous nous sommes tous rendu compte à la suite de ces analyses qu'une durée de deux ans de recherche est celle qui semble être optimale. Cela semble en effet logique, en l'espace de deux ans, une équipe a le temps de modifier considérablement son équipe, pensons notamment aux changements réalisés dans les équipes nationales belge et française aux cours des deux dernières années. Cette modification peut donc potentiellement entraîner une nette amélioration des résultats qui seraient, peu visibles sur un an car trop peu nombreux et « noyés » dans l'ensemble des mauvais résultats précédents si l'on considère des matchs trop anciens. Nous savons maintenant également que nos classements obtiennent de meilleurs résultats qu'un quelconque classement aléatoire.

Dans la section suivante, nous allons par conséquent essayer de construire ensemble une évolution temporelle de ces classements. Nous considérerons des classements qui tiennent compte des matchs des deux dernières années et nous n'allons conserver que deux des trois classements. Le premier classement est assez similaire à celui utilisant l'algorithme du Pagerank et il est celui qui produit les moins bons résultats de manière générale.

4.3 Évolution temporelle des classements

Le but de cette section va être d'estimer le potentiel de prédiction de nos deux classements. Nous allons totalement oublier le classement FIFA et travailler uniquement avec le classement utilisant l'algorithme du Pagerank et celui utilisant le facteur h. Nous sommes partis du principe qu'il pouvait être intéressant de réaliser un graphique sur lequel nous afficherions l'évolution dans le classement d'un certain nombre d'équipes. Cela nous per-

mettrait de voir si notre classement est capable d'anticiper de potentiels bouleversements footballistiques. En cas de progression fulgurante d'une équipe dans le classement, nous pourrions potentiellement anticiper une victoire qui n'aurait pas été prévisible si nous nous étions limités à regarder la position ponctuelle de l'équipe dans ce même classement. Afin de ne pas travailler sur des équipes inintéressantes en nous basant sur le top de l'un de nos classements, nous avons décidé de travailler sur les 10 « meilleures équipes » internationales. Nous entendons par là, non pas les 10 premières équipes du classement FIFA mais les 10 plus récentes équipes à avoir atteint les demi-finales en coupe du monde. Pour ce faire nous avons dû considérer les 3 dernières et donc remonter à la coupe de 2010.

La liste des équipes est donc la suivante, la France, la Croatie, la Belgique et l'Angleterre en ce qui concerne celle de 2018, nous rajoutons l'Allemagne, l'Argentine, les Pays-Bas et le Brésil pour celle de 2014 et nous terminons par l'Espagne et l'Uruguay qui se rajoutent aux précédentes grâce à la coupe de 2010.

Afin de pouvoir construire cette « continuité » des résultats, nous avons exécuté nos codes afin d'obtenir, pour chaque année entre 2009 et 2018, les différents classements. Ces derniers tiennent chacun compte des deux années précédentes, ainsi les classements ne sont pas discontinus. Le classement de janvier 2013 par exemple tient compte de l'ensemble des matchs joués en 2011 et en 2012, les matchs de cette dernière année sont donc, comme vous l'avez compris, également comptabilisés dans le classement de janvier 2014 qui tient compte des matchs joués en 2012 et en 2013. Nous allons commencer nos analyses avec le classement dépendant du Pagerank et nous réaliserons ensuite des analyses similaires pour le classement utilisant le facteur h, nous terminerons par une analyse concernant le classement FIFA afin de pouvoir effectuer une comparaison correcte.

4.3.1 Le classement Pagerank

La Figure 4.13 nous permet d'apercevoir l'évolution du placement des dix équipes que nous venons de mentionner pendant dix ans [7] dans le cadre du classement Pagerank. Il est assez difficile de constater quelque chose sur un tel graphique étant donné que les différentes positions sont bien souvent proches les unes des autres. Cet aspect global nous permet cependant de constater quelque chose qui ne nous surprend pas, la fulgurante remontée de l'équipe belge. Celle ci passe de la $142^{ème}$ place en 2009 à la $9^{ème}$ place en 2016. Cette remontée n'est pas surprenante étant donné que nous étions certainement tous devant nos écrans lors du quart de finale opposant la Belgique à l'Argentine, la progression visible sur le graphe n'est donc que le reflet de leurs performances.

Afin d'avoir un aperçu plus clair des différentes informations, nous allons réaliser trois graphiques différents qui s'attarderont chacun sur quatre équipes en particulier. Nous allons former les groupes de quatre équipes en fonction de celles ayant accédé aux demi-finales d'une des trois coupes du monde concernées. Il nous sera alors possible de concentrer nos analyses sur le comportement des quatre demi-finalistes aux alentours de l'année qui leur correspond.

7. Nous avons relié les différents points par simplicité d'analyse mais il est évident que nous ne disposons que de l'information au début de chaque année et non pas d'une information continue.

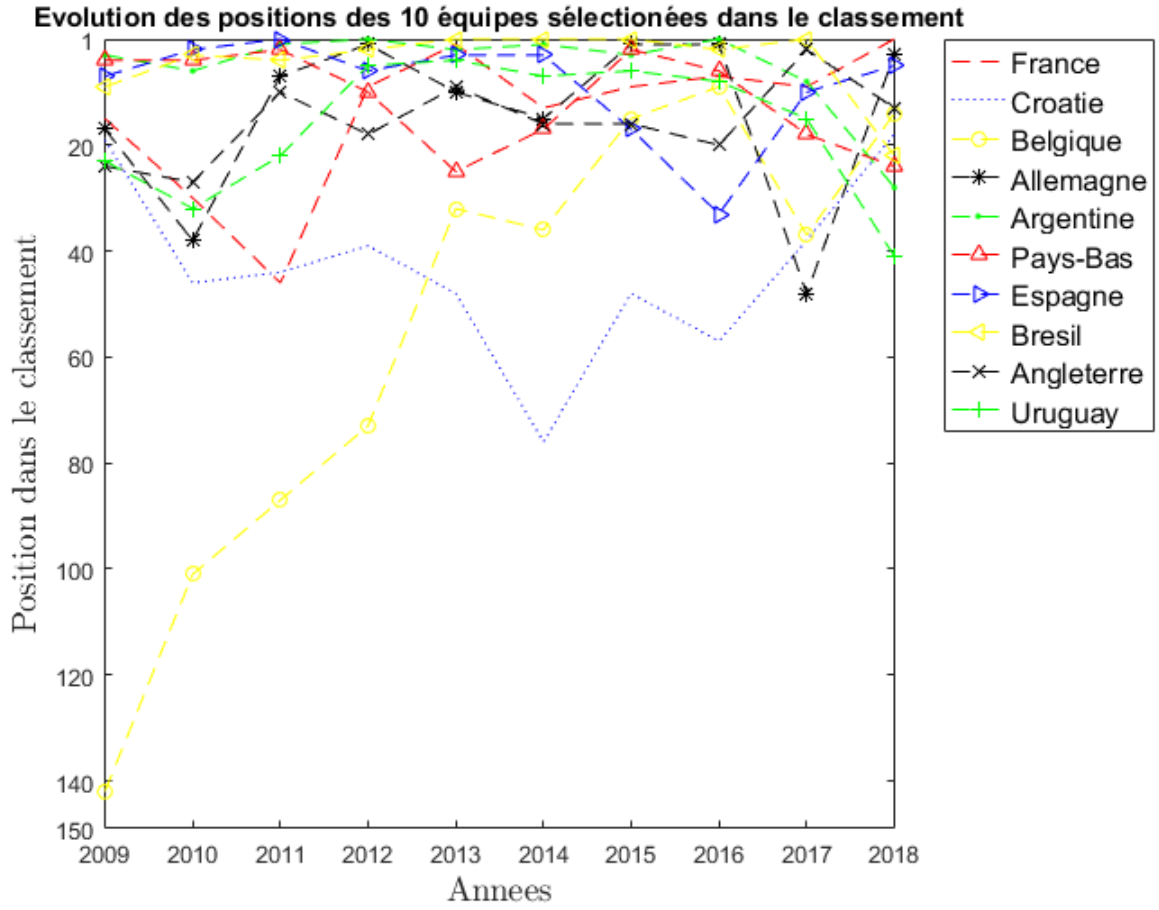

FIGURE 4.13 – Évolution du classement de 10 équipes pendant 10 ans. Classement Pagerank. Coupe du monde 2010 à 2018

Coupe du monde 2010 pour le classement Pagerank.

Rappelons dans un premier temps, pour ceux qui ne sont pas intéressés par le football en général, que les quatre équipes sont dans l'ordre, l'Espagne, les Pays-Bas, l'Allemagne et enfin l'Uruguay. Si l'on s'intéresse à la « courbe » [8] bleue de la Figure 4.14, on peut se rendre compte que celle-ci croît vers la première position lorsque l'on se rapproche des années 2010 et 2011, cela correspond parfaitement au résultat attendu. De plus, nous remarquons que les Pays-Bas qui finissent deuxième de la compétition se placent très bien dans le classement, l'Allemagne et l'Uruguay essuient cependant une lourde chute dans le classement juste avant la coupe du monde avant de tout deux remonter dans le classement. On peut également pousser plus loin l'analyse et remarquer que l'Espagne perd peu à peu sa place sur le devant de la scène avec une grosse perte de places en 2016 avant de remonter modestement. Celle-ci confirme d'ailleurs toujours sa place dans les différentes compétitions mais a réalisé des résultats beaucoup moins satisfaisants lors des coupes du monde de 2014 et de 2018. Cette perte d'efficacité semble se refléter de manière très précise sur la Figure 4.14 entre les années 2014 et 2016. Continuons directement sur la coupe du monde 2014 afin de voir si celle-ci nous permet des analyses aussi satisfaisantes.

