



**HAL**  
open science

## Imputation crédibiliste pour la prédiction de charge interne de joueurs de football

Rayane Elimam, Nicolas Sutton-Charani, Jacky Montmain, S. Perrey

### ► To cite this version:

Rayane Elimam, Nicolas Sutton-Charani, Jacky Montmain, S. Perrey. Imputation crédibiliste pour la prédiction de charge interne de joueurs de football. LFA'2020 - 29èmes Rencontres Francophones sur la Logique Floue et ses Applications, Oct 2020, Sète, France. hal-02969311

**HAL Id: hal-02969311**

**<https://imt-mines-ales.hal.science/hal-02969311v1>**

Submitted on 28 Apr 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Imputation crédibiliste pour la prédiction de charge interne de joueurs de football

Rayane ELIMAM<sup>1</sup>  
Jacky MONTMAIN<sup>3</sup>

Nicolas SUTTON-CHARANI<sup>2</sup>  
Stéphane PERREY<sup>4</sup>

EuroMov, Digital Health in Motion  
Univ. Montpellier, IMT Mines Ales, France

<sup>1</sup> 700 Avenue du Pic Saint Loup, 34090, Montpellier, France, rayane.elimam@mines-ales.fr

<sup>2</sup> 6 avenue de Clavières, F-30 319, Alès Cedex, France, nicolas.sutton-charani@mines-ales.fr

<sup>3</sup> 700 Avenue du Pic Saint Loup, 34090, Montpellier, France, stephane.perrey@umontpellier.fr

<sup>4</sup> 6 avenue de Clavières, F-30 319, Alès Cedex, France, jacky.montmain@mines-ales.fr

## Résumé :

L'objectif de cet article est de comparer trois approches de gestion des données manquantes dans un contexte de classification avec étiquettes séquentielles à pas de temps variable. Les données concernent le suivi longitudinal de vingt-sept joueurs professionnels de football en termes de planning d'entraînement et de charge interne ressentie. Une approche sans imputation utilise l'ancienneté des données, les deux autres approches sont basées sur des modèles d'incertitude lors de l'imputation des données manquantes. L'Evidential K-Nearest Neighbors (EKNN) est utilisé pour la prédiction de charge interne en prenant en compte l'incertitude des étiquettes, et l'approche sans incertitude avec K-Nearest Neighbors standard. Les résultats montrent une certaine amélioration des prédictions pour les modèles utilisant l'incertitude des données imputées.

## Mots-clés :

Imputation, données temporelles, pas de temps variable, charge interne, football, fonctions de croyance, EKNN

## Abstract:

The objective of this article is to compare three approaches of missing data management in a classification context with sequential labels at variable time steps. The data concern the longitudinal monitoring of twenty-seven professional soccer players in terms of training schedules and perceived internal perceived load. One approach without imputation using the age of the data, the other two approaches are based on uncertainty models at the missing data imputation step. The Evidential K-Nearest Neighbors (EKNN) is used for internal load prediction taking into account labels uncertainty, and the K-Nearest Neighbors for the approach without uncertain labels. Results show moderate prediction improvement for models based on imputed data uncertainty.

## Keywords:

Imputation, time series, variable time step, internal load, soccer, belief functions, EKNN

## 1 Introduction

L'imputation de données manquantes est une étape souvent nécessaire dans les applications où la quantité de données est limitée et où les modèles prédictifs utilisés sont prévus pour des jeux de données complets [12] [13]. Il existe différentes techniques d'imputation (par omission, moyenne, apprentissage) et leur choix a un impact non-négligeable sur les performances des modèles prédictifs mis en œuvre par la suite [14].

