Euro 2020 - prédictions à l'aide d'un modèle de statistiques bayésiennes

Melchior Prugniaud, Julien Foënet

23 janvier 2020

Introduction

Le but de ce projet est de faire des prédictions pour le prochain grand tournois de football, l'Euro 2020. Pour cela nous avons suivis la méthodologie proposée par F.Louzada, A.H.Suzuki, L.Salazar, A.Ara et J.Leite, qui est un modèle de statistiques Bayesienne basé sur le rang FIFA des équipes (un nombre de point est attribué par la fédération internationale de football, la FIFA, à chaque équipe selon son niveau actuel) et une distribution a priori donnée par des avis d'experts sur le score des futures matches. Nous avons décidé de remplacer dans notre modèle les avis d'experts par des données historiques car ces avis étaient trop difficiles à récolter et que nous voulions utiliser une approche différente. Avec cette méthode nous pouvons calculer analytiquement les probabilités de victoire/égalité/défaite de chaque équipe ainsi que les probabilités de gagner le tournois en faisant un grand nombre de simulations.

Modelisation

Nombre de but dans un match

Considérons un match équipe A vs équipe B, on modélise le nombre de but marqué par l'équipe A dans ce match par une loi de Poisson dont le paramètre dépend de la supériorité de l'équipe A sur l'équipe B (exprimé par le rang FIFA des deux équipe, R_A et R_B) :

$$X_{AB} \sim Poisson\left(\lambda_A \frac{R_A}{R_B}\right) \tag{1}$$

On peut faire de même avec l'équipe B pour modéliser le nombre de but qu'elle marquerait contre l'équipe A dans ce match en intervertissant R_A et R_B et en remplaçant λ_A par λ_B. Ici, λ_A peut être vue comme le nombre moyen de but que marque l'équipe A contre une équipe de même niveau qu'elle.

Distribution a priori

A la place des avis d'expert sur les futurs matchs, on a décidé de prendre en compte des données historiques dans notre distribution a priori. Pour les m matchs du tournois que l'on veut prédire où l'équipe A joue contre les adversaire $(OA_i)_{i=1...m}$, nous prenons les scores (uniquement le nombre de but marqué par l'équipe A) de S précédents matchs de l'équipe A. Pour les données correspondantes à chaque match i nous prenons autant que possible des matchs où l'équipe A est opposée à l'adversaire OA_i ou à des équipes de même rang FIFA que l'adversaire OA_i, dans une limite de 4 ans et uniquement dans des grandes compétitions si possible (sinon nous prenons des matchs amicaux). Nous prendrons $S = 3$ matchs historiques pour chaque match à prédire, ce qui semble être un bon nombre compte tenu du peu

de données disponibles pour certaines équipes. Par exemple, pour la phase de groupe (4 équipes dans chaque groupes, chaque équipe joue donc $m = 3$ matchs), on aura, $\forall i = 1...m$:

$$A \text{ vs } OA_i => y_{i1}, ..., y_{iS}$$

où y_{is} représente le nombre de but marqué par l'équipe A dans le match s que l'on a pris comme donnée pour prédire le match i.

Avant d'inclure ces données historiques, nous commençons avec une distribution a priori non informative ($Gamma(\delta_0, \beta_0)$) pour λ_A qui représente l'incertitude que l'on a sur ce paramètre :

$$\pi_0(\lambda_A) \propto \lambda_A^{\delta_0 - 1} \exp(-\beta_0 \lambda_A)$$

Puis on incorpore les données historiques $D_0 = \{y_{11}, ..., y_{mS}\}$ pour obtenir l'a posteriori $\lambda_A|D_0$, que nous utiliserons plus tard comme nouvelle a priori informative. On considère, comme dans notre première hypothèse, que $y_{ij} \sim Poisson\left(\lambda_A \frac{R_A}{RO_{Ai}}\right)$ où y_{ij} est le nombre de but marqué par l'équipe A contre OA_i (ou une équipe de niveau équivalent) dans le match historique j. On a alors d'après la formule de Bayes :

$$\begin{aligned}
\pi(\lambda_A|D_0) &= L(D_0; \lambda_A)\pi_0(\lambda_A) \\
&\propto \prod_{i=1}^{m}\prod_{j=1}^{S}\left[\left(\lambda_A \frac{R_A}{R_{OA_i}}\right)^{y_{ij}} e^{-\lambda_A \frac{R_A}{R_{OA_i}}}\right] \lambda_A^{\delta_0 - 1} e^{-\beta_0 \lambda_A} \\
&\propto \lambda_A^{\sum_{i=1}^{m}\sum_{j=1}^{S} y_{ij} + \delta_0 - 1} exp\left(-\lambda_A \left(S\sum_{i=1}^{m}\frac{R_A}{R_{OA_i}} + \beta_0\right)\right) \qquad (2) \\
\Rightarrow \lambda_A|D_0 &\sim Gamma\left(\sum_{i=1}^{m}\sum_{j=1}^{S} y_{ij} + \delta_0, S\sum_{i=1}^{m}\frac{R_A}{R_{OA_i}} + \beta_0\right)
\end{aligned}$$

Dans la prochaine sections, $\lambda_A|D_0$ sera simplifié par λ_A car nous prendrons cette distribution qui inclus l'information des données historiques comme notre nouvelle distribution a priori informative.