8. Nous allons utiliser ce terme afin de faciliter la lecture

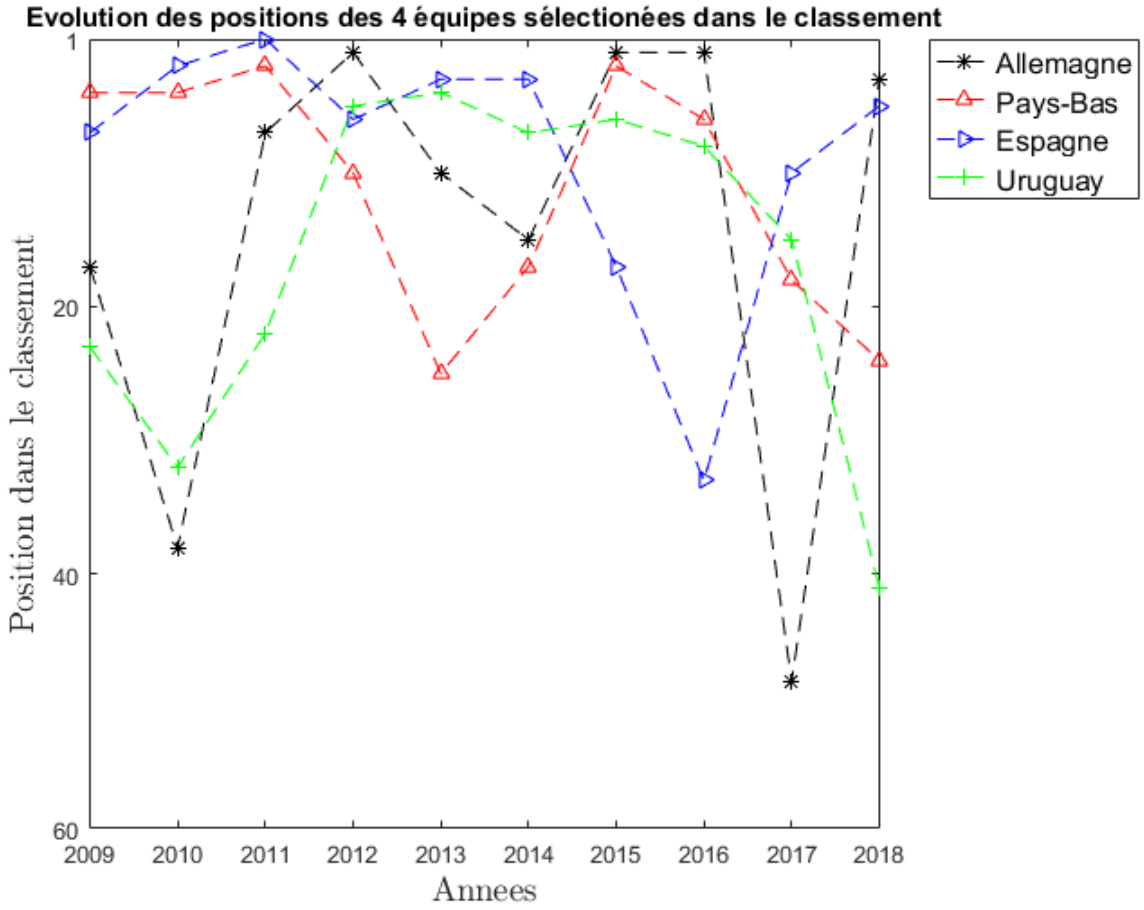

FIGURE 4.14 – Évolution du classement de 4 équipes pendant 10 ans. Classement Pagerank. Coupe du monde 2010.

Coupe du monde 2014 pour le classement Pagerank.

Lors de cette coupe du monde, les quatre équipes ayant atteints les demi-finales sont dans l'ordre, l'Allemagne, l'Argentine, les Pays-Bas et le Brésil. Sur la Figure 4.15, la victoire de l'Allemagne est cependant moins anticipable que ne l'était celle de l'Espagne sur l'analyse précédente. En effet, l'équipe allemande chute dans le classement après une grosse remontée en 2012 avant de reprendre de nombreuses positions sur l'année 2014. Chaque classement rendant compte de la place atteinte au premier mois de l'année, il est possible que l'Allemagne ait réalisé de mauvais résultats avant l'année de la coupe du monde 2014 et soit ensuite repartie du bon pied, assez pratique pour une équipe de football, pour les matchs précédant la coupe. Toujours est il que pour les autres résultats, nous remarquons que l'Argentine et le Brésil sont tous les deux très bien positionnés dans le classement et auraient d'ailleurs pu prétendre au titre, chose moins évidente pour les Pays-Bas qui occupent malgré tout une position moins importante du classement. Nous pouvons par contre constater qu'à partir de l'année 2013, le classement de cette équipe ne cesse de s'améliorer et continue d'ailleurs de le faire lors de la coupe du monde. Il aurait donc été possible avec notre graphique de prédire une bonne performance de l'équipe hollandaise mais il aurait été difficile de la prétendre apte à la troisième place.

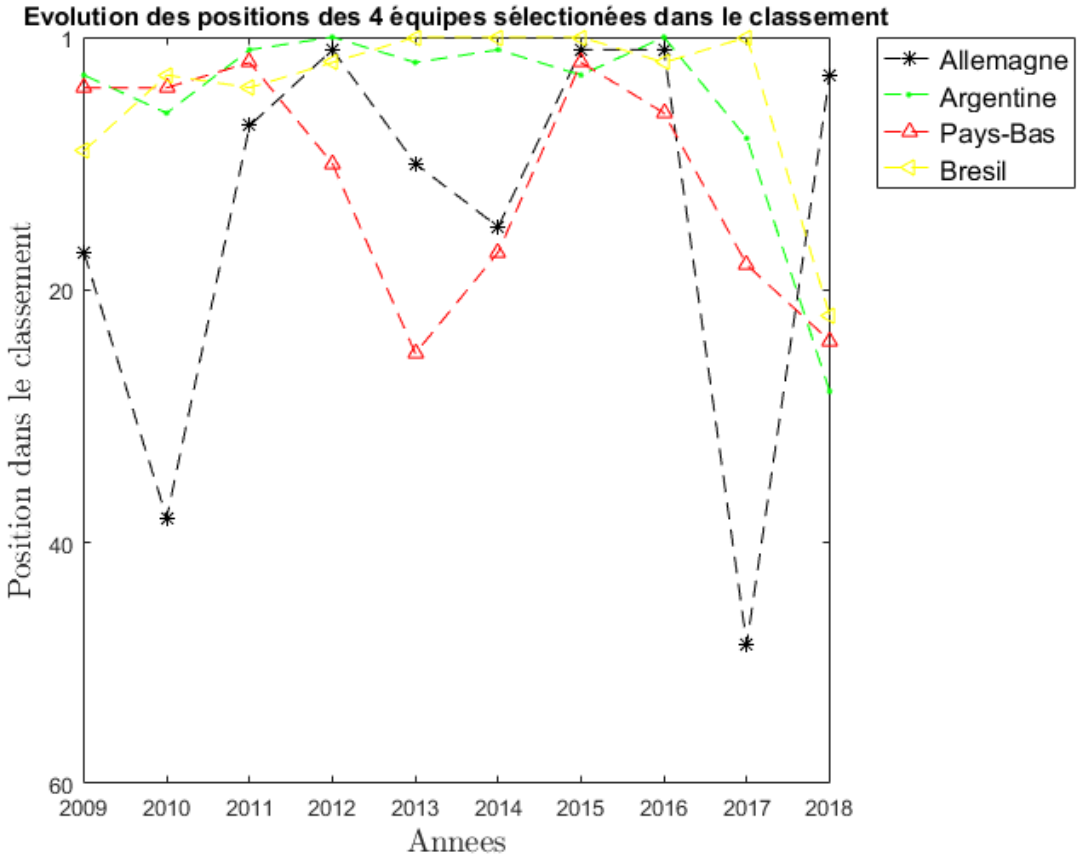

FIGURE 4.15 – Évolution du classement de 4 équipes pendant 10 ans. Classement Pagerank. Coupe du monde 2014.

Coupe du monde 2018 pour le classement Pagerank.

En ce qui concerne la dernière coupe du monde, nous rappelons malheureusement que les quatre équipes ayant atteints les demi-finales sont dans l'ordre, la France, la Croatie, la Belgique et l'Argentine. Nous nous sommes permis une légère modification du graphique sur la Figure 4.16, en effet, plutôt que de considérer les années entre 2009 et 2018, nous avons calculé une année supplémentaire afin d'obtenir l'information du classement à la date du 1^{er} janvier 2019. Deux raisons à cela, une plus scientifique et une autre plus personnelle, nous avons jugé intéressant de considérer l'année pendant laquelle ont eu lieu les matchs de la coupe afin de voir les modifications de classement de cette année. La raison personnelle est que nous étions frustré de voir la Belgique si « mal » positionnée alors que nous avons fini troisième de la compétition, prendre en compte l'année 2018 nous a permis, tout comme à l'Angleterre, de venir arracher quelques places supplémentaires dans le classement.

Si nous analysons maintenant la Figure 4.16, on se rend dans un premier temps compte que notre classement était à même de prédire la victoire de la France avec une telle certitude qu'elle a conservé la première place du classement deux années consécutives. Les performances de la France pendant l'année 2017 lui ont valu de voir son classement s'améliorer et atteindre la première place pour le classement de janvier 2018. L'équipe française était donc d'après notre classement, la mieux placée pour conquérir la coupe.

