Désagrégation temporelle du cumul annuel de croissance de l'herbe

Laurent Spillemaecker*, Thomas Guyet**,
Simon Malinowski***, Anne-Isabelle Graux****

*ENSAI/IRISA
**Inria, Centre Inria de Lyon
thomas.guyet@inria.fr
***Université Rennes 1/Inria/IRISA
**** PEGASE, INRAE, Institut Agro, 35590, Saint Gilles, France

Résumé. L'information sur la croissance de l'herbe au cours de l'année est essentielle à certains modèles simulant l'utilisation de cette ressource herbagère pour la production de fourrages conservés ou pour l'alimentation des animaux au pâturage. Malheureusement, cette information n'est que rarement disponible. Le défi réside dans la reconstruction de la croissance de l'herbe à partir de deux sources d'informations : les données journalières sur le climat (pluie, rayonnement, etc.) et la croissance cumulée sur l'année. Dans cet article, nous formulons ce défi comme un problème de désagrégation du cumul annuel en une série temporelle. Pour résoudre ce problème, on applique une méthode de prévision de série temporelle en s'aidant des informations sur le climat. Plusieurs variantes de la méthode sont proposées et comparées expérimentalement à partir d'une base de données issue d'un modèle de simulation des prairies. Les résultats montrent que notre méthode reconstruit précisément la série temporelle indépendamment de l'utilisation de l'information sur le cumul annuel de la croissance.

1 Introduction

En 2019, les prairies s'étendaient sur 12,7 millions d'hectares en France soit a peu près 44% de la surface agricole utile et 20% du territoire national. De ce fait, les prairies jouent un rôle important, notamment en assurant des services écosystémiques tels que la production de fourrages, l'atténuation du changement climatique par le stockage de carbone, le maintien de la biodiversité, etc. Le service de production fourragère assuré par les prairies est intimement lié à la façon dont l'herbe pousse et donc aux conditions spécifiques de l'année considérée. La croissance de l'herbe des prairies dépend en effet de différents facteurs (Lemaire, 2007) : les ressources en eau et en nutriments du sol, le climat (en premier lieu le rayonnement utile à la photosynthèse, la température qui régit le fonctionnement des plantes et les pluies), la gestion appliquée par l'éleveur (fauche, pâturage, fertilisation) ou encore le type de prairie.

L'information sur la croissance de l'herbe au cours de l'année est essentielle à certains modèles de simulation des troupeaux bovins. Ces modèles simulent l'utilisation de cette ressource herbagère pour la production de fourrages conservés ou pour l'alimentation des animaux au

pâturage. Mais l'information sur la croissance de l'herbe n'est malheureusement pas facilement accessible. Le défi est donc de proposer un outil qui estime cette croissance au cours de l'année, par exemple par période de 10 jours (décades).

Les prairies sont exploitées par les éleveurs pour nourrir les herbivores qu'ils élèvent. Elle peuvent être pâturées et/ou fauchées pour produire des fourrages conservés de type foin ou ensilage d'herbe. L'ensemble de l'herbe pâturée ainsi que de l'herbe fauchée dans l'année correspond à ce que l'on appelle la *valorisation annuelle* de la prairie (exprimée en tonnes de matière sèche à l'hectare et à l'année). L'information sur la valorisation annuelle des prairies françaises est disponible dans les statistiques agricoles françaises. Dans ce travail, nous avons fait l'hypothèse qu'elle permettait d'estimer le *cumul annuel de la croissance* de l'herbe (c'est-à-dire, de la pousse totale de l'herbe sur une année), et donc nous faisons l'hypothèse que le cumul annuel de la croissance est disponible. Nous reviendrons dans les discussions des résultats et la conclusion sur l'hypothèse de disponibilité de cette information sur le cumul de croissance.

Avec cette hypothèse, notre problématique est alors de désagréger l'information annuelle de croissance de l'herbe pour reconstruire la croissance au cours de l'année (par décades). Pour cela on dispose de l'information sur le climat. Autrement dit, ayant des séries temporelles décrivant le climat de l'année et sachant la croissante totale de l'herbe, on souhaite reconstruire la série temporelle de l'évolution intra-annuelle de la croissance.

Nous proposons d'aborder ce problème de désagrégation d'une information annuelle comme une tâche de prévision d'une série temporelle. Notre problématique n'est pas de prévoir des valeurs futures de la croissance de l'herbe, mais les méthodes de prévision de séries temporelles permettent d'estimer des valeurs d'une série temporelle, notamment en fonction de variables exogènes, telles que celles du climat. L'objectif est donc d'adapter ces méthodes pour notre problématique.

Dans la suite de l'article, nous présentons tout d'abord les données utilisées dans cette étude. La Section 3 présente ensuite la formalisation du problème ainsi que les différentes méthodes proposées. Ces méthodes sont ensuite évaluées dans la Section 4. Avant de conclure, la Section 5 positionne notre approche parmi celles de l'état de l'art.