Distribution prédictive

Dans cette section nous aurons besoin de distinguer deux cas : l'un (i) est lorsque nous nous plaçons avant le début du tournois, où nous n'avons alors à disposition que les données historiques, puisqu'aucun match du tournois n'a été joué. L'autre (ii) est lorsque des matchs du tournois on déjà été joués, où nous avons alors à disposition de l'information supplémentaire (ex : après la phase de groupe, avant la demi-finale, etc...).

i) Avant le début du tournois

proposition 1 :

$$Si \;\; X|\lambda \sim Poisson(\lambda c) \;\; et \;\; \lambda \sim Gamma(\alpha, \beta)$$
$$alors \;\; X \sim BinomialNegative\left(\alpha, \frac{\beta}{\beta + c}\right)$$

preuve : voir annexe.

Avec la proposition ci-dessus, et à partir de (1) et (2) on obtient la prédictive :

$$X_{AB} \sim BN\left(\sum_{i=1}^{m}\sum_{j=1}^{S} y_{ij} + \delta_0, \frac{S\sum_{i=1}^{m}\frac{R_A}{R_{OA_i}} + \beta_0}{S\sum_{i=1}^{m}\frac{R_A}{R_{OA_i}} + \beta_0 + \frac{R_A}{R_B}}\right) \tag{3}$$

ii) Après que des matchs aient été joués

Supposons que l'équipe A ait joué k matchs dans le tournois (ex : A est en quart de final, et a donc joué déjà trois matchs en phase de groupe et un en huitième de final) contre les adversaires $C_1, ..., C_k$. On a alors les données $D = x_A^1, ..., x_A^k$ où x_A^i est le score de l'équipe A contre C_i.

Puisque $x_A^i \sim Poisson\left(\lambda_A \frac{R_A}{R_{C_i}}\right)$, on écrit la vraisemblance :

$$\begin{aligned} L(\lambda_A; D) &\propto \prod_{l=1}^{k} e^{-\lambda_A \frac{R_A}{R_{C_l}}} \left(\lambda_A \frac{R_A}{R_{C_l}}\right)^{x_A^l} \\ &\propto exp\left(-\lambda_A \sum_{l=1}^{k} \frac{R_A}{R_{C_l}}\right) \lambda_A^{\sum_{l=1}^{k} x_A^l} \end{aligned} \tag{4}$$

La distribution a posteriori est maintenant :

$$\begin{aligned} \pi(\lambda_A|D) &= \overbrace{L(\lambda_A; D)}^{(4)}\overbrace{\pi(\lambda_A)}^{(2)} \\ &\propto e^{-\lambda_A \sum_{l=1}^{k}\frac{R_A}{R_{C_l}}} \lambda_A^{\sum_{l=1}^{k} x_A^l} e^{-\lambda_A\left(S\sum_{i=1}^{m}\frac{R_A}{R_{OA_i}}+\beta_0\right)} \lambda_A^{\sum_{i=1}^{m}\sum_{j=1}^{S} y_{ij}+\delta_0-1} \\ &\propto \lambda_A^{\sum_{i=1}^{m}\sum_{j=1}^{S} y_{ij}+\delta_0+\sum_{l=1}^{k} x_A^l-1} e^{-\lambda_A\left[S\sum_{i=1}^{m}\frac{R_A}{R_{OA_i}}+\sum_{l=1}^{k}\frac{R_A}{R_{C_l}}+\beta_0\right]} \\ \Rightarrow \lambda_A|D &\sim Gamma\left(\sum_{i=1}^{m}\sum_{j=1}^{S} y_{ij} + \sum_{l=1}^{k} x_A^l + \delta_0, S\sum_{i=1}^{m}\frac{R_A}{R_{OA_i}} + \sum_{l=1}^{k}\frac{R_A}{R_{C_l}} + \beta_0\right) \end{aligned} \tag{5}$$