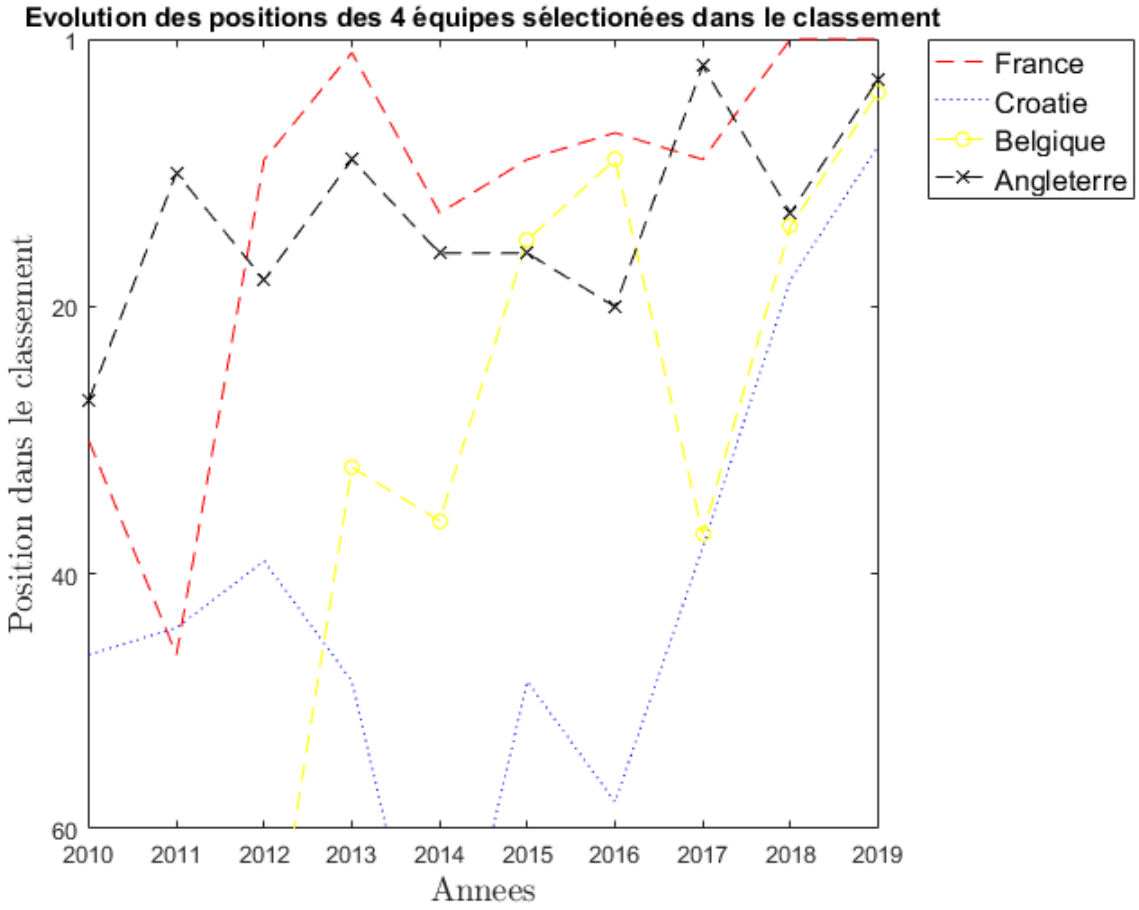

FIGURE 4.16 – Évolution du classement de 4 équipes pendant 10 ans. Classement Pagerank. Coupe du monde 2018.

En ce qui concerne les trois autres équipes, on remarque que les placements respectifs de la Croatie et de la Belgique se comportent de la même manière. Ceux-ci subissent une nette amélioration à partir de l'année 2017 qui se poursuit avec l'année 2018 et leurs performances remarquables lors de la coupe. L'Angleterre quant à elle voit son classement chuter légèrement entre 2017 et 2018 [9] ce qui peut laisser présager un petit passage creux pour l'équipe avant que celle-ci ne se reprenne lors de l'année 2018 afin de maintenir sa domination sur l'équipe belge d'une place (supposition que nous avions faite d'instinct en observant la proximité des deux points sur l'image et qui a été vérifiée numériquement).

9. Etant donné son excellente position en 2017, il était difficilement faisable de faire mieux si ce n'est en dépassant la France.

4.3.2 Le classement utilisant le facteur h

Passons maintenant à notre deuxième classement, celui utilisant pour sa construction le facteur h. La Figure 4.17 permet de réaliser la même analyse que la Figure 4.13. Il semble cependant encore plus difficile de constater une quelconque information. Une remarque que nous pouvons déjà émettre à ce stade est que ce classement semble beaucoup plus « désordonné » que le précédent. On y retrouve de très fortes variations des différentes positions pour presque toutes les équipes.

Dans le même but de clarté que pour les analyses sur le classement Pagerank, nous allons subdiviser cette analyse en 3 parties et voir si les constatations précédentes se confirment. L'aspect général du graphique ne semble cependant pas très engageant et il est fort à parier que même si les résultats obtenus sont de temps à autres corrects, ce classement ne sera pas aussi fiable que le précédent.

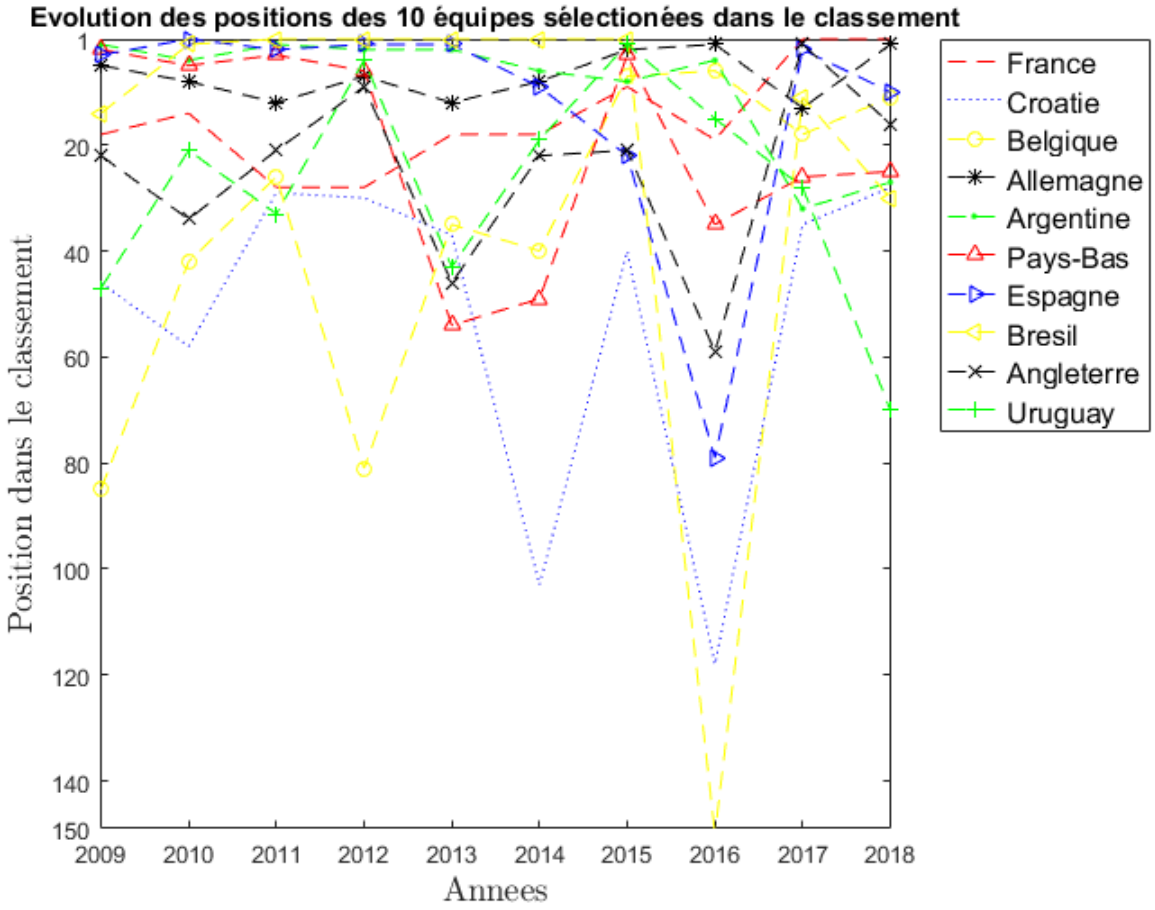

FIGURE 4.17 – Évolution du classement de 10 équipes pendant 10 ans. Classement facteur h. Coupe du monde 2010 à 2018

Coupe du monde 2010 pour le classement facteur h.

Les rappels d'équipes ne seront plus faits dans les trois sections qui suivent, les équipes sont les mêmes que pour les analyses précédentes et sont, en cas de besoin, présentes dans les légendes de chaque figure. Intéressons nous, comme nous l'avons fait pour l'autre classement, à la courbe bleue de la Figure 4.18. On remarque que tout comme précédemment, l'équipe espagnole affiche des résultats plus que satisfaisants aux alentours de l'année 2010, confirmant donc sa première place. Les Pays-Bas gardent également la même allure que précédemment mais on remarque cependant que l'Allemagne se porte beaucoup mieux que dans le classement utilisant le Pagerank. L'équipe uruguayenne présente pour sa part une remontée intéressante l'année précédant la coupe du monde mais n'atteint malgré tout que le top 20 du classement. La perte de niveau des équipes hollandaise et espagnole se fait également ressentir mais de manière beaucoup plus abrupte sur la Figure 4.18 que sur la Figure 4.14. Nous voyons donc à la suite de cette analyse que les résultats ne sont pas mauvais, mais ils semblent cependant beaucoup plus sensibles aux changements que ne l'était le classement précédent. Continuons notre analyse avec la coupe du monde 2014 afin de vérifier si cette « sensibilité » se confirme.

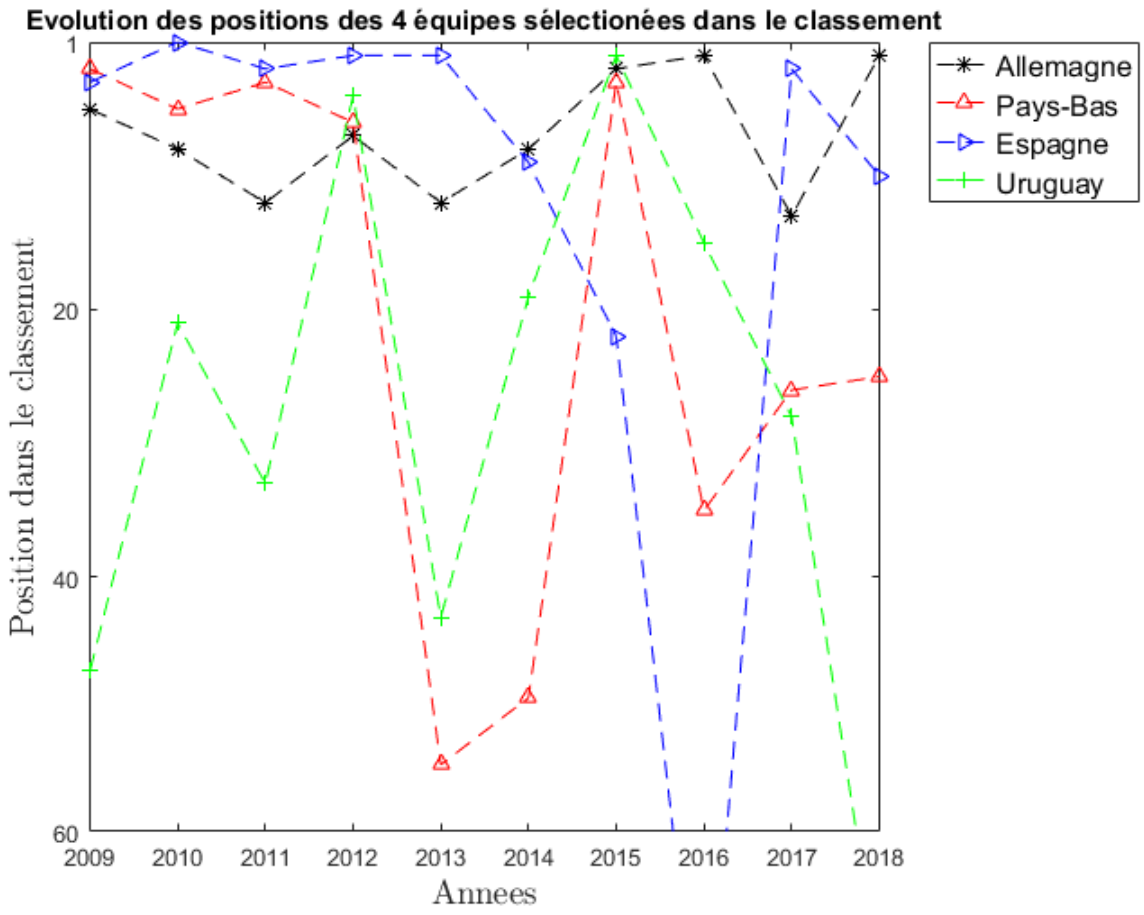

FIGURE 4.18 – Évolution du classement de 4 équipes pendant 10 ans. Classement facteur h. Coupe du monde 2010.