Un pas de temps irrégulier dans un contexte de données évoluant dans le temps peut être assimilé à un problème de données manquantes. De telles données peuvent être transformées (extrapolation, etc) de manière à se ramener à un pas de temps constant [15]. Dans le milieu sportif, les préparateurs physiques et entraîneurs doivent gérer la charge interne de leurs effectifs. Cette charge correspond à la somme des stimuli physiologiques et psychologiques (stress) pendant l'entraînement [17]. Prédire cette valeur est utile pour les responsables qui peuvent alors optimiser et personnaliser les entraînements, cependant peu d'études s'intéressent à ce problème d'un point de vue multivarié [4, 3, 7, 5, 6]. Pour estimer la charge d'entraînement, des questionnaires comme le "Rating of Perceived Exertion" (RPE) [16] sont souvent utilisés. Après chaque session d'en-

traînement les sportifs remplissent le formulaire RPE en renseignant la charge interne ressentie. Nous désignerons cette charge interne ressentie, seulement par charge interne, on parle de ressentie car le RPE est une mesure subjective. Le RPE peut être interprété de façon dynamique, la charge d'entraînement à l'instant  $t$  étant notamment une conséquence des charges précédentes. En pratique, les joueurs ne remplissent qu'occasionnellement les RPE, traiter la charge interne comme un processus à pas de temps constant est donc impossible et l'imputation des RPE manquants représente alors un objectif important pour sa prédiction si on la base sur son passé.

Dans cet article, l'imputation de données manquantes au sein de séries temporelles est revisitée sous un angle d'incertitude de manière à pouvoir gérer les pas de temps irréguliers pour la prédiction de RPE. Ce travail est une étude préliminaire, le RPE est de ce fait discrétisé en 3 classes de proportions comparables, les 3 classes représentent 3 niveaux de difficulté, faible, normal, élevée.

La nature subjective du RPE et l'irrégularité dans le remplissage du formulaire impliquent une forte incertitude quant à la vraie valeur de la charge interne d'entraînement supportée par les joueurs. Le formalisme des fonctions de croyances [9, 10] permet une modélisation souple de l'incertitude relative au RPE et certains travaux associés permettent l'apprentissage de classifieurs capables de prendre en compte l'incertitude inhérente aux données d'apprentissage [20, 18, 19]. Le modèle de l'EKNN [1], permet ainsi de prendre en compte l'incertitude des étiquettes modélisée sous forme de fonctions de croyance.

En section 2 la problématique de prédiction de la charge interne des footballeurs est présentée, la section 3 détaille l'approche incertaine proposée en termes de prédiction et de modélisation de l'incertitude des étiquettes, la section 4 est dédiée aux expériences et des conclusions et perspectives sont finalement pro-

posées en section 5.

## 2 Prédiction de la charge interne des footballeurs professionnels

Après avoir décrit le formalisme et les différentes variables de la problématique de la prédiction de la charge interne, les données disponibles sont rapidement décrites.

### 2.1 Modélisation du problème

La charge interne est une conséquence directe de l'entraînement. En outre, on fait l'hypothèse qu'elle constitue un processus dynamique dont les valeurs aux différentes dates ne sont pas indépendantes. La manière de réagir aux entraînements dépend aussi des caractéristiques physiques d'un joueur. On pose alors :

$$Y_{t_i}^p = f(Y_{t_{i-1}}^p, \dots, Y_{t_{i-q}}^p, X_{t_i}^p, X_{t_{i-1}}^p, \dots, X_{t_{i-q}}^p, Z^p) \quad (1)$$

avec

- $Y_{t_i}^p \in \mathcal{C} = \{C_1, C_2, C_3\}$  le RPE discrétisé du joueur  $p$  au temps  $t_i$
- $X_{t_i}^p = ({}^1X_{t_i}^p, \dots, {}^{14}X_{t_i}^p)$  l'entraînement suivi par le joueur  $p$  à la date  $t_i$  avec  ${}^jX_{t_i}^p \in \{0, 1\}$  une variable binaire spécifiant si l'exercice  $j$  a été réalisé par  $p$  en  $t_i$
- $Z^p = (Z_1^p, Z_2^p, Z_3^p) \in R_+^{*3}$  le vecteur d'attributs personnels du joueur  $p$  contenant son âge (années), son poids (kg) et sa taille (cm)
- $q$  la taille de l'historique des données considérées

L'enjeu est alors d'estimer la meilleure fonction  $f$  capable, pour chaque joueur  $p$ , de prédire  $Y_{t_i}^p$  à partir de son historique de charge interne  $(Y_{t_{i-1}}^p, \dots, Y_{t_{i-q}}^p)$ , de son historique d'entraînement  $(X_{t_i}^p, X_{t_{i-1}}^p, \dots, X_{t_{i-q}}^p)$  et de ses attributs personnels  $Z^p$ .