Grâce à la proposition 1 on obtient la prédictive :

$$X_{AB} \sim BN\left(\sum_{i=1}^{m}\sum_{j=1}^{S} y_{ij} + \sum_{l=1}^{k} x_A^l + \delta_0, \frac{S\sum_{i=1}^{m}\frac{R_A}{R_{OA_i}} + \sum_{l=1}^{k}\frac{R_A}{R_{C_l}} + \beta_0}{S\sum_{i=1}^{m}\frac{R_A}{R_{OA_i}} + \sum_{l=1}^{k}\frac{R_A}{R_{C_l}} + \beta_0 + \frac{R_A}{R_B}}\right)$$

Méthode

Données

Les données des matchs historiques ont été récupérés sur le jeu de données "results.csv" provenant du site Kaggle, qui donne les scores de tous les matchs internationaux depuis le début des années 1900, avec évidemment le nom des équipes mais aussi le contexte du match (Euro, coupe du monde, copa america, match amical, etc...). Grâce à ce dataset nous avons pu extraire, pour chaque équipe et chaque match de

phase de groupe joué par cette équipe, trois matchs historiques contre un adversaire similaire en terme de rang FIFA. Par exemple pour la coupe du monde 2014, le Brésil jouait en phase de groupe contre la Croatie (903 poins FIFA), le Mexique (882 points) et le Cameroon (558 points). Pour prédire le match Brésil-Croatie nous avons pris les scores de matchs précédents du Brésil contre les Pays-Bas (981 points), le Chili (1026 points), et la France (913 points). Nous avons fait de même pour les deux autres matchs :

FIgure 1 Données utilisées pour les matchs du Brésil en phase de groupe (coupe du monde 2014)

TeamA	TeamB	RatingA	RatingB	GoalsA	GoalsB
Brazil	Croatia	1242	903		
Brazil	Netherlands	1242	981	1	2
Brazil	Chile	1242	1026	3	0
Brazil	France	1242	913	3	0
Brazil	Mexico	1242	882		
Brazil	Ivory Coast	1242	809	3	1
Brazil	Mexico	1242	882	2	0
Brazil	Russia	1242	893	1	1
Brazil	Cameroon	1242	558		
Brazil	Paraguay	1242	575	2	2
Brazil	Japan	1242	626	3	0
Brazil	Venezuela	1242	675	0	0

Analyses et simulations

Avec notre modèle, nous pouvons calculer plusieurs choses :

- Pour un match spécifique, nous pouvons calculer analytiquement grâce à la distribution prédictive la probabilité de victoire/égalité/défaite (voir annexe).
- Pour le tournois entier, nous pouvons simuler le score de chacun des matchs, et ainsi obtenir les résultats d'un tournois complet, avec le vainqueur, les demi-finalistes, etc... Si l'on répète cette simulation un grand nombre de fois, nous obtenons alors les probabilités pour chaque équipe de gagner le tournois, d'arriver en finale, de sortir des groupes, etc...

Pour juger de la qualité de notre modèle sur les différents tournois précédents pris en compte et ainsi s'assurer que nos prédictions pour l'Euro 2020 sont bonnes, nous avons utilisé deux mesures : la mesure de DeFinetti, et le pourcentage de bonne prédiction. Nous avons comparés nos résultats à ceux que l'on trouverait si l'on s'appuyait uniquement sur le rang FIFA, c'est à dire à une méthode qui donnerait vainqueur l'équipe avec le plus haut rang FIFA pour chacun des matchs.

La mesure de DeFinetti pour un tournois est la moyenne des distances de DeFinetti sur chacun des matchs, où la distance de DeFinetti est : $||pred - true||^2$ avec $pred = (\mathbb{P}_{win}, \mathbb{P}_{draw}, \mathbb{P}_{loss})$ le vecteur de prédiction des probabilités du résultat du match, et $true = (win, draw, loss)$ le vecteur représentant le résultat du match. Par exemple si notre prédiction pour le résultat d'un match est (0.3, 0.6, 0.1), c'est à dire que l'équipe A a 30% de chance de gagner, 60% de chance de faire un nul et 10% de chance de perdre, et que le vrai résultat est une égalité, c'est à dire $true = (0, 1, 0)$, la distance de DeFinetti sera : $(0.3 - 0)^2 + (0.6 - 1)^2 + (0.1 - 0)^2 = 0.26$.

Le but est évidemment d'avoir la mesure de DeFinetti la plus faible possible et le pourcentage de bonne prédiction le plus haut possible sur l'ensemble d'un tournois.

Résultats

Dans cette section tous les calculs ont été fait avec les paramètres $\beta_0 = 0.01, \delta_0 = 0.5$. Ces paramètres ont été choisit de manière à obtenir les meilleurs résultats.