Coupe du monde 2014 pour le classement facteur h.

Sur la Figure 4.19, la prédominance de l'équipe du Brésil est encore plus prononcée que pour le classement précédent. L'équipe allemande n'est cependant pas mal classée durant l'année de la coupe du monde et avait en plus entamé une remontée dans le classement à partir de 2013. Cette remontée a donc l'air plus intéressante et prévisible dans ce cas ci car, rappelons le, dans le classement utilisant le Pagerank, l'Allemagne voyait son classement chuter même entre 2013 et 2014. L'Argentine, qui finit deuxième de la compétition, avait semble-t-il entamé une chute dans les positions du classement et continue cette légère descente même lors de l'année de la coupe, ce qui est un peu étrange.

Une chose est certaine, si nous avions dû prédire un vainqueur pour cette coupe du monde, le Brésil aurait été sélectionné directement. Nous pouvons avant de passer à la suite, réaliser la même constatation concernant les Pays-Bas que dans la Figure 4.15, en effet, à partir de l'année 2013, le classement de l'équipe commence à s'améliorer avant de subir une nette propulsion due certainement aux bons résultats réalisés durant la coupe du monde. On peut toutefois se mettre d'accord sur le fait que cette figure, est toujours aussi sensible aux changements comme le prouve le maximum local atteint par la courbe des Pays-Bas en janvier 2015.

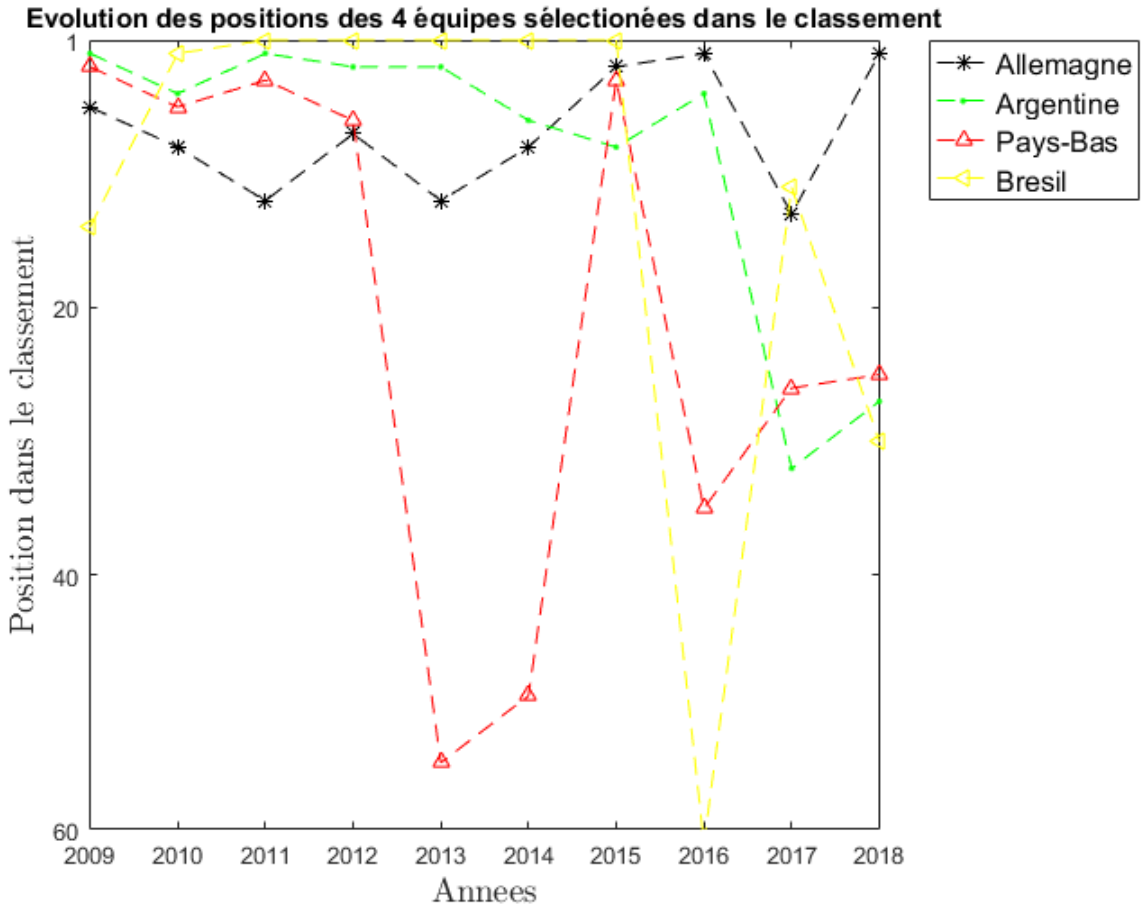

FIGURE 4.19 – Évolution du classement de 4 équipes pendant 10 ans. Classement facteur h. Coupe du monde 2014.

Coupe du monde 2018 pour le classement facteur h.

Pour cette dernière analyse, nous avons réalisé la même petite modification que pour le classement du Pagerank de manière à avoir l'information concernant l'année de la coupe 2018. Sur la Figure 4.20, le caractère abrupt du classement ne se fait pas prier pour se faire ressentir. Les classements sont très peu constants et semblent pour la plupart, changer de comportement années après années. Les informations qui nous intéressent sont cependant assez satisfaisantes étant donné que nous retrouvons toujours la France au top du classement (bien qu'elle retombe à la deuxième place en 2019) et que la Belgique ainsi que l'Angleterre sont toujours aux coudes à coudes comme c'était le cas sur la Figure 4.16. La Croatie reste comme précédemment légèrement en retard par rapport à la Belgique et l'Angleterre.

Le reste de l'allure concernant les pertes de place dans le classement de l'Angleterre sont toujours visibles et sont cette fois-ci précédées d'une chute similaire pour l'équipe belge.

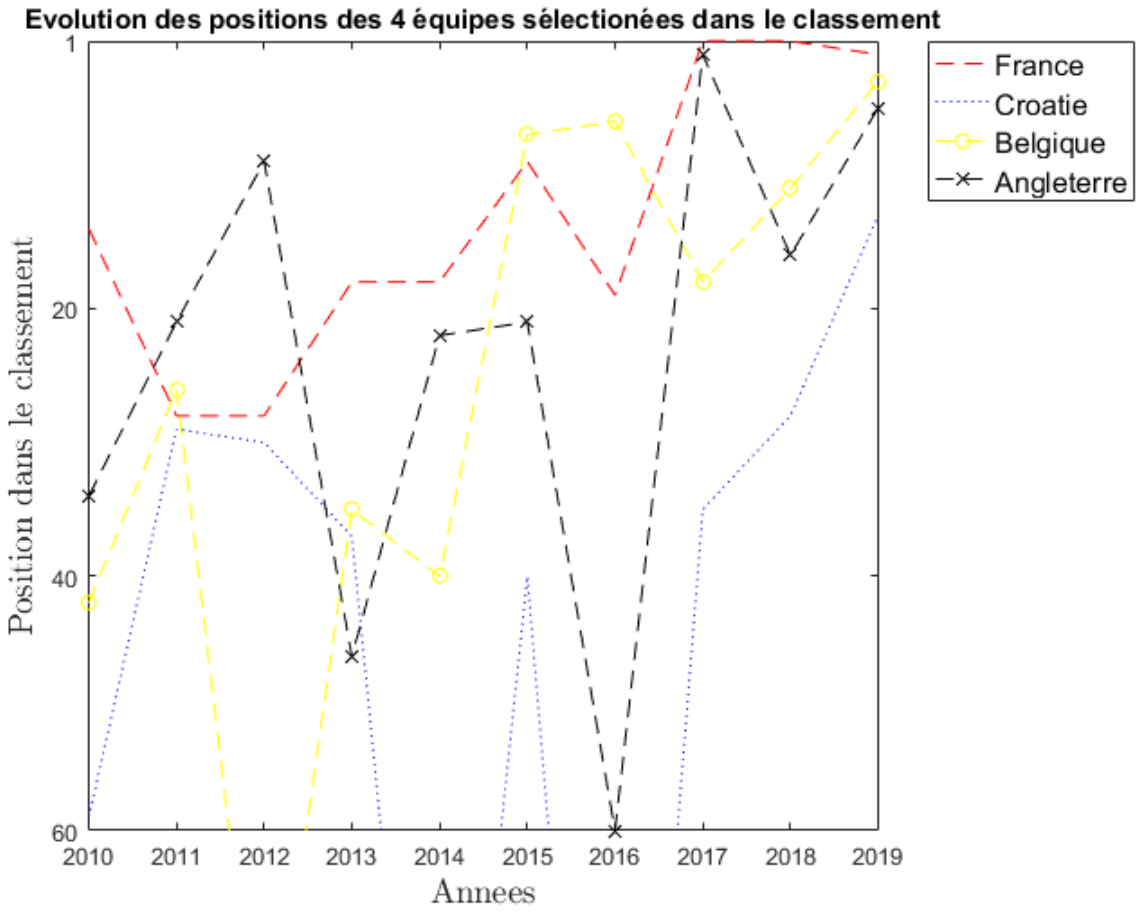

FIGURE 4.20 – Évolution du classement de 4 équipes pendant 10 ans. Classement facteur h. Coupe du monde 2018.