2 Données

Pour l'apprentissage et l'évaluation d'un modèle de prévision d'une série temporelle de croissance de l'herbe, on dispose d'une base de données issue de simulations de la croissance de l'herbe par le modèle STICS (Brisson et al., 2003) [1], qui est un modèle mécaniste et déterministe. La Figure 1 illustre des courbes de croissances ayant des comportements très différents (voir légende).

Les simulations ont été réalisées à l'échelle de la France à une haute résolution spatiale correspondant à des unités pédoclimatiques (UPC), issues du croisement de la résolution de l'information climatique (maille safran) et pédologique (unité cartographique de sol ou UCS), et pour lesquelles la surface de prairies est significative. Les sorties des simulations correspondent à des séries temporelles de 30 années (1984-2013) au pas de temps journalier. À chaque UPC est associé un climat de ces 30 années, un à deux sols majoritaires, un à deux

1. Modèle STICS : `https://www6.paca.inra.fr/stics/`.

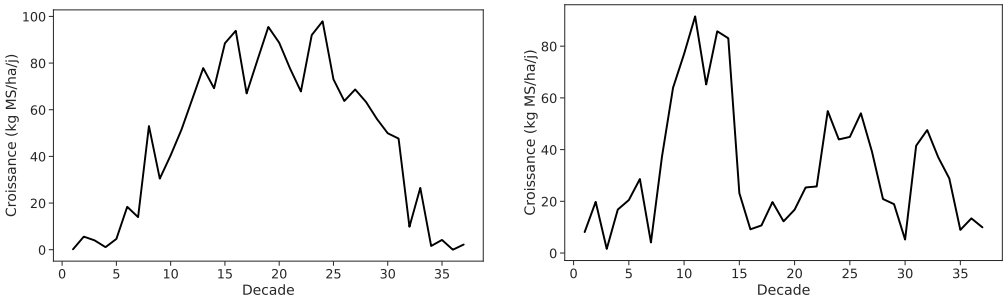

FIG. 1 – *Exemples de courbe annuelle de croissance de l'herbe pour deux UPC différentes avec des conditions climatiques différentes. Dans le cas de droite, il y a une forte diminution de la croissance en été due probablement au manque de pluie.*

TAB. 1 – *Description des différentes variables du jeu de données.*

Variable	Définition (unité)
id	Numéro de l'UPC simulée
$annee$	Année des observations
$decade$	Numéro de la décade (période de dix jours consécutifs)
$Tmin$	Température minimale (°C)
$Tmax$	Température maximale (°C)
$Tmoy$	Moyenne de $Tmin$ et de $Tmax$ (°C)
$Rain$	Cumul des pluies sur la décade (mm)
RG	Cumul du rayonnement sur la décade ($J.cm^{-2}$)
im	Indice de Martonne définit par : $I = 37 \times \frac{Rain}{Tmoy+10}$ (mm/C)
$croissance$	Croissance jounalière nette moyenne sur la décade ($kg\,MS/ha/jour$, MS signifiant matière sèche)

types de prairie majoritaires, et pour chacun des types de prairie, 1 à 18 modes d'exploitation. Au sein d'une même UPC, il peut donc y avoir plusieurs simulations (une dizaine en moyenne par UPC).

Dans le cadre de ce travail, nous nous sommes limités aux données de la région Bretagne pour les cinq dernières années (2009 à 2013). Nous avons sélectionné uniquement les variables utiles à savoir la croissance journalière des prairies ainsi que les données climatiques journalières. Les cumuls de croissance annuels sont obtenus par sommation des croissances journalières.

Les données journalières de croissance et de climat ont été agrégées à la décade (en cumul ou en moyenne, cf Table 1), c'est-à-dire sur une période de dix jours consécutifs, période pendant laquelle la croissance varie peu, et suffisante pour la précision attendue de la désagrégation. Chaque série annuelle comporte ainsi 37 valeurs. Le jeu de données se compose de 477 439 séries (Graux, 2021). Chaque série comporte 10 variables résumées dans la Table 1 : le couple $(id, annee)$ identifie une série de 37 décades, la variable de mesure de la croissance (à reconstruire) et les autres variables qui décrivent le climat (température, pluie et rayonnement). L'indice de Martonne (im) est un indice d'aridité et l'intérêt de disposer de cet indice est qu'il peut être intégrateur de l'information sur la température et les pluies (de Martonne, 1926).

3 Méthodologie

Dans cette section, on commence par présenter la formalisation du problème de désagrégation d'une série temporelle en un problème d'apprentissage d'un modèle de prévision. On présente ensuite la préparation des données pour réaliser l'apprentissage des paramètres du modèle. La Section 3.2 présente des variantes du problème de désagrégation en utilisant des représentations alternatives d'une série temporelle (série différenciée ou cumulée). Finalement, la Section 3.3 revient sur les choix des valeurs initiales de la série reconstruite.