Coupe du monde 2014

a) Résultats analytiques

Le tableau ci-dessous montre, pour les matchs à partir des huitièmes de finale (les matchs de groupes sont en annexe), les probabilités de victoire/égalité/défaite calculées avec la distribution prédictive pour chaque match (pred1), ainsi que la distance de DeFinetti par rapport au vrai résultat (deFi1). Nous avons également fait les même calculs pour la méthode "FIFA ranking" qui donnerait vainqueur pour chaque match l'équipe ayant le plus haut rang FIFA (pred2, deFi2).

Figure 2 Prédictions pour chaque match à partir des huitièmes de finale - coupe du monde 2014

teamA	score	teamB	pred1	deFi1	good1	pred2	deFi2	good2
Brazil	1-1	Chile	(0.47 0.25 0.3)	0.88	0	(1 0 0)	2	0
Colombia	2-0	Uruguay	(0.43 0.26 0.37)	0.535	1	(0 0 1)	2	0
France	2-0	Nigeria	(0.56 0.23 0.3)	0.338	1	(1 0 0)	0	1
Germany	2-1	Algeria	(0.81 0.12 0.07)	0.054	1	(1 0 0)	0	1
Netherlands	2-1	Mexico	(0.66 0.17 0.22)	0.197	1	(1 0 0)	0	1
Costa Rica	1-1	Greece	(0.16 0.32 0.56)	0.797	0	(0 0 1)	2	0
Argentina	1-0	Switzerland	(0.45 0.26 0.28)	0.443	1	(1 0 0)	0	1
Belgium	2-1	United States	(0.22 0.26 0.52)	0.953	0	(1 0 0)	0	1
Brazil	2-1	Colombia	(0.38 0.25 0.33)	0.555	1	(1 0 0)	0	1
France	0-1	Germany	(0.11 0.15 0.76)	0.092	1	(0 0 1)	0	1
Netherlands	0-0	Costa Rica	(0.83 0.12 0.09)	1.477	0	(1 0 0)	2	0
Argentina	1-0	Belgium	(0.55 0.25 0.21)	0.314	1	(1 0 0)	0	1
Brazil	1-7	Germany	(0.25 0.22 0.52)	0.339	1	(0 0 1)	0	1
Netherlands	0-0	Argentina	(0.39 0.22 0.42)	0.936	0	(0 0 1)	2	0
Germany	1-0	Argentina	(0.57 0.2 0.24)	0.288	1	(1 0 0)	0	1

Nous obtenons pour l'ensemble du tournois une mesure de DeFinetti de 0.52 et un pourcentage de bonne prédiction de 68% , alors que la méthode "FIFA ranking" donne une mesure de DeFinetti plus forte (0.70) et un pourcentage de bonne prédiction plus faible (65%). Ceci montre que notre modèle est plus à même de prédire l'issue d'un match, probablement parce que les données historiques apportent de l'information supplémentaire sur la dynamique d'une équipe et sur sa capacité à marquer (ou au contraire a ne pas marquer) contre des adversaire de rang différent.

b) Tournois - Simulations

Dans cette partie qui vise a obtenir les probabilités pour chaque équipe de gagner le tournoi, nous avons effectués $n = 5000$ simulations du tournois entier, en utilisant pour chaque match la distribution binomiale pour simuler le nombre de but des deux équipes. Les deux meilleures équipes de chaque groupes continuent en huitième de finale (pour l'Euro il y'a une différence : il y a six groupes de quatre équipes, il faut donc en plus prendre les quatre meilleurs troisième de tous les groupes).

Figure 3 probabilités d'atteindre différents stades du tournois - coupe du monde 2014

Team	winner	final	semi final	quarter final	eigth final
Germany	0.37	0.5002	0.6864	0.8434	0.9374
Argentina	0.1108	0.2302	0.4038	0.7152	0.9456
Spain	0.0838	0.1662	0.3106	0.4802	0.7396
Netherlands	0.0772	0.1594	0.2998	0.4806	0.738
United States	0.0642	0.1576	0.302	0.499	0.6758
Brazil	0.0622	0.127	0.2758	0.4572	0.894
Uruguay	0.0392	0.099	0.2118	0.4398	0.755
Colombia	0.0376	0.0852	0.1918	0.4122	0.7436
Chile	0.0264	0.0638	0.1516	0.2786	0.5132
Switzerland	0.0248	0.0676	0.1576	0.4204	0.7592
England	0.0176	0.0504	0.1254	0.3152	0.616
Italy	0.0162	0.0468	0.124	0.2982	0.5766
Portugal	0.0128	0.0348	0.0812	0.1732	0.2922
Belgium	0.01	0.0312	0.0822	0.2098	0.719
Greece	0.0094	0.029	0.0792	0.2128	0.478
Mexico	0.0084	0.031	0.0872	0.1948	0.6234
Bosnia and Herzegovina	0.0056	0.024	0.0772	0.291	0.606
Ivory Coast	0.0048	0.0176	0.056	0.1716	0.415
Croatia	0.0032	0.0112	0.0388	0.0942	0.3736
Japan	0.003	0.0118	0.042	0.1394	0.3634
Ecuador	0.0026	0.0116	0.0382	0.1534	0.4378
France	0.0026	0.0108	0.0402	0.1632	0.4582
Russia	0.0026	0.0086	0.0284	0.0962	0.4732
Algeria	0.0018	0.0094	0.0326	0.1032	0.5362
Honduras	0.0016	0.0062	0.0238	0.1056	0.3448
Ghana	0.0008	0.0032	0.0154	0.0462	0.0946
Nigeria	0.0006	0.0038	0.0206	0.1194	0.328
Costa Rica	0.0002	0.0004	0.0024	0.0108	0.0524
Australia	0	0	0.0002	0.0014	0.0092
Cameroon	0	0.0006	0.0034	0.013	0.109
Iran	0	0.0004	0.0046	0.0318	0.1204
South Korea	0	0.001	0.0058	0.029	0.2716