4.3.3 Le classement FIFA

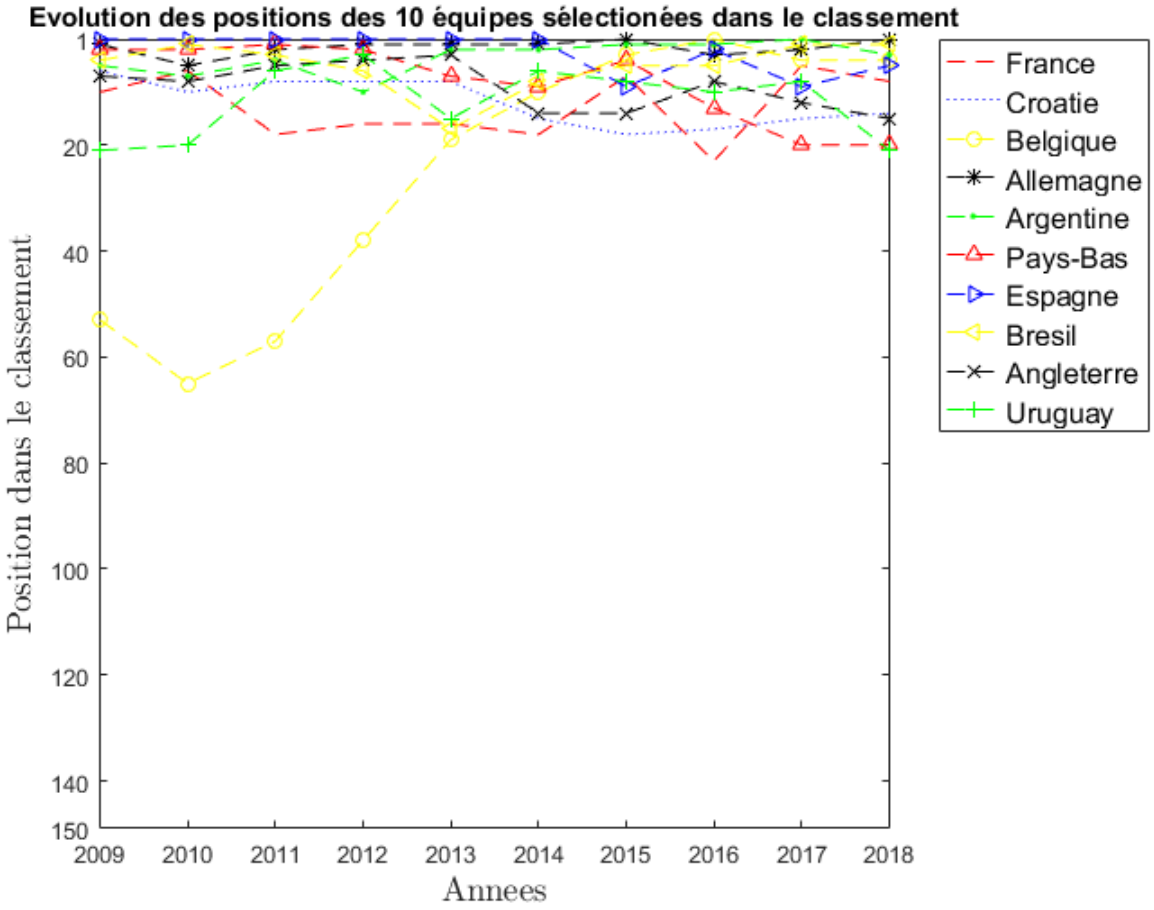

FIGURE 4.21 – Évolution du classement de 10 équipes pendant 10 ans. Classement FIFA. Coupe du monde 2010 à 2018

Nous allons maintenant nous intéresser un instant à l'évolution du classement FIFA sur la même période que nos deux classements. Afin de réaliser cette analyse, nous avons considéré lorsque c'était possible le classement du mois de janvier pour chacune des années entre 2009 et 2018. Lorsqu'aucun classement n'était établi pour le mois de janvier, nous nous sommes arrangés pour prendre celui du mois de décembre de l'année précédente.

Sur la Figure 4.21, on remarque immédiatement un élément que nous n'avions pas au préalable. Cet élément n'est autre que la constance au sein des résultats, en effet, nous remarquons très peu d'irrégularités et celles-ci, lorsqu'elles surviennent, sont minimes. Nous pouvons observer que 9 des 10 équipes restent en quasi-permanence dans le top 20 du classement FIFA. La seule équipe à ne pas respecter cela est la Belgique qui n'entre dans le top 20 qu'à partir de l'année 2013.

Analysons maintenant les graphes les uns après les autres comme nous l'avons fait précédemment afin de pouvoir réaliser une analyse plus fine de ce qu'il se passe pour les différentes équipes en fonction des coupes du monde.

Coupe du monde 2010 pour le classement FIFA.

On remarque directement sur la Figure 4.22 que l'Espagne, championne du monde en 2010, semble effectivement mériter son titre étant donné qu'elle est première du classement pendant 6 années de suite [10]. On remarque ensuite sa chute dans le classement lors de l'année 2015, cette chute à similaire à celle que nous observons dans nos deux classements à la différence qu'elle ne perdure pas dans le temps. Nous remarquons que dès l'année 2016, l'Espagne reprend des places dans le classement et semble osciller dans le top 10.

En ce qui concerne les trois autres équipes, on constante qu'elles sont correctement ordonnées par rapport aux résultats de la coupe du monde. L'Allemagne et les Pays-Bas se situent dans le top du classement ce qui n'est pas le cas de l'Uruguay. Ce dernier réussit cependant à gagner plus d'une dizaine de places dans le classement après ces bons résultats lors de la coupe du monde 2010.

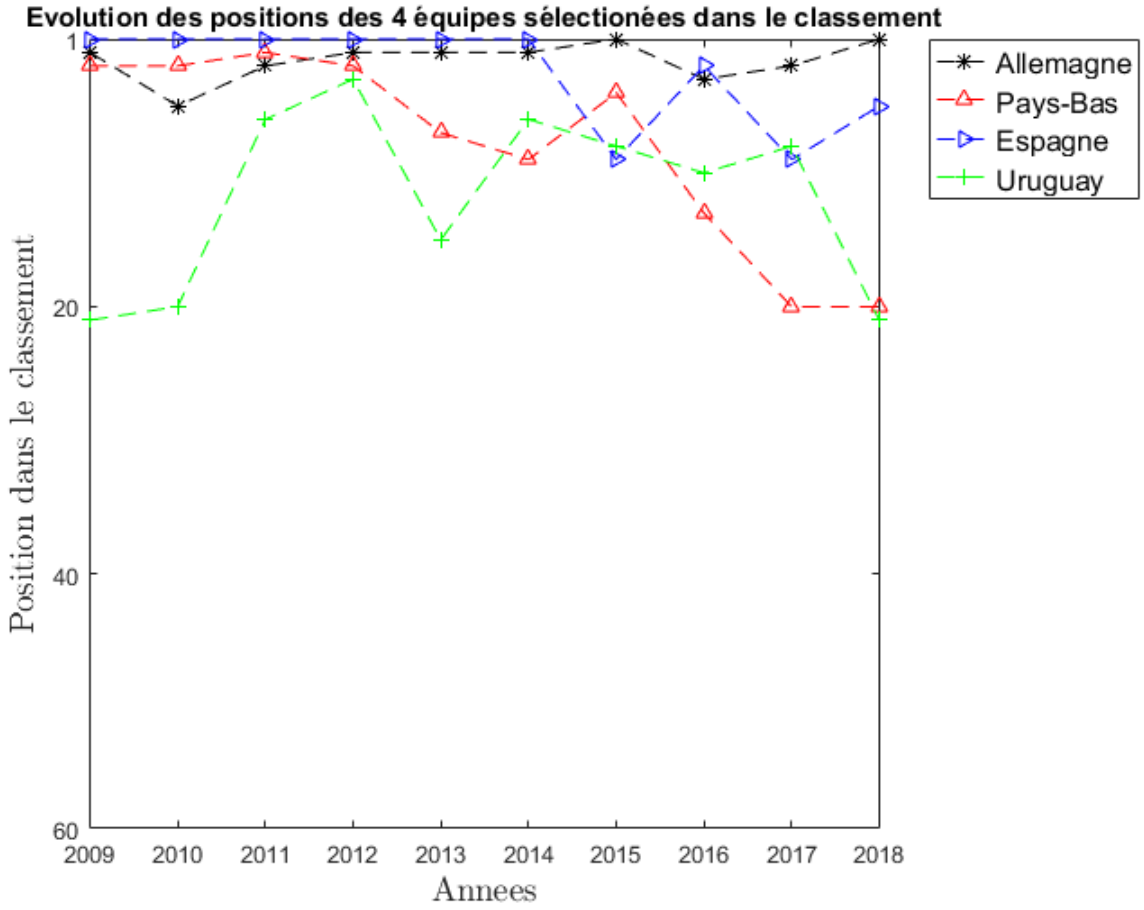

FIGURE 4.22 – Évolution du classement de 4 équipes pendant 10 ans. Classement FIFA. Coupe du monde 2010.

10. Elle n'est pas forcément restée première pendant les 6 ans tous les mois mais l'était en tout cas pour les classements annuels considérés.

Coupe du monde 2014 pour le classement FIFA.

Nous retrouvons sur la Figure 4.23, l'Allemagne et les Pays-Bas qui étaient présents sur la figure précédente et nous pouvons remarquer que l'Allemagne est dans le top des équipes lors de la période de la coupe du monde 2014, ce qui n'est pas le cas des Pays-Bas. On remarque effectivement la « même » allure que dans nos deux classements, en effet les Pays-Bas subissent une perte de places entre l'année 2012 et l'année 2014, regagnent des places en 2015 et reprennent ensuite le chemin du top 20 à partir de 2016. Cette allure était également visible dans nos classements à l'énorme différence que les variations que nous apercevons ici sont beaucoup moins directes. La transition se fait d'une manière beaucoup plus douce ce qui renforce notre idée principale qui était que le classement FIFA est très peu sensible aux variations.

On peut également remarquer sur la Figure 4.23, qu'à l'approche de la coupe du monde, l'Argentine remonte dans le classement et talonne l'Allemagne. Le Brésil cependant, subit une légère perte de places avant la coupe du monde et semble se reprendre en mains à partir de 2013, date à partir de laquelle leur classement ne va faire que croître.