### 2.2 Données disponibles

Dans cette section, les données utilisées pour les expérimentations décrites en section

4 sont décrites. Elles concernent vingt-sept joueurs professionnels de football d'une équipe française de deuxième division. Les données collectées couvrent la période allant du 26 Juillet 2017 au 15 Novembre 2017. Sur cette période nous avons 51 dates, ou il y a au moins un joueur qui a attribué un RPE à sa séance d'entraînement

Soit  $D = (X, Y, Z)$  avec

$$- X = \begin{pmatrix} x_{t_1}^1 & \dots & x_{t_1}^P \\ \vdots & & \vdots \\ x_{t_T}^1 & \dots & x_{t_T}^P \end{pmatrix}$$

la matrice de variables d'entraînements ordonnées chronologiquement de taille  $T \times P$  avec  $P = 27$  où chaque élément contient 14 variables binaires associées aux différents exercices possibles lors d'un entraînement.

$$- Y = \begin{pmatrix} y_{t_1}^1 & \dots & y_{t_1}^P \\ \vdots & & \vdots \\ y_{t_T}^1 & \dots & y_{t_T}^P \end{pmatrix}$$

la matrice d'étiquettes ordonnées chronologiquement de taille  $T \times P$  avec  $P = 27$  où chaque colonne représente les étiquettes d'un joueur évoluant avec le temps.

$$- Z = \begin{pmatrix} z_1^1 & \dots & z_1^P \\ z_2^1 & & z_2^P \\ z_3^1 & & z_3^P \end{pmatrix}$$

une matrice contenant les valeurs des âges ( $z_1^p$ ), tailles ( $z_2^p$ ) et poids ( $z_3^p$ ) des  $P$  joueurs.

Un planning d'entraînement  $(x_{t_i}^1, x_{t_i}^2, \dots, x_{t_i}^J)$  contient  $J$  variables binaires représentant des exercices physiques ou évènements (Soin) ayant eu lieu ou pas. Le tableau 1 présente les distributions des variables d'entraînement.

Le tableau 2 expose la moyenne et l'écart-type de chaque attribut personnel. La matrice  $Y$  contient 503 RPE remplis, la figure 1 expose la distribution du RPE avant discrétisation.

Dans la figure 1 on constate une concentration des RPE autour de la valeur 5. La discrétisation suivante est donc adoptée pour la suite de l'article :

Tableau 1 – Distributions pour les variables d'entraînement

Variable	0	1
Versus	0.45	0.55
Toro	0.84	0.16
Étirements	0.96	0.04
Réath	0.93	0.07
Abdos	0.96	0.04
Gainage	0.95	0.05
Aerobie	0.77	0.23
Brz	0.98	0.024
Mob	0.98	0.02
Mult	0.72	0.28
Non	0.56	0.45
PPG	0.93	0.066
Tech	0.84	0.16
Soins	0.86	0.14

Tableau 2 – Moyenne et écart-type des attributs personnels

Attribut	Moyenne	écart-type
Poids	76.2	8.3
Taille	180.3	7.7
Âge	29	6

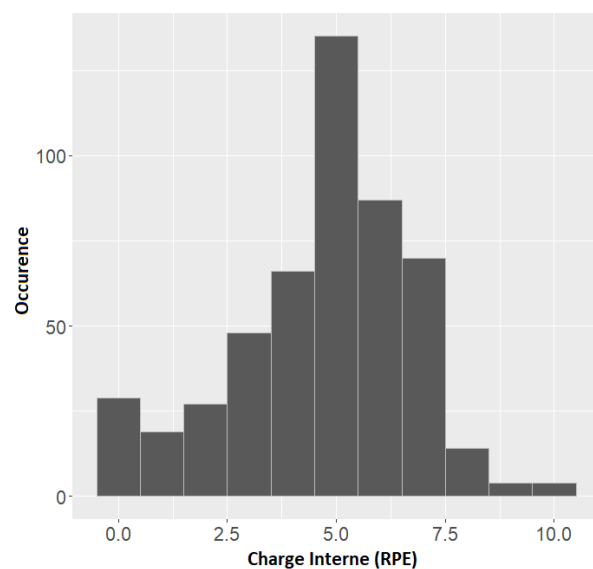


Figure 1 – Distribution du RPE

- Classe 0 :  $RPE < 5$
- Classe 1 :  $RPE = 5$
- Classe 2 :  $RPE > 5$

On obtient ainsi 3 classes d'effectifs relativement proches.