3.1 Formalisation et approche générale

Dans notre problème de désagrégation, un exemple est représenté sous la forme $\langle C, \boldsymbol{Y} \rangle$ où $C \in \mathbb{R}$ est la valeur d'un cumul de croissance et $\boldsymbol{Y} = \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ est une série temporelle multivariée de taille n où $\boldsymbol{y}_t \in \mathbb{R}^k$ pour tout $t \in [1, n]$. $\boldsymbol{Y}$ représente les données climatiques de l'année considérée.

L'objectif de la désagrégation est de construire une série temporelle $\widehat{X} = \hat{x}_1, \ldots, \hat{x}_n$ telle que $\sum_{t=1}^{n} \hat{x}_t = C$ et que ces valeurs soient aussi proches que possible des données originales X. On cherche donc à minimiser l'erreur quadratique moyenne entre X et $\widehat{X}$.

L'approche proposée dans ce travail consiste à utiliser une technique de prévision de série temporelle pour reconstruire la série $\widehat{X}$ de proche en proche. Il s'agit donc d'estimer la valeur de $\hat{x}_t$ en fonction des valeurs précédentes de la série, mais également des valeurs des séries exogènes. Si on considère un modèle auto-régressif d'ordre p ($AR(p)$), la valeur $\hat{x}_t$ s'obtient en utilisant l'équation suivante :

$$\hat{x}_t = c + \sum_{i=1}^{p} \varphi_i \hat{x}_{t-i} + \sum_{j=0}^{p} \boldsymbol{\psi}_j . \boldsymbol{y}_{t-j} = f_{(\varphi,\boldsymbol{\psi})} \left(\hat{x}_{t-1,\ldots,t-p}, \boldsymbol{y}_{t,\ldots,t-p} \right) \tag{1}$$

où $\varphi_i \in \mathbb{R}$ et $\boldsymbol{\psi}_j \in \mathbb{R}^k$ sont les paramètres du modèle. L'ordre du modèle, p, désigne le nombre de valeurs dans le passé qui sont prises en compte pour la prédiction.

On peut noter que l'équation 1 ne tient compte que des valeurs strictement avant t pour prédire $\hat{x}_t$, mais les valeurs exogènes étant connues, on peut également prendre en compte les données de $\boldsymbol{Y}$ à la date t. Plus généralement, on souhaite estimer $\hat{x}_t$ en fonction de $\hat{x}_{t-1}, \ldots, \hat{x}_{t-p}$, et on note f_θ cette fonction d'estimation où θ représente les paramètres de cette fonction.

Notre problème de désagrégation est donc vu comme un problème d'apprentissage automatique pour lequel il s'agit d'estimer les paramètres θ d'un modèle de prévision f_θ à partir des données d'apprentissage. La reconstruction d'une série temporelle $\widehat{X}$ se fait alors en appliquant de manière récursive le modèle de prévision. Néanmoins, il sera nécessaire d'ajouter des hypothèses sur les valeurs initiales de la série temporelle pour appliquer une première fois le modèle de prévision (voir Section 3.3).

3.2 Différents pré-traitements des données

Deux prétraitements ont été proposés comme alternatives à l'utilisation des données de la série temporelle : le calcul de la série « différenciée » et le calcul de la série « cumulée ». Dans les deux cas, on transforme la série X sans modifier les séries exogènes ($\boldsymbol{Y}$).

TAB. 2 – *Initialisations des 3 premières valeurs pour la reconstruction d'une série* $X = \langle x_1, \ldots, x_n \rangle$.

	Type de reconstruction		
Initialisation	Brute	Différenciée	Cumulée
Connue	x_1, x_2, x_3	$x_1, x_2 - x_1, x_3 - x_2$	$x_1, x_1 + x_2, x_1 + x_2 + x_3$
Moyenne	$9, 9, 9$	$0, 0, 0$	$9, 18, 27$

Le calcul de la série différenciée est effectué par la soustraction de la valeur de la croissance de la décade $d-1$ à la valeur de la croissance de la décade d. Le problème d'apprentissage consiste alors à être en mesure de reconstruire fidèlement la dérivée de la croissance, plutôt que la croissance elle-même. Pour la reconstruction de la croissance, il faut d'abord prédire la valeur de la série différenciée puis intégrer la série (cumul avec la valeur précédente) pour reconstruire la courbe de croissance. L'intérêt de cette approche est de laisser libre le choix de l'initialisation de la croissance lors de l'étape d'intégration. On peut ainsi reconstruire la croissance en garantissant la valeur du cumul C.

La valeur de la série cumulée correspond à la somme de toutes les valeurs des décades précédentes. C'est le cheminement inverse de la série différenciée. Notre problème de prévision devient ainsi de reconstruire fidèlement cette croissance partielle cumulée. Lors de la phase de reconstruction, le modèle de prévision appris reconstruit la série cumulée qui est ensuite dérivée pour obtenir la série de la croissance. L'intérêt de la méthode « cumulée » est potentiellement de rendre le problème d'apprentissage plus simple puisque la fonction à prédire est croissante.