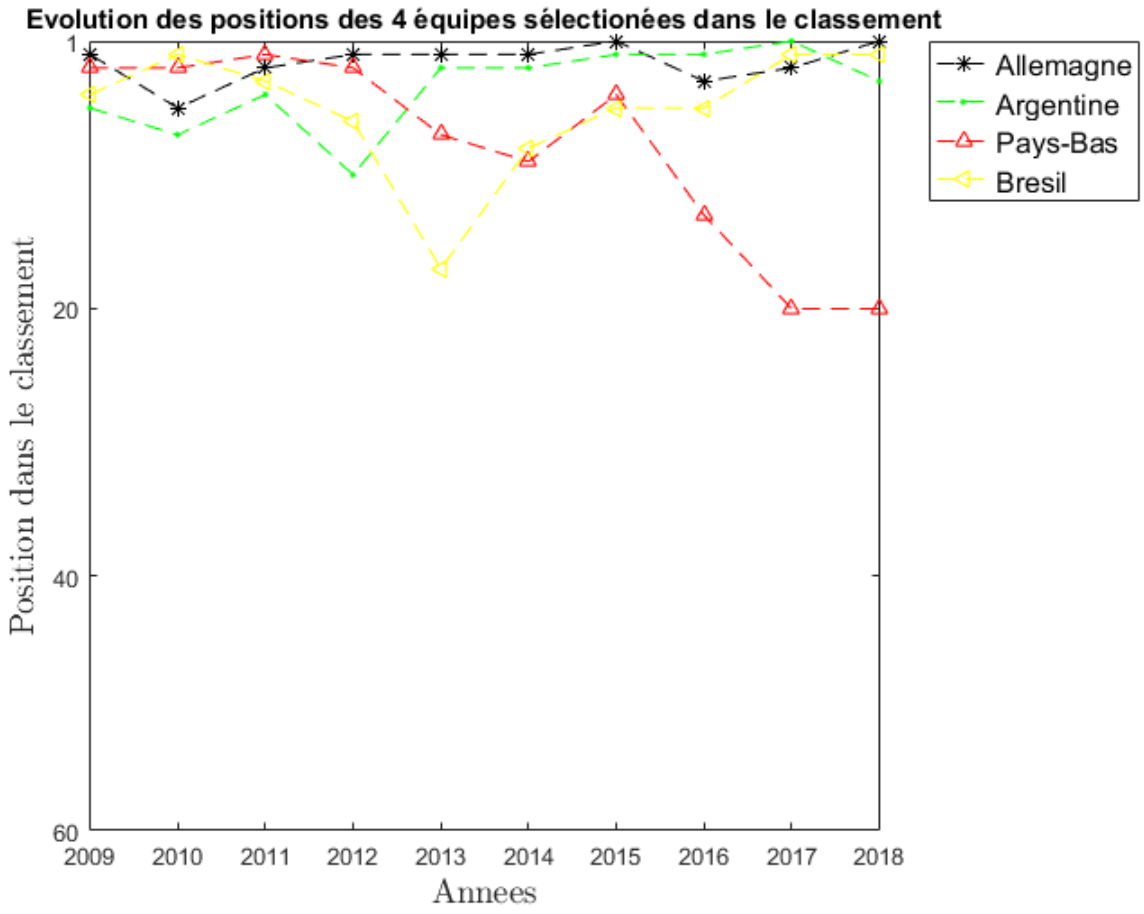

FIGURE 4.23 – Évolution du classement de 4 équipes pendant 10 ans. Classement FIFA. Coupe du monde 2014.

Coupe du monde 2018 pour le classement FIFA.

Nous arrivons pour finir cette analyse, à la Figure 4.24 sur laquelle on voit nettement l'évolution du classement de l'équipe belge. Le code permettant d'extraire le classement FIFA directement du site ne fonctionnant plus, nous n'avons malheureusement pas pu extraire les données relatives à l'année 2019. Une vérification scolaire du classement sur le site officiel de la FIFA [5], nous a permis de vérifier qu'en janvier 2019, la Belgique était passée première du classement.

On remarque que l'évolution du classement des équipes concernées par cette coupe du monde est beaucoup moins constante que ne l'était les classements des équipes précédentes. On peut en effet se rendre compte des nombreuses variations que subissent les classements de la France, dont le classement est moins bon que celui de la Belgique en 2018, de l'Angleterre et de la Croatie sont plus importants que pour les autres équipes.

En ce qui concerne la Belgique, on peut constater une légère perte de place à partir de 2016 qui pourrait attester de sa perte de forme minime à l'approche de la coupe du monde. La Croatie quand à elle suit une légère progression dans le classement à partir de 2015.

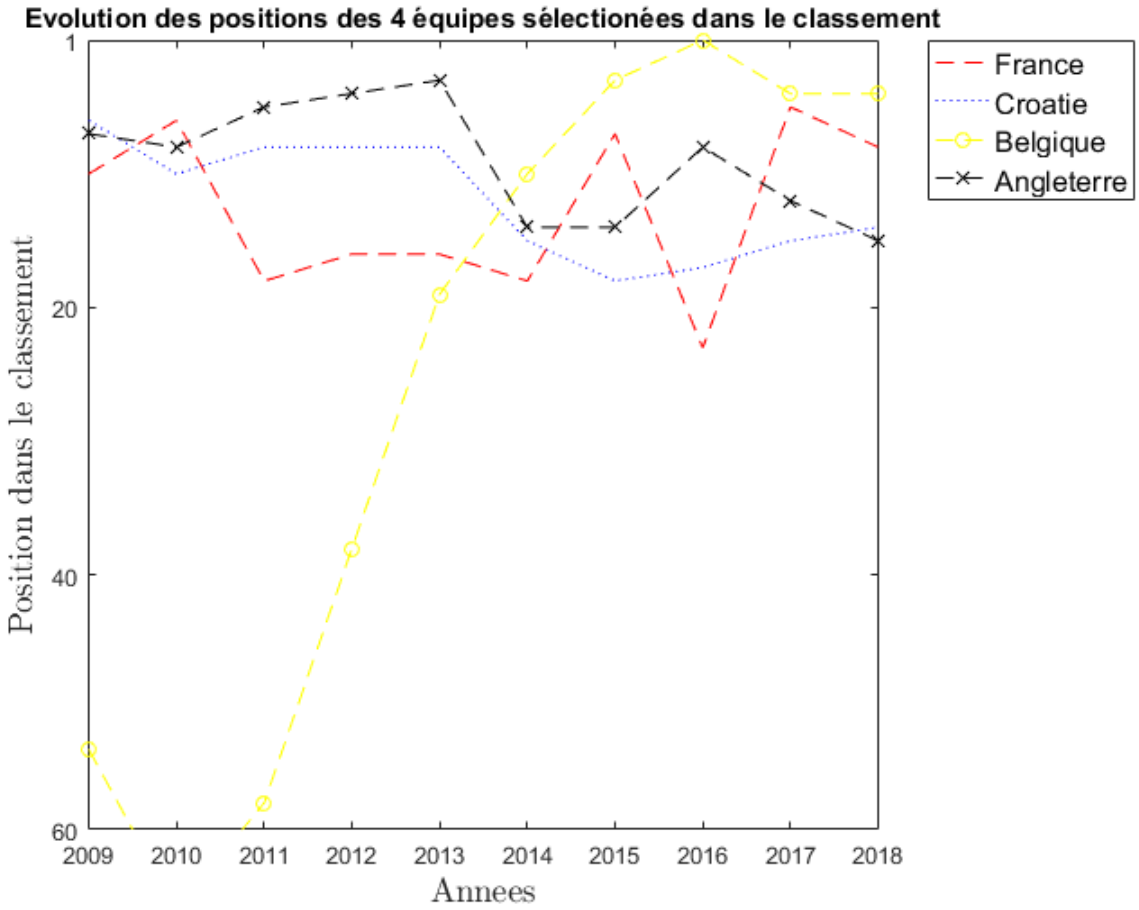

FIGURE 4.24 – Évolution du classement de 4 équipes pendant 10 ans. Classement FIFA. Coupe du monde 2018.

4.3.4 Conclusion relative à l'évolution temporelle

Suite à cette partie d'analyse, on se rend finalement bien compte que nos deux classements sont suffisamment satisfaisants, en tout cas dans une moindre mesure. Nous sommes cependant face à un dilemme qui concerne l'efficacité de ces deux classements, en effet, nous avons des résultats similaires lors de nos différentes constatations mais nous avons d'un côté le classement utilisant l'algorithme du Pagerank qui semble être constant et peu capricieux. De l'autre côté, nous avons le classement construit simplement sur la base du facteur h qui semble parfois plus efficace que le précédent et d'autres fois moins efficace en fonction des petites modifications. Ce dernier semble être beaucoup plus sensible à des petites perturbations.

Nous pouvons également nous mettre d'accord sur le fait que nos classements, bien que satisfaisants, sont moins efficaces que celui de la FIFA. Celui possède vraisemblablement de meilleurs résultats globaux que nos deux classements. Lors de certaines années, nos classements ont réussi à coller de manière plus formelle à la réalité pour l'une ou l'autre équipe mais était en général moins performant que celui utilisé par la FIFA.

Nous poursuivrons nos différents ressentis à la toute fin de ce mémoire mais nous allons avant tout passer au chapitre des perspectives qui permettront peut-être à d'autres de poursuivre ce qui a été fait jusqu'ici.

Chapitre 5

Perspectives

Ce chapitre permettra d'énoncer les pistes d'améliorations auxquelles nous avons pensé durant ce mémoire. Elles permettraient potentiellement d'améliorer ce qui a été fait jusqu'à présent en affinant certains paramètres de l'analyse. La majorité de ces modifications sont des éléments très minimes qui n'ont pas été pris en compte car les décisions initiales n'ont pas été dans ce sens et qu'il était difficile de tout refaire pour les corriger.

La première manipulation qui pourrait être réalisée serait de permettre l'utilisation de durées de recherche différentes que des années. En effet si l'on prend notre exemple de classement, nous étions fatalement obligés de constituer nos classements sur des années complètes et donc considérer les classements FIFA de décembre de la même année ou janvier de l'année suivante. Ce genre de contrainte nous a empêché par exemple de générer des classements à la date du début de la coupe du monde. En comparant nos classements avec les résultats des différentes coupes du monde, nous avions donc toujours en permanence 6 mois de décalage entre notre classement et le lancement de la coupe du monde. Cette manipulation n'est pas bien compliquée à mettre en place en soit, mais l'entièreté du code général avait été construit sur des périodes annuelles, il aurait donc fallu entièrement le remanier après avoir déjà généré des résultats. Nous avons donc fait le choix de conserver ces analyses tout en sachant que les résultats en seraient moins bons.