Pour la coupe du monde 2014, nous obtenons de très bons résultats. Nous présentons ci-dessus un tableau avec la probabilité de chaque équipe de gagner le tournois, d'atteindre la finale, les demi-finales, etc... Lors de ce tournoi, l'Allemagne est sortit vainqueur et a battu l'Argentine en finale, qui sont dans cet ordre en première et deuxième position dans notre tableau. Les Pays-Bas ainsi que le Brésil étaient également présents en demi-finale de cette compétition, et ils font partie d'après notre modèle des six équipes les plus à mêmes d'accéder aux demi-finales.

Coupe du monde 2018 & Euro 2016

Nous donnerons ici seulement quelques informations sur ces deux tournois, mais nous avons fait les même calculs et simulations que pour la coupe du monde 2014. Pour l'Euro 2016, nous obtenons une mesure de DeFinetti de 0.62 et un pourcentage de bonne prédiction de 45% (toujours mieux que ceux de la méthode "FIFA ranking" qui sont de 1.17 et 41% Le Portugal, qui a gagné le tournois, était en 3ème position dans notre tableau et la France qui est arrivé en finale était en 5ème position. Pour la coupe du monde 2018, nous obtenons une mesure de DeFinetti de 0.60 et un pourcentage de bonne prédiction de 55% (toujours mieux que ceux de la méthode "FIFA ranking" qui sont respectivement de 1.01 et 49%). La France, qui a gagné le tournois, était en 3ème position dans notre tableau, mais la Croatie qui est

arrivé en finale était seulement en 13ème position pour accéder à la finale. En effet la Croatie était une "surprise" comme il en arrive parfois dans le football, qui reste un sport extrêmement imprévisible.

Euro 2020 - prédictions

Nous nous intéressons désormais à la prédiction des résultats pour l'Euro 2020, compétition qui se déroulera dans plusieurs pays européen en Juin 2020. Pour le moment, la majorité des pays qualifiés pour cette compétition est connu mais il reste trois pays qui doivent encore disputer des play-offs pour jouer la compétition. Etant donné que les pays restant ne sont pas des grandes nations du football, nous avons choisis de qualifier les pays avec la plus grande chance de gagner le tournoi de qualification d'après l'avis de journalistes sportifs. Ces pays sont : l'Islande, la Macédoine du Nord et la Slovaquie.

Nous allons dans un premier temps montrer les chances de qualifications de chaque équipe dans son groupe pour les huitièmes de finale, c'est à dire les probabilités de finir 1er, 2ème, 3ème ou 4ème de son groupe. Il est important de noter que certains groupes sont beaucoup plus indécis que les autres notamment le groupe F qui contient le champion du monde en titre (la France), le champion du monde 2014 (l'Allemagne), le champion d'Europe en titre (le Portugal) et l'Islande. Les journalistes et experts s'accordent pour dire qu'il s'agit d'un groupe de la "mort". Le tableau ci-dessous montre donc les chances de chaque équipes de se qualifier.