Finalement, dans le cas où il n'y a pas de pré-traitements, on évalue aussi un traitement de remise à l'échelle de la prédiction pour faire coïncider celle-ci avec l'information sur le cumul de croissance C. Ce traitement consiste à appliquer le facteur $\frac{C}{\sum_{i=1}^{n} \hat{x}_i}$ à toutes les valeurs de la série obtenue par le modèle de prévision.

3.3 Ordre des modèles et initialisation

Il faut maintenant fixer l'ordre p de la fonction de prévision, c'est-à-dire le nombre de valeurs passées à prendre en compte pour prédire la valeur suivante. Il faut choisir ce nombre de telle sorte qu'il soit suffisamment grand pour donner une prédiction précise et en même temps relativement faible, car si on apprend sur p décades cela signifie qu'on ne peut pas prédire la croissance des p premières décades de l'année. Il faut donc optimiser ce nombre pour qu'il remplisse ces deux conditions. Nous avons défini à dire d'expert ce nombre à $p = 3$ décades, ce qui correspond approximativement à un mois (30 jours). Ce chiffre semble être un bon compromis entre avoir suffisamment d'informations antérieures pour un bon apprentissage et de limiter l'absence de prédiction du début d'année. En fixant ce nombre à 3, on ne prédit donc pas la croissance du mois de janvier, mois pendant lequel la croissance est faible et peu variable.

Deux initialisations différentes de la croissance sont possibles utilisant soit les valeurs originales de la série, soit une valeur constante [2]. Dans ce second cas, on a choisi de fixer cette

2. L'utilisation d'une valeur constante évite d'être dépendant d'une information dont l'utilisateur final n'est pas censé disposer.

constante à 9 kg MS/ha/jour, valeur qui correspond à la moyenne de la croissance connue sur chacune des trois premières décades et pour l'ensemble des séries disponibles. Le premier type d'initialisation nécessite une information dont on ne dispose pas en réalité, on aimerait donc que le second type d'initialisation donne des résultats aussi précis.

Pour les séries cumulées et différenciées, les valeurs d'initialisation se déduisent des hypothèses précédentes. Les choix pour les initialisations sont résumés dans le Tableau 2.

4 Expérimentations et résultats

L'ensemble des traitements et des analyses ont été implantés en Python. Trois modèles de régression ont été explorés : la régression linéaire (*lm*), la Support Vector Regression (SVR) (Awad et Khanna, 2015) (noyau Gaussien, configuré avec $C = 100$) et les forêts aléatoires (Breiman, 2001) (limité à 100 arbres). Devant la quantité de données, l'apprentissage des forêts aléatoires et des SVR, a nécessité de sous échantillonner aléatoirement le jeu d'apprentissage (utilisation de seulement 10% à 15% du jeu total).

En combinant les différents types de régression et prétraitements on obtient neuf modèles de prévision de la croissance différents :

— trois pré-traitements (variantes de série temporelle à reconstruire) : brute, différenciée (*diff*) et cumulée (*cumul*) ;
— trois modèles de régression : régression linéaire (*lm*), SVR et forêts aléatoires (*RF*) ;

Ces modèles sont combinés ensuite avec différentes possibilités pour la reconstruction :

— deux initialisations : initialisation connue ou moyenne (à $9\,kg\ MS/ha/j$) des trois premières valeurs.
— utilisation de deux post-traitements pour faire correspondre à la valeur du cumul annuel prédit à celui connu : facteur d'échelle (*scale*) ou translation (*trans*).

Chacun de ces modèles a été appris sur le jeu d'apprentissage, puis a été évalué sur les données du jeu de test. Le jeu d'apprentissage représente 70% du jeu original (30% pour le jeu de test). Dans la mesure où on souhaite disposer d'un modèle qui généralise bien pour des données exogènes différentes, nous avons choisi de découper le jeu de données selon le climat : les séries du jeu de test et du jeu d'apprentissage ont des climats (séries exogènes) qui sont distincts. [3]

Pour estimer la précision de la reconstruction d'une série temporelle, nous avons calculé la racine de l'erreur quadratique moyenne (RMSE). On cherche alors à déterminer le modèle qui a la RMSE la plus petite. Dans le cas de valeurs de X positives, la RMSE est également un indicateur direct de l'erreur commise sur le cumul C.

À titre comparatif, nous avons commencé par calculer la RMSE moyenne commise dans le cas de l'utilisation de l'estimation moyenne de la croissance. Pour ce modèle naïf, on calcule la moyenne des croissances de chaque décade sur toutes les données. On obtient ainsi une courbe de croissance moyenne utilisée pour toutes les désagrégations. L'erreur moyenne de ce modèle naïf est de $20.6\,kg\ MS/ha/j$.

Dans la suite de cette section, on commence par donner les résultats des comparaisons de nos 9 modèles avec une initialisation moyenne. Dans un second temps, on analyse l'erreur

3. On dispose pour cela d'une information complémentaire sur les prairies de la simulation : la maille *safran*. Il y a 469 mailles *safran* différentes dans le jeu de données, les séries issues des simulations de 70% des mailles servent à l'apprentissage, et le reste des séries sert au test.