Un deuxième élément qu'il pourrait être intéressant de modifier concerne les différentes confédérations de football. Dans notre cas, nous avons décidé de travailler avec presque toutes les équipes, les « seules » équipes exclues étaient celles qui ne faisaient pas partie de la composante connexe principale. Il était en effet inutile d'intégrer deux équipes dans le classement général si celles-ci n'avaient joué qu'entre elles. Le problème qui s'est cependant posé au début de nos analyses, étaient que certaines équipes étaient à des places qui ne leur correspondait dans le classement, tout cela dû à de nombreux matchs gagnés dans leur confédération (autre que l'UEFA).
Certaines équipes, comme Guernesey, se sont retrouvés dans le classement alors que cela est totalement incohérent car si nous prenons l'exemple de l'équipe que nous venons de citer, elle n'est membre ni de la FIFA, ni de l'UEFA. Son classement au sein même d'un classement mondial n'aurait donc aucun intérêt dû à l'absence de match officiel l'impliquant.
Nos analyses n'ont globalement pas été impactées par ces équipes « spéciales » mais peut-être seraient-elles un problème dans le cas d'analyses plus poussées.

Le troisième point qu'il faudrait absolument modifier concerne le classement numéro 1 qui utilisait simplement la centralité de degré, en effet, cette partie des codes est une des premières à avoir été réalisée et nous nous sommes rendu compte il y a peu qu'il aurait été préférable de diviser chaque degré par le nombre de matchs joués afin d'éviter que certaines équipes n'aient trop d'importance acquise uniquement en ayant réalisé un grand nombre de matchs.

Nous avions également réfléchi à une amélioration du code qui consisterait à rendre celui-ci plus « dynamique ». En effet, dans l'implémentation de nos différents codes, le nombre d'années que nous imposons est fixe et le classement est par conséquent très rigide, il pourrait être envisagé de donner une certaine durée au code pour que celui-ci génère plusieurs classements tout au long de l'exécution et que ces différents classements n'aient pas une durée fixe et identique. Nous avions prévu dans notre code général, une option qui permet de générer un classement sur base, non pas d'une durée, mais d'un nombre de matchs fixes, l'idée serait donc de reprendre cette idée mais de la rendre automatique afin d'avoir des classements plus étendus en cas de période creuse et des classements plus fréquents en cas de grosse quantité de matchs.

L'idée précédente est évidemment faite dans l'idée de générer le classement de manière différente afin de vérifier si des variations d'implémentations permettraient d'améliorer les résultats, ce n'est cependant pas la seule. Nous avions également pensé à réaliser des classements dans lesquels chaque équipe aurait réalisé le même nombre de matchs afin qu'elles soient chacune sur un même pied d'égalité. Cette réflexion a rapidement été abandonnée car nous avons jugé la réalisation trop complexe et surtout trop casse-tête. Si nous avions pu sélectionner nous-mêmes les matchs à comptabiliser dans le classement, cela aurait encore pu être possible, mais en respectant le suivi des différents matchs, il est quasi impossible de réaliser cette idée avec l'organisation actuelle des différents matchs officiels et amicaux. Si par la suite, une telle modification est possible, cela serait certainement très intéressant à analyser car nous aurions alors un classement dans lequel chaque équipe a eu le même « nombre » d'occasions pour briller.

Nous avions également pensé à une autre analyse qui pouvait être réalisée sur la section 4.3. Celle-ci consistait à calculer les indices de Jaccard des différents classements utilisés pour représenter les Figures 4.13, 4.17 et 4.21. Le but n'était pas de calculer l'indice de Jaccard sur les classements globaux mais bien de créer des mini-classements uniquement pour les 10 équipes concernées. Il nous aurait ensuite fallu calculer les différents indices de Jaccard en comparant chaque fois deux classements entre eux. Cette manipulation n'a pas été faite, bien que le code de la fonction Jaccard ait été implémenté, car cela nécessitait énormément de manipulations et nous étions obligés de recréer les classements afin d'éliminer les éléments parasites. Etant donné que cela nous aurait simplement apporté l'information concernant les taux de similitudes entre ces classements, nous avons préféré nous concentré sur d'autres choses.

Nous aimerions avant de proposer une dernière amélioration, émettre deux critiques relatives à ce qui a été évoqué pendant nos analyses. La première concerne un abus de langage que nous avons ouvertement utilisé à de nombreuses reprises. Nous avons énormément parlé de prédiction en ce qui concerne ce mémoire. Si nous avions voulu être plus rigoureux et moins enthousiaste, nous aurions en réalité dû parler uniquement de constatations. En effet,

nous avons vérifié si les différents classements que nous générions était en adéquation avec les résultats observés lors des différentes coupes du monde. Un modèle prédictif se baserait sur l'ensemble des données antérieures et nous fournirait par chaque équipe, une probabilité que celle-ci se retrouve dans le trio de tête. Il serait donc possible, afin de continuer ce mémoire, de constituer un modèle prédictif sur base de nos analyses précédentes.

La deuxième critique concerne la génération de nos matrices aléatoires. Nous n'avions au départ pas prévu de générer ces matrices et nous nous sommes rendu compte de leurs utilités très tard dans le mémoire. Nous avons donc implémenté celles-ci d'une manière simpliste sans tenir compte de paramètres importants. La génération actuelle de nos matrices permet de créer des matrices de la même taille que nos matrices d'analyses, c'est à dire le nombre total d'équipe. Les nombres aléatoires qui servent d'éléments à cette matrice sont simplement générés avec la fonction rand de matlab, le nombre obtenu pour chaque élément a été multiplié par un coefficient pour pouvoir être utilisé dans nos différents réseaux. Il aurait fallu, pour que ces matrices permettent de conclure de réelles informations concernant nos véritables matrices, qu'elles soient associées à un véritable réseau de matchs de foot aléatoire. Cela signifie qu'il nous faut deux éléments importants, premièrement que le nombre d'équipes soit le même que pour un classement réel, ce qui est le cas, et deuxièmement, que le nombre de matchs joués soit également identique à celui d'un classement réel. Ce dernier point n'est pas vérifié dans nos matrices et limite donc l'utilisation de nos analyses utilisant l'aléatoire. Il faudrait donc, dans le but d'affiner l'ensemble des analyses, être capable de générer des matrices aléatoires qui respectent ces critères.

Une dernière chose que nous pourrions préciser dans ce chapitre relève plus de la remarque générale que des perspectives. Tous les différents classements que nous avons générés ont nécessité l'utilisation de certains paramètres, le facteur des matchs amicaux à 0.2, le paramètre « followprobability » du Pagerank qui a été laissé par défaut, etc. Tous ces paramètres peuvent évidemment être modifiés en fonction de la volonté de l'utilisateur, ne voulant pas participer à un jeu d'essai erreur infini, nous n'avons évidemment pas testé de nombreuses variations de paramètres. Il est donc possible en modifiant certaines valeurs d'un côté ou de l'autre des codes, que les résultats obtenus soient encore plus probants. Nous sommes cependant persuadés au vu de ces résultats que les potentielles améliorations seraient minimes et nous ne sommes pas certains que cela en vaille donc la peine à moins de vraiment vouloir proposer ce modèle à la FIFA.

Conclusion

Au terme de ce mémoire, nous aimerions revenir sur l'ensemble de ce qui a été fait jusqu'à présent, sur les différents résultats que nous avons obtenus et également sur le choix que nous ferions si nous étions à la place de la FIFA.

Après avoir introduit la théorie des réseaux et s'être attardés sur différentes centralités, nous nous sommes quand même concentrés sur les concepts qui nous semblaient être les plus adéquats pour ce travail. Nous avons détourné l'utilisation littéraire du facteur h pour en faire un facteur sportif et nous avons utilisé une modification de l'indice de Jaccard afin de pouvoir le manier correctement. L'utilisation du Pagerank qui est pourtant très connu dans le milieu du web nous a semblé évidente lorsque nous avons commencé à réfléchir à une manière de classer nos différentes équipes.

Tout cela nous a permis d'obtenir des résultats qui au final nous satisfont. Nous étions dans un premier temps assez peu rassurés en voyant les résultats car nous avancions à tâtons et bien que les codes soient opérationnels, nous nous rendions compte à première vue que les classements obtenus n'avaient pas grand chose à voir avec celui de la FIFA qui nous a, pendant tout un temps, servi de modèle. Au fur et à mesure de l'avancement du mémoire, nous étions dépités car nous avions l'impression de foncer dans une impasse et de produire du travail parfois « inutile ». Entendons nous bien, aucun travail n'est inutile en soit, mais nous étions attristés à l'idée de ne pas pouvoir fournir de résultats probants.
En effet, les taux de similitudes entre nos différents classements et celui de la FIFA n'étaient pas très élevés et cela nous paraissait étrange. Nous ne désirions pas réaliser une copie conforme du classement FIFA étant donné que nous avions pour but de l'améliorer, mais obtenir des classements aussi peu ressemblants nous étonnait car certaines informations du classement FIFA n'était, malgré tout, pas négatives.
Les analyses concernant la fiabilité de nos classements ne nous ont pas non plus apporté de soulagement car nous nous sommes dans un premier temps attardés sur les chiffres. Nous étions fixé sur le fait que le classement FIFA trouvait le trio gagnant de la coupe du monde dans ses 15 premières places alors tous nos classements faisaient pire ! Ce n'est que lorsque nous nous sommes intéressés à un autre critère assez différent que nous avons obtenu les premiers vrais bons résultats. Lorsque nous avons eu l'idée de comptabiliser combien d'équipes des 32 qualifiées pour une coupe du monde se retrouvaient dans le top 32 de nos classements, nous nous sommes rendu compte que nos résultats étaient en fait similaires à ceux de la FIFA. Nous retrouvions dans certains de nos classements, le même nombre d'équipe dans certains de nos classements que ce que nous en retrouvions dans celui de la FIFA.
Nous avons cependant pris le parti à ce moment là d'oublier un de nos trois classements, le plus simple, car celui-ci était fort semblable à celui utilisant l'algorithme du Pagerank et réalisait de moins bons résultats. Nous avons donc poursuivi en utilisant uniquement le classement que nous appelons Pagerank et le classement utilisant le facteur h. Une de nos

idées était de voir si nos classements étaient capables de prédire un quelconque résultat d'une manière plus avérée que le classement réalisé par la FIFA. Nous nous sommes donc intéressés aux graphiques que vous pouvez voir dans la section 4.3 et nous nous sommes attardés sur des critères différents. Nous regardions quelle équipe atteignait la première place du classement et également comment les différentes équipes se comportaient en fonction des matchs qu'elles avaient joués. Ces résultats nous ont agréablement surpris lorsque nous avons eu complété les différentes analyses. Effectivement, la majorité de nos analyses et de nos graphes semblaient correspondre avec les résultats observés lors des différentes coupes du monde, les résultats n'étaient évidemment pas parfaits et nos analyses étaient « orientée par le résultat » [1], mais cela ne nous empêche pas de dire que nos classements sont loin d'être aussi mauvais que nous le pensions au départ. Il faut d'ailleurs oser le dire, certains de nos classements ont réalisé de meilleures prévisions que celles réalisées par la FIFA à l'époque. Le fait de voir la France première de plusieurs de nos classements avant la coupe du monde qui la sacrera championne n'est, à notre avis, pas un hasard. Il faut cependant être honnête, les résultats, lorsqu'ils sont bons, le sont ponctuellement et lorsque l'on tient compte de l'ensemble du classement et non pas d'une seule équipe, le classement FIFA obtient de biens meilleurs résultats que nos classements.