Figure 4 - prévision pour les phases de groupe de l'Euro 2020

Groupe	Equipe	Premier	Deuxieme	Troisieme	Quatrieme	Qualification
Groupe A	**Turquie**	0.33	0.27	0.22	0.18	0.78
	Italy	0.32	0.27	0.23	0.18	0.78
	Suisse	0.24	0.26	0.26	0.25	0.70
	Pays de Galles	0.12	0.20	0.29	0.39	0.54
Groupe B	**Russie**	0.40	0.32	0.21	0.07	0.90
	Belgique	0.36	0.32	0.24	0.08	0.87
	Dannemark	0.21	0.29	0.35	0.15	0.76
	Finlande	0.03	0.07	0.21	0.70	0.24
Groupe C	**Pays Bas**	0.68	0.23	0.08	0.01	0.97
	Autriche	0.21	0.42	0.29	0.08	0.83
	Ukraine	0.10	0.30	0.48	0.12	0.71
	Macédoine du Nord	0.01	0.05	0.16	0.79	0.13
Groupe D	**Angleterre**	0.35	0.28	0.22	0.15	0.80
	République tchèque	0.30	0.28	0.25	0.18	0.78
	Croatie	0.26	0.26	0.26	0.22	0.73
	Serbie	0.10	0.18	0.27	0.45	0.48
Groupe E	**Espagne**	0.70	0.23	0.06	0.02	0.97
	Suède	0.24	0.48	0.21	0.07	0.84
	Slovaquie	0.05	0.21	0.43	0.30	0.49
	Pologne	0.01	0.08	0.30	0.61	0.19
Groupe F	**Allemagne**	0.53	0.33	0.10	0.04	0.93
	France	0.39	0.40	0.15	0.06	0.88
	Portugal	0.06	0.18	0.42	0.35	0.45
	Islande	0.02	0.10	0.33	0.55	0.24

En nous concentrant sur le groupe F, nous pouvons observer que nos résultats donnent comme vainqueur du groupe l'Allemagne avec 93.30% de chance de se qualifier pour la suite de la compétition, suivit de la France en deuxième position. A noter que le Portugal, champion d'Europe en titre a moins d'une chance sur deux de sortir de la phase de groupe, ce qui est le plus mauvais résultat pour un troisième dans la compétition.

Comme pour les autres précédentes compétitions, nous avons effectués des simulations du tournois entier pour obtenir les probabilités de chaque équipe de gagner le tournois, d'arriver en finale, etc... Nous

avons pour cela fait 5000 simulations, toujours avec les paramètres $\beta_0 = 0.01$ et $\delta_0 = 0.5$, et en utilisant comme données les matchs les plus pertinents depuis moins de quatre ans et les rangs FIFA actuels de chaque équipes.

La Belgique arrive en tête de nos probabilités de gagner, suivit par la France, ce qui est plutôt raisonnable au vue des dernières performances des deux équipes : la France a gagné la dernière coupe du monde, la Belgique y est arrivée en demi-finale, et toutes deux ont fini première de leur groupe de qualification pour l'Euro 2020. La présence de l'Espagne, des Pays-Bas et de l'Allemagne dans le haut du classement n'est également pas une surprise : ce sont des grosses nations du football qui ont bien gérés leurs match de qualifications. On peut par contre s'étonner de la présence de la Croatie en 3ème position et de la Russie en 5ème position, qui ne font clairement pas partie des favoris pour les bookmakers, mais qui, selon notre modèle, ont tout de même de grandes chance d'aller loin dans le tournois.

Figure 5 probabilités d'atteindre chaque phase de l'Euro 2020

Team	winner	final	semi final	quarter final	eigth final
Belgium	0.2058	0.3202	0.4952	0.7198	0.954
France	0.1158	0.2008	0.3478	0.5588	0.9076
Croatia	0.0918	0.1782	0.317	0.5426	0.9048
Spain	0.0894	0.168	0.3008	0.5524	0.8864
Russia	0.0852	0.1752	0.3332	0.574	0.8644
Netherlands	0.065	0.1292	0.2528	0.515	0.8618
Germany	0.0638	0.1322	0.2586	0.4646	0.8342
England	0.0626	0.1292	0.2658	0.4892	0.8694
Ukraine	0.0542	0.1112	0.2264	0.4596	0.8184
Portugal	0.0288	0.0724	0.1634	0.326	0.7164
Switzerland	0.0274	0.0718	0.1706	0.4006	0.8586
Sweden	0.0248	0.0592	0.1462	0.3366	0.7214
Turkey	0.016	0.0434	0.1198	0.3168	0.7844
Denmark	0.0142	0.0394	0.1064	0.2552	0.5624
Austria	0.0132	0.0406	0.0926	0.2432	0.5758
Poland	0.0102	0.0272	0.0774	0.2092	0.561
Slovakia	0.0072	0.019	0.0632	0.1846	0.506
Czech Republic	0.0056	0.016	0.0494	0.1504	0.4364
Romania	0.0056	0.017	0.0478	0.1604	0.4452
Serbia	0.0056	0.0174	0.0522	0.151	0.4496
Wales	0.0038	0.015	0.052	0.1756	0.6194
Finland	0.002	0.007	0.0224	0.08	0.236
Italy	0.001	0.0056	0.0232	0.0828	0.392
North Macedonia	0.001	0.0048	0.0158	0.0516	0.2344

Conclusion

Au vue des résultats obtenus par notre modèle sur d'ancien tournois, nous pouvons avoir relativement confiance en nos prédictions pour l'Euro 2020, car cela reste extrêmement difficile de prévoir des résultats sur un seul ou très peu de matchs, en effet comme on dit "dans le football, sur un match, tout peut arriver". Pour améliorer notre modèle nous pourrions tenter d'utiliser à la fois nos données historiques et les opinions d'experts (qui sont utilisés dans le modèle proposé par F.Louzada), pour refléter au mieux la dynamique des équipes, mais aussi prévoir éventuellement les bons résultats de "petites équipes" comme la Croatie en 2018, ou au contraire les mauvais résultats d'équipe favorites.