### 3 Approche crédibiliste

Dans cette section nous justifions l'utilisation de théories modernes de l'incertain, les principes de l'algorithme utilisé pour la prédiction de la charge interne (EKNN) [1] sont brièvement rappelés, et enfin deux modèles d'imputation d'étiquettes avec prise en compte d'incertitude sont proposés.

#### 3.1 Motivations

En général les différentes incertitudes des données applicatives sont soit ignorées, soit prises en compte par des modèles ensemblistes ou probabilistes. Il existe néanmoins d'autres formalismes plus souples permettant une modélisation fine et spécifique à la nature de l'incertitude en question :

- sous ensembles flous [2]
- possibilités [8]
- fonctions de croyance [9] [10]
- probabilités imprécises [11]

Dans notre contexte, les données contiennent plusieurs niveaux d'incertitude :

- pas de temps irrégulier : de l'irrégularité "naturelle" des entraînements (weekends, jours fériés, vacances) et du remplissage occasionnel des RPE par les joueurs résulte une matrice  $Y$  creuse.
- nature subjective du RPE : la perception de la charge interne en est une mesure naturellement imprécise.

Même si le RPE est préalablement discrétisé de manière notamment à diminuer le bruit créé par une mesure trop précise associée à une perception subjective, il est souhaitable de modéliser l'incertitude relative au RPE à un niveau plus fin.

#### 3.2 EKNN

L'Evidential K-Nearest Neighbors (EKNN) [1] est un modèle de classification développé dans le formalisme des fonctions de croyance. Il tire profit de la richesse et la souplesse des règles de combinaison crédibilistes et permet de prendre en compte les étiquettes incertaines dans les données d'apprentissage. Pour un nouvel exemple à classer  $x$ , une fois le voisinage déterminé en calculant les distances entre exemples, chaque voisin est considéré comme une source d'information et est associé à une fonction de masse déterminée par sa distance à  $x$ . Au moment de la prédiction, au lieu d'un vote classique (pondéré ou non), les masses des voisins sont combinées conjonctivement par la règle de Dempster en une fonction de croyance prédictive. Sa transformation en probabilité pignistique permet une prédiction précise en considérant finalement l'étiquette la plus probable.

Le package 'evclass' de R, disponible sur le CRAN, permet l'implantation de l'EKNN sur données précises. Dans le cadre de cet article il a été étendu de manière à pouvoir être utilisé sur des données contenant des étiquettes incertaines (modélisées par fonctions de croyance), les codes sont disponibles à [https://github.com/lgi2p/evclass\\_uncertain\\_labels](https://github.com/lgi2p/evclass_uncertain_labels).

#### 3.3 Modèle d'incertitude des étiquettes

Nous proposons ici deux approches d'imputation impliquant une modélisation de l'incertitude associée. La première correspond à une interpolation linéaire associée à un degré d'incertitude déterminé par la fenêtre temporelle durant laquelle les joueurs n'ont pas rempli de RPE. La deuxième peut être assimilée à de l'imputation par apprentissage basée sur l'EKNN. Les charges internes (RPE : formulaire à remplir) collectés couvrent une période de 114 jours, sur cette période seulement 51 jours ont au moins 1 RPE rempli.

### Interpolation incertaine (evid\_imp)

La charge interne étant considérée comme un processus temporel il est envisageable d'imputer les RPE manquants à partir des précédents et des suivants.