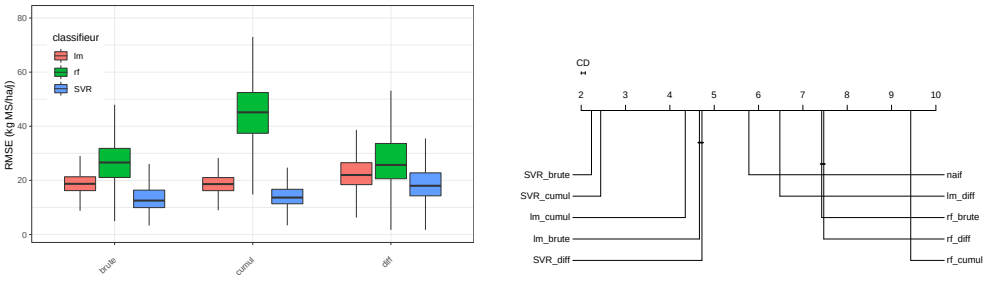

FIG. 2 – *Comparaison des RMSE entre les différents classifieurs et pré-traitements : à gauche, distribution des erreurs, à droite, diagramme de différences critiques.*

commise par le modèle du fait du choix de l'initialisation, puis on analyse l'apport de l'utilisation de l'information du cumul C sur la qualité des prédictions. Finalement, la Section 4.4 illustre qualitativement les résultats de la désagrégation.

4.1 Comparaison des 10 modèles de prévision

Dans cette partie, on compare les neuf modèles de prévision (trois pré-traitements de données et 3 types de régresseur), ainsi que le modèle naïf. Ils sont comparés ici dans la situation d'une initialisation avec les valeurs moyennes et sans ajustement post-prévision.

La Figure 2 sur la gauche représente les résultats de RMSE obtenus. Pour les classifieurs RF, SVR et lm (tous pré-traitements confondus), les RMSE moyennes sont respectivement de $35.3\,kg\ MS/ha/j$, $16.8\,kg\ MS/ha/j$ et $20.7\,kg\ MS/ha/j$ pour les séries brutes. On constate qu'on obtient de meilleurs résultats avec SVR. Le modèle par forêts aléatoires montre étonnement de mauvaises performances par rapport aux autres approches. En comparaison avec le modèle moyen, seul le SVR semble réellement faire mieux en moyenne.

On constate ensuite que les versions avec cumul ont des RMSE plutôt réduites pour les modèles linéaires (lm) et SVR, en revanche les performances des RF empirent dans ce cas.

La Figure 2 sur la droite illustre ces résultats de manière synthétique par un diagramme de différences critiques. Le diagramme confirme les remarques précédentes : ce sont les approches avec SVR qui sont significativement meilleures. Les solutions basées sur le cumul sont également plutôt intéressantes (sauf combinées avec RF). On constate que la plupart des modèles basés sur les modèles de régression linéaire et de SVR sont significativement meilleurs que le modèle moyen naïf. La comparaison se base ici sur des différences pair à pair pour chaque série à désagréger (test de Nemenyi avec $\alpha = 5\%$) et donne une conclusion différente de la comparaison des moyennes seules.

4.2 Effet de l'approximation des premières valeurs

On regarde maintenant les erreurs liées à l'approximation des premières décades pour savoir si l'approximation par la constante à un impact sur la précision. Pour cela, on s'intéresse aux ratios de MSE avec et sans approximation pour chacune des configurations. On se place dans la situation où il n'y a pas de post-traitement.

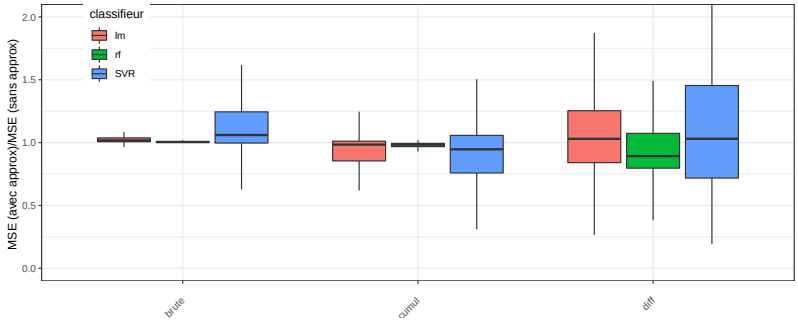

FIG. 3 – *Ratio de la MSE avec initialisation moyenne sur la MSE avec initialisation connue. Plus le ratio est faible et moins l'utilisation de l'approximation a de l'effet sur l'erreur. Une valeur inférieure à 1 indique une amélioration de l'erreur.*

On reprend alors le même organisation de graphique que précédemment mais, à la place des RMSE, la Figure 3 visualise les ratios de MSE avec initialisation moyenne et initialisation connue. Plus ce ratio est grand, et plus l'initialisation moyenne apporte de l'erreur par rapport à l'utilisation des données originales. On cherche donc à avoir les plus petits ratios possibles.