Malgré cela, notre avis sur les différents classements est donc positif. Nous trouvons cependant que le classement utilisant l'algorithme du Pagerank est beaucoup plus adapté dans cette situation que ne l'est celui utilisant le facteur h. Bien que les résultats obtenus avec ce dernier soient bons également, les derniers graphes dénotent une certaine fragilité quant aux résultats. Les positions des équipes dans le classement semblent en effet pouvoir bouger de manière très abrupte en faisant ainsi subir à l'équipe, de lourdes pertes de places ou au contraire d'incroyables gains de places. Un classement aussi volatile serait très contraignant à suivre et c'est là que réside une des forces selon nous du classement FIFA, si modification il y a, elle se réalise sur un temps assez long et ne modifie pas totalement la structure du classement. Notre classement utilisant le Pagerank semble être un compromis entre les deux. Les variations de positions sont légères et les résultats prédits sont en général suffisamment corrects que pour tenir la route en cas d'utilisation. Un tel classement assure donc une meilleure robustesse du classement et permet à priori de prédire de manière assez objective les résultats des différentes rencontres. En effet, rappelons ici que lors de nos analyses de graphes, nous étions à même de relever les moments ou une équipe semblait revenir sur le devant de la scène. De telles informations sont beaucoup plus intéressantes qu'une simple position dans un classement qui ne tient pas compte de la dynamique antérieure. Si nous devions prédire le résultat d'un match, nous analyserions donc notre deuxième classement afin de voir si les équipes concernées sont dans une bonne ou une mauvaise période, un peu comme le font les différents journalistes sportifs à la télé lors des émissions d'avant match. Attention toutefois que ce que nous appelons prédictions, sont en réalité des constatations sur des faits qui se sont déjà déroulés. Il serait grandement intéressant d'être capable de fournir une probabilité que telle ou telle équipe soit dans le trio gagnant d'une coupe du monde. Nous ne nous limiterions cependant pas à la seule étude de nos classements et nous croiserions ces résultats avec ceux du classement FIFA, la robustesse et la fiabilité de ce dernier compenserait les défauts de notre classement. Pour appuyer notre explication, nous insistons sur le fait qu'une analyse combinée de nos résultats avec ceux de la FIFA aurait

1. Nous signifions par cela que nos analyses sont réalisées en ayant à l'avance l'information du résultat réel et de l'issue des différentes coupes. Le but de ce mémoire est cependant de prédire ces résultats et non pas de s'y adapter.

potentiellement permis de prédire une victoire française et par conséquent éviter une désillusion à un bon nombre de supporters belges.

Nous tenons simplement pour finir ce mémoire, à émettre une réflexion très humble et finir sur une touche un peu « philosophique ». Le classement FIFA existe depuis 1993 et a déjà été remanié, étant donné les enjeux économiques et sociaux que représentent le football, nous espérons fortement que les différents créateurs du classement FIFA se sont intéressés à son fonctionnement et ne l'ont pas cependant créé de toutes pièces. Il est donc rassurant que notre classement, réfléchi, construit et analysé en l'espace de deux ans ne soit pas totalement apte à remplacer le classement FIFA. De plus, il ne faut également pas oublier un facteur primordial dans l'étude des différents classements et même celui de la FIFA. Le football reste un sport pratiqué par des êtres humains, il ne s'agit pas d'un jeu contrôlé par une intelligence artificielle qui soit parfaite et ne commettant pas d'erreur. Il faut donc absolument prendre en compte que, peu importe ce que disent les différents classements, il est possible pour des raisons humaines, médicales ou autres, que le match se solve par un résultat inattendu. C'est aussi cela qui fait la beauté du football au grand dam de ceux qui font de ce sport un terrain de jeu économique autant que sportif.

Nous espérons que ce mémoire aura pu vous apprendre des choses et vous ouvrir à certaines possibilités au niveau de l'utilisation des réseaux. Quant à nous, nous nous tenons prêt à analyser les nouvelles données lorsque celles-ci seront disponibles pour espérer prédire une victoire des Belges lors de la prochaine coupe du monde.

Bibliographie

[1] Barabási, Albert-László, *Network Science*, Presses Universitaires de Cambridge, 2016.

[2] Bibliothèque du Centre universitaire de santé de McGill, *Quel est votre impact ? En savoir plus sur l'indice h*, 2015

[3] Chikhi Nacim Fateh, *Calcul de centralité et identification de structures de communautés dans les graphes de documents.* Interface homme-machine [cs.HC]. Université Paul Sabatier - Toulouse III, 2010.

[4] Collombat, Benoît, *Foot, argent sale et paradis fiscaux : pourquoi le prix des transferts explose*, franceinfo, 1 mai 2020.

[5] Fifa, *Classement Masculin*, disponible sur https ://fr.fifa.com/fifa-world-ranking/ranking-table/men/, 29 avril 2020.

[6] FIFA, *Men's Ranking Procedure*, www.fifa.com/fifa-world-ranking/procedure/men, 10 mai 2020.

[7] Fifa, *Révision du Classement mondial FIFA/Coca-Cola*, https ://resources.fifa.com/image/upload/revision-of-the-fifa-coca-cola-world-ranking.pdf ?cloudid=em15xdivyscqfwhqxcqt, 10 mai 2020.

[8] Freeman, Linton C., *Centrality in Social Networks Conceptual Clarification*, Social Networks, 1, 215-239, 1979.

[9] Gargiulo, F., Caen, A., Lambiotte, R. et al. *The classical origin of modern mathematics*, EPJ Data Sci, 5, 26, https ://doi.org/10.1140/epjds/s13688-016-0088-y, 2016.

[10] Jeong, Hawoong, et al., *The large-scale organization of metabolic networks*, Nature 407, 651–654, 2000.

[11] Jürisoo Mart, *Internationnal football results from 1872 to 2020*, base de données, Kaggle, https ://www.kaggle.com/martj42/international-football-results-from-1872-to-2017, consulté le 18 mai 2020

[12] Kunegis, Jérôme, *Théorie des graphes*, Syllabus, Université de Namur, 2017.

[13] Lemaire, Sophie, *Introduction aux chaînes de Markov*, Syllabus, Université Paris-Saclay, 2012.

[14] Maquinay, Franck, et al., Rankspirit, *Le Pagerank : Qu'est-ce que c'est ?*, https ://www.rankspirit.com/pagerank, 2 mai 2020.

[15] Newman, Mark, *Networks*, Presses Universitaires d'Oxford, Université du Michigan, seconde édition, 2018.

[16] Sen, Parongama, et al., *Small-world properties of the Indian railway network.*, Phys.Rev.E67, 2003.

[17] Vega-Redondo, Fernando, *Complex Social Networks*, Presses Universitaires de Cambridge, Université d'Alicante et Université de l'Essex, 2007.

[18] Wikipédia, *Indice h*, mis à jour le 9 mai 2020, 9 mai 2020.

Annexes

Synthèse des différentes similitudes entre les classements.

	Classement numéro 1	Classement numéro 2	Classement numéro 3	Classement FIFA
Classement numéro 1	/	14.5235%	4.2764%	3.8021%
Classement numéro 2	14.5235%	/	3.837%	3.6789%
Classement numéro 3	4.2764%	3.837%	/	4.5756%
Classement FIFA	3.8021%	3.6789%	4.5756%	/

FIGURE 1 – Tableau comparatif des résultats pour une matrice aléatoire et 20 équipes.

Synthèse des différentes similitudes entre les classements.

	Classement numéro 1	Classement numéro 2	Classement numéro 3	Classement FIFA
Classement numéro 1	/	28.4404%	25%	11.1111%
Classement numéro 2	28.4404%	/	28.8344%	27.2727%
Classement numéro 3	25%	28.8344%	/	13.8211%
Classement FIFA	11.1111%	27.2727%	13.8211%	/

FIGURE 2 – Tableau comparatif des résultats pour un laps de temps de un an et 20 équipes.

Synthèse des différentes similitudes entre les classements.

	Classement numéro 1	Classement numéro 2	Classement numéro 3	Classement FIFA
Classement numéro 1	/	36.3636%	22.449%	32.9114%
Classement numéro 2	36.3636%	/	35.4839%	32.0755%
Classement numéro 3	22.449%	35.4839%	/	22.449%
Classement FIFA	32.9114%	32.0755%	22.449%	/

FIGURE 3 – Tableau comparatif des résultats pour un laps de temps de deux ans et 20 équipes.

Synthèse des différentes similitudes entre les classements.

	Classement numéro 1	Classement numéro 2	Classement numéro 3	Classement FIFA
Classement numéro 1	/	32.0755%	30.031%	18.9802%
Classement numéro 2	32.0755%	/	36.8078%	24.2604%
Classement numéro 3	30.031%	36.8078%	/	28.0488%
Classement FIFA	18.9802%	24.2604%	28.0488%	/

FIGURE 4 – Tableau comparatif des résultats pour un laps de temps de trois ans et 20 équipes.

Synthèse des différentes similitudes entre les classements.

	Classement numéro 1	Classement numéro 2	Classement numéro 3	Classement FIFA
Classement numéro 1	/	26.1261%	26.1261%	26.1261%
Classement numéro 2	26.1261%	/	35.4839%	35.4839%
Classement numéro 3	26.1261%	35.4839%	/	38.6139%
Classement FIFA	26.1261%	35.4839%	38.6139%	/

FIGURE 5 – Tableau comparatif des résultats pour un laps de temps de quatre ans et 20 équipes.