Pour chaque RPE manquant  $y_{t_i}^p \in \mathcal{C}$  :

- on calcule la moyenne arithmétique entre les RPE suivants et précédents  $y_{t_{i-j}}^p$  et  $y_{t_{i+k}}^p$  :

$$\widehat{y}_{t_i}^p = \frac{1}{2}y_{t_{i-j}}^p + \frac{1}{2}y_{t_{i+k}}^p \quad (2)$$

- on définit ensuite la fonction de masse suivante :

$$\begin{cases} m(\widehat{y}_{t_i}^p) &= \alpha \\ m(\mathcal{C}) &= 1 - \alpha \end{cases} \quad (3)$$

où  $\alpha$  représente le degré de certitude de  $\widehat{y}_{t_i}^p$  défini par

$$\alpha = \frac{\beta}{\delta t} \quad (4)$$

avec  $\beta \in [0, 2]$  une constante à fixer (empiriquement ou de façon experte) et  $\delta t = t_{i+k-1} - t_{i-j+1}$  correspond à la période durant laquelle le joueur  $p$  n'a pas rempli de RPE.

Pour les RPE remplis (charges internes disponibles) on a  $m(y_{t_i}^p) = \alpha = 1$ .

$\beta$  est une valeur qui permet de gérer la décroissance de la masse, plus  $\beta$  est proche de 2, plus la masse décroît lentement.

### Imputation basé sur l'EKNN (eknn\_imp)

Dans cette approche l'idée est d'utiliser les joueurs qui ont suivi le même entraînement au temps  $t$  comme voisinage du joueur qui n'a pas rempli son RPE (pas de charge interne attribuée) en  $t$ , le nombre de voisins varie selon le nombre de joueurs présents à l'entraînement  $t$ . Les attributs personnels des voisins permettent le calcul des distances entre le joueur pour lequel la charge interne est à prédire/imputer et ses voisins. On utilise ensuite l'EKNN pour prédire le RPE manquant.

## 4 Expériences

Pour évaluer les 2 méthodes d'imputation crédibiliste proposées en section 3.3 dans un cadre de classification, nous définissons ici les alternatives de référence (baseline), nous décrivons le protocole d'évaluation défini de façon chronologique et présentons les résultats obtenus.

### 4.1 Méthodes alternatives

Une méthode de référence naïve (*baseline*) consistant à prédire systématiquement la classe la plus probable sera ici mise en œuvre de manière à vérifier que les prédictions calculées possèdent un minimum de potentiel prédictif.

Nous définissons aussi une méthode de prédiction sans imputation qui utilise, pour chaque joueur et chaque date de prédiction, les  $q$  derniers RPE et variables d'entraînement (en plus des attributs personnels des joueurs) associées à leurs anciennetés. L'information d'ancienneté constitue en effet un autre moyen de pondérer "automatiquement" les données trop anciennes. Cette méthode permet ainsi de mesurer l'intérêt d'une modélisation de la charge interne en série temporelle à pas de temps variable dans un contexte où on dispose malgré tout de certaines données passées. Elle n'implique aucun modèle d'incertitude et sera nommée "precise".

### 4.2 Protocole d'évaluation

Nous choisissons de nous placer dans un contexte de prédiction chronologique du futur à partir du passé car il correspond à la réalité terrain du préparateur physique. L'intervalle temporel sur lequel les RPE (charges internes remplis via le formulaire RPE) ont été remplis s'étale sur 114 jours pendant lesquels 51 dates d'entraînement sont renseignées. De manière à disposer de suffisamment de données d'apprentissage et de test, les 25 premières et 11 dernières dates sont respectivement toujours utilisées en apprentissage et en test, les

dates intermédiaires sont initialement contenues dans l'échantillon de test puis progressivement ajoutées aux données d'apprentissage. La prédiction naïve (baseline) à chaque itération d'apprentissage se base sur la classe majoritaire dans le jeu d'entraînement. Pour l'imputation par EKNN (*eknn\_imp*), la taille du voisinage correspond, pour chaque joueur, au nombre de joueurs ayant des données de RPE lors du même entraînement. Pour les prédictions par KNN ou EKNN le nombre de voisins considérés est de 5 par défaut (une rapide étude de sensibilité a montré un faible impact de ce paramètre sur les performances prédictives).