On constate qu'en moyenne les ratios sont très proches de 1. Ceci signifie que l'utilisation de l'initialisation moyenne n'introduit pas d'erreur importante. Néanmoins, on constate une forte présence de valeurs aberrantes, en particulier lors de l'utilisation du cumul ou de la série différenciée. Ces valeurs aberrantes sont également observées sur l'utilisation du SVR sur les séries brutes.

Finalement, on peut également constater que dans certains cas, l'utilisation de l'initialisation connue améliore les résultats (ratio < 1).

4.3 Apport de l'utilisation d'un post-traitement

On regarde maintenant l'apport d'un post-traitement sur notre approche. Pour rappel de la Section 3.2, nous avons la possibilité de faire des post-traitements par une mise à l'échelle (quelque soit le pré-traitement de la série) et, dans le cas de la série différentiée, il est possible d'utiliser une translation (choix de la valeur initiale lors de l'intégration).

On commence tout d'abord par regarder l'apport de l'application du facteur d'échelle. On constate sur le graphique de la Figure 4 que l'utilisation de ce post-traitement améliore les MSE dans la plupart des configurations. À la médiane, pour le SVR appliqué sans pré-traitement, l'amélioration est de $\approx 84\%$. La Figure 6 illustre qualitativement cette amélioration. L'utilisation des RF avec le cumul continue à donner des résultats à l'inverse des autres sans explication particulière.

On complète finalement l'analyse avec la comparaison dans le cas du pré-traitement de différenciation avec les trois alternatives de post-traitements : translation, facteur d'échelle ou sans post-traitement. Les résultats sont donnés dans la Figure 5.

Par rapport au graphique précédent, cette figure ajoute le cas de la transformation avec translation. On constate alors que les performances de la mise à l'échelle sont semblables à celles de la translation. De plus, on a constaté que la translation peut induire des valeurs de

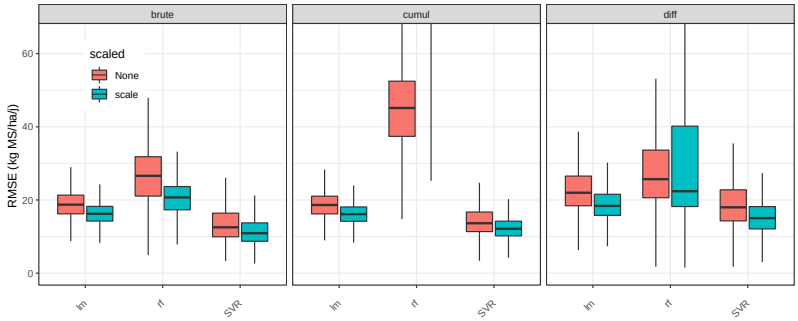

FIG. 4 – *Comparaison des RMSE avec et sans l'utilisation d'un facteur d'échelle en post-traitement (utilisation du l'information de cumul C).*

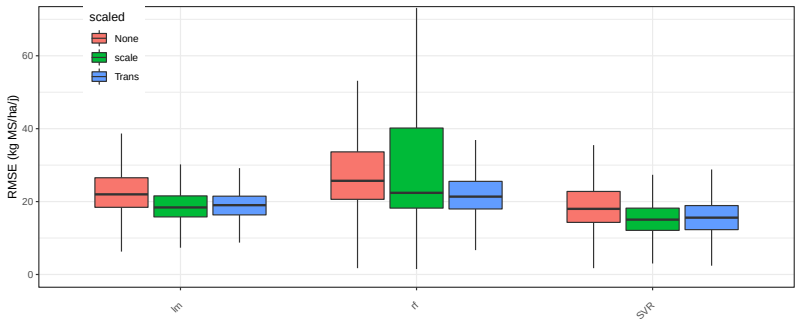

FIG. 5 – *Comparaison des RMSE en fonction des post-traitements appliqué sur la série différentiée : sans post-traitement, mise à l'échelle ou translation de la courbe (utilisation du l'information de cumul C).*

croissance négatives. Ces valeurs ne sont pas acceptables du point de vue biologique. C'est une situation que la mise à l'échelle ne provoque pas.

Par conséquent, bien qu'attrayant d'un point de vue méthodologique, la solution avec translation se montre peu intéressante en pratique. En revanche, l'utilisation pratique de la mise à l'échelle tend à montrer un intérêt pour l'amélioration de la désagrégation en utilisant la connaissance du cumul.