Annexe

Annexe A - preuve

proposition 1 :

$$Si \;\; X|\lambda \sim Poisson(\lambda c) \;\; et \;\; \lambda \sim Gamma(\alpha, \beta)$$

$$alors \;\; X \sim BinomialNegative\left(\alpha, \frac{\beta}{\beta + c}\right)$$

preuve :

$$\begin{aligned}
P(X = x) &= \int_0^\infty P(X = x|\lambda)\pi(\lambda)d\lambda \\
&= \int_0^\infty \frac{e^{-\lambda c}(\lambda c)^x}{x!}\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}d\lambda \\
&= \frac{c^x\beta^\alpha}{x!\Gamma(\alpha)}\int_0^\infty e^{-\lambda(c+\beta)}\lambda^{\alpha+x-1}d\lambda \\
&= \frac{c^x\beta^\alpha}{x!\Gamma(\alpha)}\frac{\Gamma(\alpha+x)}{(c+\beta)^{\alpha+x}}\underbrace{\int_0^\infty \frac{(c+\beta)^{\alpha+x}}{\Gamma(\alpha+x)}e^{-\lambda(c+\beta)}\lambda^{\alpha+x-1}d\lambda}_{density\ of\ Gamma(\alpha+x,c+\beta)=1} \\
&= \underbrace{\frac{\Gamma(\alpha+x)}{x!\Gamma(\alpha)}\left(\frac{c}{c+\beta}\right)^x\left(\frac{\beta}{c+\beta}\right)^\alpha}_{BN\left(\alpha,\frac{\beta}{\beta+c}\right)}
\end{aligned}$$

□

Annexe B - probabilité de victoire/égalité/défaite

$$\begin{aligned}
\mathbb{P}_{Win} = \mathbb{P}(X_{AB} > X_{BA}) &= \sum_{k=1}^{\infty}\mathbb{P}(X_{AB} = k)\ \underbrace{\mathbb{P}(k > X_{BA})}_{\sum_{i=0}^{k-1}\mathbb{P}(X_{BA}=i)} \\
&= \sum_{k=1}^{\infty}\sum_{i=0}^{k-1}\mathbb{P}(X_{AB=k})\mathbb{P}(X_{BA} = i) \\
\mathbb{P}_{Draw} = \mathbb{P}(X_{AB} = X_{BA}) &= \sum_{k=1}^{\infty}\mathbb{P}(X_{AB} = k)\mathbb{P}(X_{BA} = k) \\
\mathbb{P}_{Loss} = \mathbb{P}(X_{AB} < X_{BA}) &= \sum_{k=1}^{\infty}\mathbb{P}(X_{BA} = k)\ \underbrace{\mathbb{P}(X_{AB} < k)}_{\sum_{i=0}^{k-1}\mathbb{P}(X_{AB}=i)} \\
&= \sum_{k=1}^{\infty}\sum_{i=0}^{k-1}\mathbb{P}(X_{AB=i})\mathbb{P}(X_{BA} = k)
\end{aligned}$$

Annexe C - Distance de DeFinetti (phase de groupe - coupe du monde 2014)

Figure 6 Prédictions pour chaque match de la phase de groupe - coupe du monde 2014