### 4.3 Résultats

Nous nous sommes restreints à de petites valeurs de  $q$  car la plupart des joueurs n'ont rempli que peu de RPE sur la période des 51 entraînements. La valeur  $q = 0$  indique qu'aucun historique n'est pris en compte, seul l'entraînement présent ainsi que les attributs personnels sont utilisés comme variables prédictives. De manière à quantifier d'une part les apports des 2 modèles d'imputations (*evid\_imp*, *eknn\_imp*) et de la méthode précise utilisant l'ancienneté des données (*precise*), et d'autre part ceux des modèles d'incertitude associés aux étiquettes basés sur la taille des périodes sans données de RPE (3-4) et aux distances entre exemples comme initialement proposé dans l'EKNN [1], 3 niveaux de comparaisons sont ici proposés. Une première comparaison est effectuée entre les modèles *baseline*, *precise*, *evid\_imp*, et *eknn\_imp* pour différentes tailles d'historique ( $q$ ) en termes de justesse moyenne de prédiction ( $\pm$  écarts-types) par KNN (Tableau 3). L'apport du modèle d'incertitude inhérent à l'EKNN est identifié en comparant les prédictions des mêmes modèles mais par EKNN (Tableau 4). Enfin, les 2 modèles d'imputation crédibiliste sont évalués en comparant les prédictions qu'ils produisent par EKNN (Tableau 5).

On constate que les données possèdent un pou-

voir prédictif, les différentes approches obtenant presque toujours de meilleures performances que la baseline. Globalement, les meilleures justesses de prédiction, à savoir 0.50 et 0.51 (Tableaux 5 et 4) sont obtenues par l'approche *precise* sans le modèle d'incertitude (3-4) pour  $q = 0$ . On remarque que cette approche devient moins efficace quand un historique est pris en compte même si l'EKNN permet d'obtenir de meilleures prédictions, la différence étant plus marquée lorsque des données historiques sont prises en compte ( $q > 0$ ). La prise en compte d'incertitude semble donc ne présenter ici qu'un léger intérêt et uniquement au niveau de la pondération des voisins dans l'approche EKNN.

Les performances des modèles avec imputation *evid\_imp* et *eknn\_imp* semblent plus adaptés à l'EKNN qu'au KNN et la prise en compte du modèle d'incertitude leur permet presque systématiquement d'augmenter leurs justesses. Le modèle d'imputation *evid\_imp* donne des prédictions plus justes en moyenne que *eknn\_imp*. On peut noter que les justesses moyennes obtenues par EKNN avec imputation de type *evid\_imp* associée au modèle d'incertitude (3-4) pour de petits historiques sont légèrement inférieures à celles obtenues avec l'approche *precise* par EKNN mais presque toujours supérieures à celles obtenues par KNN pour la même configuration. Le modèle d'incertitude (3-4) semble donc adapté à l'approche *evid\_imp* même si son utilisation n'apporte aucun gain prédictif sur ce jeu de données.

Les figures 2 et 3 montrent l'évolution des taux de bonne classification (justesse) au cours du temps des différentes approches obtenues respectivement pour pour  $q = 0$  et  $q = 1$  selon les mêmes conditions que dans les Tableaux 4 et 5, i.e. par EKNN respectivement sans et avec modèle d'incertitude (3-4). Le niveau de justesse de l'approche *precise* augmente légèrement au cours du temps sauf en fin de période. Ceci est probablement dû au fait qu'il bénéficie de l'ajout de données d'apprentissage mais que la quantité de données de

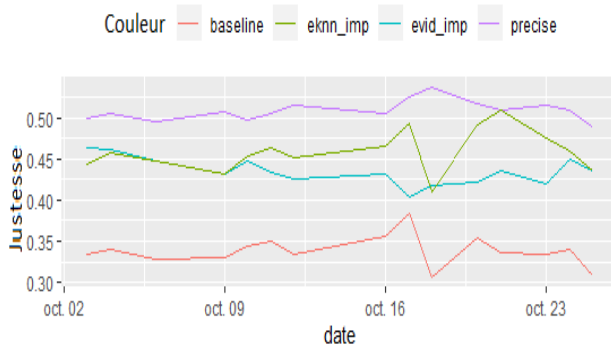


Figure 2 – Évolution de la justesse par EKNN pour  $q = 0$

tests devient insuffisante en fin de période pour obtenir des estimations fiables de son pouvoir prédictif.