À l'issue de ces expérimentations, on peut conclure que la meilleure solution de désagrégation est celle basée sur un classifieur SVR, sans pré-traitement de la série de croissance. L'erreur moyenne de ce modèle est de $12.4 \pm 4.3\, kg\ MS/ha/j$. Cette solution de désagrégation peut être utilisée avec une initialisation moyenne des trois premières décades : cette approximation est validée expérimentalement. L'utilisation de l'information de cumul se montre intéressante pour réduire la RMSE, mais l'amélioration reste faible. On rappelle que cette information de cumul annuel n'est pas directement accessible : à terme, on peut envisager de l'estimer à partir de l'information de valorisation annuelle, mais l'erreur qui pourrait être commise par cette estimation pourrait faire perdre le bénéfice de l'utilisation de cette information.

Il faut noter que si le SVR requiert un long temps de calcul (plusieurs heures en apprentis-

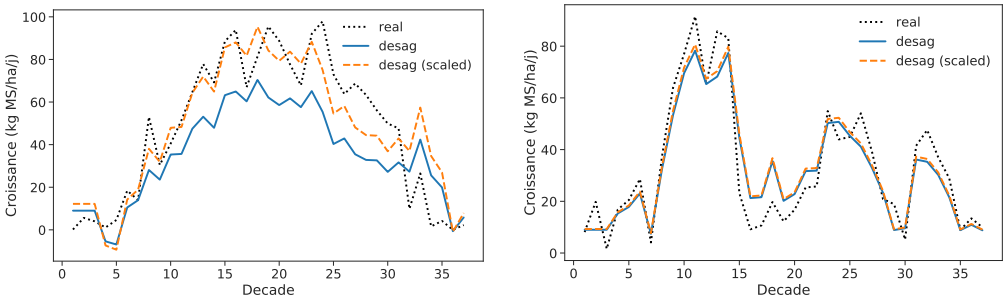

FIG. 6 – *Exemples de désagrégation avec le modèle linéaire (sur la gauche) et le modèle SVR (sur la droite, RMSE :* 6.9 $kg\,MS/ha/j$*). L'exemple de gauche illustre l'effet du post-traitement (RMSE avant :* 17.5 $kg\,MS/ha/j$*, après :* 11.1 $kg\,MS/ha/j$*).*

sage et une dizaine de secondes en inférence). L'apprentissage d'un modèle linéaire nécessite seulement quelques secondes.

4.4 Évaluation qualitative des reconstructions

La Figure 6 illustre les résultats de désagrégation obtenus pour les deux séries exemples de l'introduction (voir Figure 1). Dans les deux cas, les traitements ont été menées sans pré-traitement et avec une initialisation moyenne (on constate que les trois premières valeurs sont fixes). Cette figure illustre le résultat obtenu avec un modèle linéaire (graphique de gauche) qui sous estime la croissance en été. Le facteur d'échelle corrige ce défaut, mais induit des erreurs supplémentaires en hiver. Sur le graphique de droite, on constate que la courbe est très bien reconstruite malgré certains changements importants. Sur cette courbe, le modèle linéaire ne s'adapte pas bien à ces changements.

5 Travaux connexes

Le terme de *désagrégation* correspond parfois à une tâche de traitement du signal qui vise à séparer les signaux de différentes sources et qui sont mélangés dans un même signal, par exemple dans un signal de consommation électrique (Figueiredo et al., 2012). Ce type de problème ne correspond pas au nôtre.

Pour des séries temporelles, la *désagrégation temporelle* désigne des méthodes pour reconstruire une série temporelle à haute fréquence cohérente avec une série temporelle à plus basse fréquence (la somme ou la moyenne des valeurs de la série à haute fréquence). Il s'agit d'une sorte de sur-échantillonnage. Elle est utilisée en économétrie pour affiner des estimations annuelles, mensuelles ou trimestrielles d'indicateurs à des échelles temporelles plus fines (Moauro et Savio, 2005). La méthode brute dans ce domaine est l'algorithme de Chow et Lin (1971) dont de nombreuses variantes ont été proposées. Néanmoins, dans cette approche du problème, il est nécessaire d'avoir des données régulières sur la croissance de l'herbe. Hors, nous ne disposons que d'une seule valeur agrégée dans notre cas. L'application de ce type de problème risque donc d'être peu performante.

Le problème de désagrégation peut aussi se voir comme un cas particulier d'un problème plus général : celui d'estimer des valeurs individuelles à partir de données agrégées. Ce problème est également fréquemment rencontré pour des applications économiques où des données statistiques (par exemple, des statistiques de vote) ne sont connues que pour des groupes de personnes, et non individuellement. Mathématiquement, il s'agit de reconstruire une distribution jointe à partir de distributions marginales. Pour notre application, la marginale serait le cumul et on rechercherait à reconstruire la distribution pour les différentes décades. Ce type de problème peut être résolu par exemple par des algorithmes d'*Iterative Proportional Fitting* (IPF) (Deming et Stephan, 1940) ou d'inférence écologique [4] (King et Fox, 1999). Pour cette dernière approche, Quinn (2004) propose une variante tenant compte de dépendances temporelles entre les valeurs en faisant l'hypothèse sur la dynamique. Néanmoins, les hypothèses sur ces dynamiques peuvent être difficiles à proposer, aussi, nous avons préféré une approche centrée uniquement sur les données.