teamA	score	teamB	pred1	deFi1	good1	pred2	deFi2	good2
Brazil	3-1	Croatia	(0.63 0.23 0.14)	0.207	1	(1 0 0)	0	1
Mexico	1-0	Cameroon	(0.66 0.2 0.14)	0.176	1	(1 0 0)	0	1
Brazil	0-0	Mexico	(0.57 0.23 0.2)	0.957	0	(1 0 0)	2	0
Cameroon	0-4	Croatia	(0.21 0.27 0.52)	0.353	1	(0 0 1)	0	1
Cameroon	1-4	Brazil	(0.06 0.12 0.82)	0.049	1	(0 0 1)	0	1
Croatia	1-3	Mexico	(0.27 0.28 0.45)	0.452	1	(1 0 0)	2	0
Spain	1-5	Netherlands	(0.4 0.25 0.35)	0.645	0	(1 0 0)	2	0
Chile	3-1	Australia	(0.83 0.12 0.05)	0.046	1	(1 0 0)	0	1
Australia	2-3	Netherlands	(0.02 0.06 0.92)	0.011	1	(0 0 1)	0	1
Spain	0-2	Chile	(0.47 0.27 0.26)	0.84	0	(1 0 0)	2	0
Australia	0-3	Spain	(0.03 0.11 0.86)	0.031	1	(0 0 1)	0	1
Netherlands	2-0	Chile	(0.49 0.22 0.29)	0.39	1	(0 0 1)	2	0
Colombia	3-0	Greece	(0.47 0.28 0.25)	0.426	1	(1 0 0)	0	1
Ivory Coast	2-1	Japan	(0.41 0.21 0.37)	0.527	1	(1 0 0)	0	1
Colombia	2-1	Ivory Coast	(0.54 0.24 0.23)	0.323	1	(1 0 0)	0	1
Japan	0-0	Greece	(0.32 0.24 0.43)	0.868	0	(0 0 1)	2	0
Japan	1-4	Colombia	(0.21 0.2 0.59)	0.249	1	(0 0 1)	0	1
Greece	2-1	Ivory Coast	(0.39 0.28 0.33)	0.556	1	(1 0 0)	0	1
Uruguay	1-3	Costa Rica	(0.76 0.2 0.04)	1.543	0	(1 0 0)	2	0
England	1-2	Italy	(0.36 0.3 0.33)	0.669	0	(0 0 1)	0	1
Uruguay	2-1	England	(0.42 0.28 0.3)	0.498	1	(1 0 0)	0	1
Italy	0-1	Costa Rica	(0.67 0.28 0.06)	1.411	0	(1 0 0)	2	0
Italy	0-1	Uruguay	(0.28 0.29 0.43)	0.479	1	(0 0 1)	0	1
Costa Rica	0-0	England	(0.05 0.25 0.69)	1.038	0	(0 0 1)	2	0
Switzerland	2-1	Ecuador	(0.52 0.27 0.21)	0.344	1	(1 0 0)	0	1
France	3-0	Honduras	(0.39 0.31 0.3)	0.565	1	(1 0 0)	0	1
Switzerland	2-5	France	(0.49 0.3 0.21)	0.963	0	(1 0 0)	2	0
Honduras	1-2	Ecuador	(0.32 0.28 0.4)	0.542	1	(0 0 1)	0	1
Honduras	0-3	Switzerland	(0.18 0.26 0.57)	0.285	1	(0 0 1)	0	1
Ecuador	0-0	France	(0.34 0.31 0.35)	0.709	0	(0 0 1)	2	0
Argentina	2-1	Bosnia	(0.65 0.2 0.15)	0.188	1	(1 0 0)	0	1
Iran	0-0	Nigeria	(0.2 0.27 0.53)	0.858	0	(1 0 0)	2	0
Argentina	1-0	Iran	(0.85 0.11 0.04)	0.034	1	(1 0 0)	0	1
Nigeria	1-0	Bosnia	(0.27 0.25 0.48)	0.834	0	(0 0 1)	2	0
Nigeria	2-3	Argentina	(0.09 0.14 0.78)	0.075	1	(0 0 1)	0	1
Bosnia	3-1	Iran	(0.61 0.26 0.13)	0.234	1	(1 0 0)	0	1
Germany	4-0	Portugal	(0.73 0.17 0.1)	0.115	1	(1 0 0)	0	1
Ghana	1-2	United States	(0.12 0.17 0.71)	0.126	1	(0 0 1)	0	1
Germany	2-2	Ghana	(0.89 0.07 0.04)	1.648	0	(1 0 0)	2	0
United States	2-2	Portugal	(0.52 0.26 0.21)	0.858	0	(0 0 1)	2	0
United States	0-1	Germany	(0.19 0.18 0.63)	0.202	1	(0 0 1)	0	1
Portugal	2-1	Ghana	(0.46 0.3 0.24)	0.443	1	(1 0 0)	0	1
Belgium	2-1	Algeria	(0.42 0.33 0.25)	0.509	1	(1 0 0)	0	1
Russia	1-1	South Korea	(0.44 0.3 0.26)	0.757	0	(1 0 0)	2	0
Belgium	1-0	Russia	(0.43 0.35 0.23)	0.5	1	(1 0 0)	0	1
South Korea	2-4	Algeria	(0.25 0.28 0.48)	0.41	1	(0 0 1)	0	1
South Korea	0-1	Belgium	(0.17 0.25 0.58)	0.268	1	(0 0 1)	0	1
Algeria	1-1	Russia	(0.34 0.35 0.3)	0.628	1	(0 0 1)	2	0