De manière générale il apparaît que, pour toutes les approches, la prise en compte de données historiques nuit aux performances prédictives. Ce dernier point est surprenant d'un point de vue de l'information mais peut constituer un résultat important du point de vue cognitif et sportif. La charge interne n'apparaît ainsi pas impactée par ses niveaux antérieurs, même si on aurait pu s'attendre à ce que l'état psychologique d'un joueur ne soit pas indépendant de ses niveaux dans un passé proche. En outre, on observe aussi que les données d'entraînements passés ne permettent pas non plus d'amélioration prédictives. Une potentielle explication pourrait être lié au sommeil qui pourrait permettre une réelle régénération physico-cognitive qui rendrait l'état psychologique des joueurs surtout dépendant des conditions présentes d'entraînement.

## 5 Conclusion

Cet article propose deux approches d'imputation de données manquantes pour des séries temporelles à pas de temps variable. L'utilisation d'un classifieur crédibiliste (EKNN) pour la prédiction de la charge interne à partir de données historiques de RPE, d'entraînement et

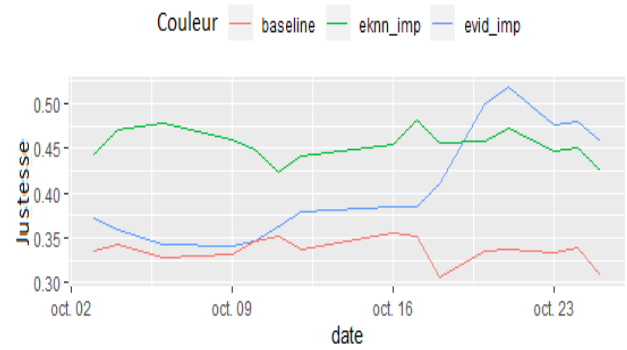


Figure 3 – Évolution de la justesse par EKNN avec étiquette incertaine pour  $q = 1$

Tableau 3 – Résultats avec étiquettes certaines par KNN

q	baseline	precise	evid_imp	eknn_imp
0	0.34 ±0.02	<b>0.50</b> ±0.02	0.43 ±0.02	0.44 ±0.02
1	0.34 ±0.01	<b>0.42</b> ±0.05	0.40 ±0.04	0.39 ±0.03
2	0.34 ±0.02	<b>0.41</b> ±0.06	0.38 ±0.03	0.35 ±0.03
3	0.36 ±0.02	<b>0.42</b> ±0.04	0.40 ±0.03	0.38 ±0.02
4	0.35 ±0.02	<b>0.42</b> ±0.05	0.38 ±0.03	0.35 ±0.02
5	0.36 ±0.02	0.40 ±0.04	<b>0.40</b> ±0.03	0.37 ±0.02

Tableau 4 – Résultats avec étiquettes certaines par EKNN

q	baseline	precise	evid_imp	eknn_imp
0	0.34 ±0.02	<b>0.51</b> ±0.01	0.44 ±0.01	0.46 ±0.02
1	0.34 ±0.02	<b>0.46</b> ±0.03	0.42 ±0.04	0.42 ±0.02
2	0.34 ±0.02	<b>0.44</b> ±0.03	0.42 ±0.02	0.38 ±0.02
3	0.36 ±0.02	<b>0.44</b> ±0.02	0.38 ±0.02	0.37 ±0.02
4	0.35 ±0.02	<b>0.43</b> ±0.04	0.40 ±0.04	0.36 ±0.02
5	0.36 ±0.02	0.40 ±0.03	<b>0.41</b> ±0.03	0.39 ±0.02

Tableau 5 – Résultats avec étiquettes incertaines

q	evid_imp	eknn_imp
0	<b>0.48</b> ±0.01	0.47 ±0.02
1	<b>0.48</b> ±0.01	0.45 ±0.02
2	<b>0.44</b> ±0.02	0.36 ±0.03
3	<b>0.41</b> ±0.01	0.38 ±0.03
4	<b>0.39</b> ±0.02	<b>0.39</b> ±0.01
5	<b>0.42</b> ±0.02	0.41 ±0.03

d'attributs personnels montre certaines limites qui ouvrent la voix à différentes perspectives :



- l'utilisation d'autres classifieurs capables de prendre en compte des données incertaines [20]
- le développement d'une approche mixte combinant imputation incertaine et l'utilisation des données d'ancienneté
- extension du protocole d'évaluation au cas de régression (le RPE étant naturellement numérique)

L'implantation de l'extension du package 'ev-class' (R) aux étiquettes incertaines pourra servir de benchmark pour différentes méthodes d'apprentissage à partir de données incertaines.