Finalement, le domaine de l'apprentissage automatique s'est également intéressé à ce type de problématique, notamment en proposant des alternatives à l'IPF en se basant sur les formes normales de Sinkhorn (Idel, 2016). Cette formulation rapproche le problème IPF de celui du transport optimal. Néanmoins, il ne tient pas compte des spécificités de données sous forme de séries temporelles.

6 Conclusion

Nous nous sommes intéressés au problème de désagrégation temporelle du cumul annuel de croissance de l'herbe. Pour cela, nous avons transformé ce problème en un problème de prévision d'une série temporelle, s'aidant de données exogènes disponibles. Au travers de nos expérimentations, nous avons pu identifier une méthode qui offre de bonnes performances. Cette méthode se base sur l'apprentissage d'un modèle SVR pour prédire la valeur de la croissance à une date donnée en fonction des données des trois dates précédentes et des données climatiques. Il est important de noter que c'est la présence de ces données climatiques, choisies pour leur lien fort avec la croissance de l'herbe qui explique la qualité de nos reconstructions. Ainsi, les expérimentations ont montré que l'initialisation de la croissance peut être faite à l'aide des valeurs moyennes des trois premières décades sans dégrader les résultats.

Finalement, les expérimentations ont montré que l'information de cumul annuel de la croissance améliore la précision de la désagrégation. Néanmoins, cette amélioration est faible et la disponibilité d'une information fiable de ce cumul annuel n'est pas garantie dans le futur. Si le problème initial était bien de désagréger cette quantité, on se rend compte que les résultats sont en fait très bons sans l'utiliser directement, mais en utilisant simplement l'information sur le climat. Par la suite, il ne semble pas nécessaire de poursuivre son utilisation. Le modèle de prédiction retenu alimentera un modèle de simulation d'élevage laitier nommé FARM-AQAL qui sera utilisé dans le cadre du projet européen GENTORE.

La perspective de ce travail sera d'explorer d'autres variantes. En particulier, le choix de la décade a été fixé initialement, il est probable qu'en réduisant cette période, par exemple à la semaine, les algorithmes auto-régressifs pourraient donner de meilleurs résultats, et donc globalement améliorer la désagrégation.

4. Le terme *ecological inference* ne fait pas particulièrement référence à des données écologiques.

Références

Awad, M. et R. Khanna (2015). Support vector regression. In *Efficient learning machines*, pp. 67–80. Springer.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Brisson, N., C. Gary, E. Justes, R. Roche, B. Mary, D. Ripoche, D. Zimmer, J. Sierra, P. Bertuzzi, P. Burger, et al. (2003). An overview of the crop model STICS. *European Journal of agronomy 18*(3-4), 309–332.

Chow, G. C. et A.-l. Lin (1971). Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. *The review of Economics and Statistics*, 372–375.

de Martonne, E. (1926). Aréisme et indice d'aridité. *Comptes rendus de L'Academie des Sciences 182*, 1395–1398.

Deming, W. E. et F. F. Stephan (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics 11*(4), 427 – 444.

Figueiredo, M., A. de Almeida, et B. Ribeiro (2012). Home electrical signal disaggregation for non-intrusive load monitoring (NILM) systems. *Neurocomputing 96*, 66–73.

Graux, A.-I. (2021). Growth and annual valorisation of breton grasslands simulated by STICS, and associated climate. https ://doi.org/10.15454/FD9FHU, Portail Data INRAE.

Idel, M. (2016). A review of matrix scaling and sinkhorn's normal form for matrices and positive maps. *arXiv preprint arXiv :1609.06349*.

King, G. et J. Fox (1999). A solution to the ecological inference problem : Reconstructing individual behavior from aggregate data. *Canadian Journal of Sociology 24*(1), 150.

Lemaire, G. (2007). Physiologie de la croissance de l'herbe : applications au pâturage. *Fourrages 112*(325–344), 69–78.

Moauro, F. et G. Savio (2005). Temporal disaggregation using multivariate structural time series models. *The Econometrics Journal 8*(2), 214–234.

Quinn, K. M. (2004). Ecological inference in the presence of temporal dependence. In *Ecological Inference : New Methodological Strategies*, pp. 207–233. Cambridge University Press.

Summary

Information on the grass growth over a year is essential for some models simulating the use of this grassland resource for the production of fodder or for feeding animals on pasture. Unfortunately, this information is rarely available. The challenge is to reconstruct grass growth from two sources of information: daily climate data (rainfall, radiation, etc.) and cumulative growth over the year. In this paper, we formulate this challenge as a problem of disaggregating the cumulative sum of growth into a time series. To address this problem, our method applies time series forecasting using climate information. Several alternatives of the method are proposed and compared experimentally using a database generated from a grassland simulator. The results show that our method can accurately reconstruct the time series, independently of the use of the cumulative sum of growth.