#### Remerciements :

Nous tenons à remercier Emmanuel Vallance pour nous avoir permis d'utiliser son jeu de données.

## Références

- [1] T. Denœux. A k-nearest neighbor classification rule based on Dempster-Shafer Theory. *IEEE Transactions on Systems, Man, and Cybernetics*, 25 :5, 1995.
- [2] L. A. Zadeh, Fuzzy sets, *Information and Control* 8 : 338-353 . 1965
- [3] J.D. Bartlett, F. O'Connor , N. Pitchford, L. Torres-Ronda, S. J. Robertson. Relationships Between Internal and External Training Load in Team-Sport Athletes : Evidence for an Individualized Approach. *Int J Sports Physiol Perform.* 12(2) : 230-234. 2017
- [4] D.L. Carey , J. Crow , K. L. Ong , et al. Optimizing Preseason Training Loads in Australian Football. *Int J Sports Physiol Perform.* 13(2) : 194-199. 2016
- [5] G. Vandewiele, Y. Geurkink, M. Lievens, et al. Enabling Training Personalization by Predicting the Session Rate of Perceived Exertion (sRPE).*Print* . 1-12. 2017
- [6] A. Jaspers, T. De Beéck , M. Brink, et al. Relationships Between the External and Internal Training Load in Professional Soccer : What Can We Learn From Machine Learning ? . *Int J Sports Physiol Perform.* 13(5) : 625-630. 2017
- [7] A. Rossi, E. Perri, L. Pappalardo, P. Cintia, . Relationship between External and Internal Workloads in Elite Soccer Players : Comparison between Rate of Perceived Exertion and Training Load. *Applied Sciences.* 9. 2019
- [8] H. Prade and D. Dubois, Possibility Theory : Qualitative and Quantitative Aspects *Quantified Representation of Uncertainty and Imprecision*, Springer, Dordrecht, 1998.
- [9] A. P. Dempster, Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics.* 38 (2) : 325-339. 1966
- [10] G. Shafer. A. Mathematical Theory of Evidence. *Princeton University Press*, 1976
- [11] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall. 1991
- [12] J. Jerez, I. Molina, L. García. Missing Data Imputation Using Statistical and Machine Learning Methods in a Real Breast Cancer Problem. *Artificial Intelligence in Medicine.* 50 : 105-115. 2010
- [13] K. Lakshminarayan, S. A. Harp, T. Samad. Imputation of Missing Data in Industrial Databases. *Applied Intelligence.* 11 : 259-275. 1999
- [14] A. Farhangfar, L. Kurgan, J. Dy. Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition.* 41 : 3692-3705. 2008
- [15] A. Beygelzimer, E. Erdogan, S. Ma, I. Rish. Statistical Models for Unequally Spaced Time Series. *Proceedings of the 2005 SIAM International Conference on Data Mining*, 2005. pp.
- [16] G. A. Borg. Psychophysical bases of perceived exertion. *Med Sci Sports Exerc.* 14 (5) : 377-81. 1982
- [17] F. M. Impellizzeri, E. Rampinini, S. M. Marcora. Physiological assessment of aerobic training in soccer. *J Sports Sci.* 23(6) : 583-592. 2005
- [18] Z. Elouedi, K. Mellouli, P. Smets. Belief decision trees : theoretical foundations. *International Journal of Approximate Reasoning.* 28(2-3) : 91-124. 2001
- [19] T. Denœux. A neural network classifier based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics*, 2000. pp. 131-150.
- [20] N. Sutton-Charani, S. Destercke, T. Denœux. Learning decision trees from uncertain data with an evidential EM approach. *International Conference on Machine Learning and Applications (ICMLA)*, 2013. pp. 111-116.