Prédiction du Rayonnement Solaire par Apprentissage Automatique

Pierrick Bruneau, Philippe Pinheiro, Yoann Didry

LIST, L-4362 Esch-sur-Alzette
prenom.nom@list.lu,
http://www.list.lu

Résumé. Cet article décrit une approche flexible pour la prédiction à court terme de variables météorologiques. En particulier, nous nous intéressons à la prédiction du rayonnement solaire à une heure. Cette tâche est d'une grande importance pratique dans l'optique d'optimiser les resources énergétiques solaires. Comme le défi EGC 2018 nous fournit des données météorologiques enregistrées sur cinq sites géographiques de l'île de la Réunion, nous utilisons ces données historiques comme base pour créer des modèles de prédiction, et nous testons la performance de ces modèles selon le site considéré. Après avoir décrit notre méthode de nettoyage de données et de normalisation, nous combinons une méthode de sélection de variables basée sur les modèles ARIMA (*AutoRegressive Integrated Moving Average*) à l'utilisation de méthodes de régression génériques, telles que les arbres de régression et les réseaux de neurones.

1 Introduction

Dans cet article, nous traitons de la prédiction des valeurs futures d'une série temporelle d'intérêt. Nous considérons un horizon de prédiction arbitraire, et un contexte multivarié, où les valeurs historiques de plusieurs séries temporelles sont disponibles en entrée. Vu le contexte particulier du Défi EGC 2018, nous nous intéressons à la prédiction du rayonnement solaire. La prédiction des valeurs futures de variables météorologiques a notamment un intérêt pratique quand il s'agit d'optimiser des sources d'énergie renouvelables (Barbounis et al., 2006).

L'approche classique à la prédiction météorologique utilise des simulations physiques initialisées par des relevés sur le terrain (Lynch, 2008). De manière alternative, dans cet article nous adoptons une approche basée sur l'apprentissage automatique, complètement agnostique de la dimension physique. Plus précisément, la tâche de prédiction est vue comme un problème de régression, avec pour variable cible le rayonnement solaire à un horizon de prédiction donné. Le vecteur d'entrée peut potentiellement utiliser l'ensemble des valeurs historiques (i.e. observées jusqu'à l'instant présent). Ce choix engendre les sous-problèmes suivants :

— *pré-traitement* : les valeurs manquantes affectent généralement les modèles d'apprentissage. L'inspection et la correction préalable des données d'apprentissage est nécessaire.

— sélection de variables : dans le contexte des séries temporelles, il faut trouver un compromis entre le modèle naif n'utilisant que les valeurs présentes comme vecteur d'entrée, et le modèle exhaustif qui utilise l'historique complet. Ce dernier contient intuitivement plus d'information, mais l'estimation de paramètres trop nombreux engendre un excès de variance et de complexité.

Après la revue des travaux existants dans la section 2, nous présentons les données du Défi EGC 2018 dans la section 3.1 [1]. Notre pré-traitement est introduit dans la section 3.2. Une procédure de sélection de variables est ensuite décrite dans la section 3.3. Des modèles de régression génériques sont ensuite entrainés avec les données résultant de ces processus selon le protocole décrit en section 4. Nos résultats expérimentaux permettent d'évaluer la performance respective de deux modèles de régression, de la procédure de sélection de variables, ainsi que de l'impact des sites géographiques où les données ont été enregistrées.

2 Travaux Existants

Considérons un ensemble de séries temporelles météorologiques, x désignant l'une d'entre elles. Ces séries sont indexées par un pas de temps t, de sorte que x_t est la valeur d'une série donnée observée à un temps donné. Le pas de temps est supposé fixe, i.e. le laps de temps entre x_t et x_{t+1} est constant pour tout t. Par commodité et cohérence vis à vis de la littérature du domaine, nous définissons la prédiction à l'horizon l pour la série temporelle x comme la tâche de prédire x_{t+l-1} connaissant les valeurs de la série jusqu'à x_{t-1}. Prédire à l'horizon 1 revient alors à prédire x_t connaissant la série jusqu'à x_{t-1}. En d'autres termes, par commodité, l'instant présent est supposé être $t-1$ dans notre article. Étant donné un ensemble de D séries temporelles $\{x_d\}_{d\in 1\dots D}$, parmi lesquelles nous isolons une série d'intérêt avec l'index $\delta \in 1\dots D$, la tâche de prédiction de $x_{\delta,t+l-1}$ étudiée a un caractère multivarié via la connaissance de $x_{d,t-1}$ pour tout d.

Les modèles ARIMA sont l'approche classique pour le traitement des séries temporelles. Ils sont optimisés selon l'horizon 1. Ils mélangent une partie auto-régressive (i.e. processus $AR(p)$) :

$$x_t = \sum_{i=1}^{p} \phi_i x_{t-i} + \sigma_t \tag{1}$$

et une moyenne mobile (i.e. processus $MA(q)$) :

$$\sigma_t = \sum_{i=1}^{q} \theta_i \sigma_{t-i} \tag{2}$$

Dans l'équation (1), σ_t est un bruit gaussien. Intuitivement, la prédiction réalisée selon l'équation (1) est une combinaison linéaire des p dernières valeurs observées (i.e. x_{t-1} à x_{t-p}), additionnée d'un résidu indépendamment distribué. La moyenne mobile permet d'introduire une dépendance entre résidus. Sous réserve de conditions formelles, les deux processus sont stationnaires. Cette propriété a pour conséquence importante que leur espérance est constante

1. Les données peuvent être téléchargées à `http://www.egc.asso.fr/news-details-1-40-Defi_EGC_2018_Un_defi_sous_le_soleil_de_lle_de_La_Reunion`

Moufia	Possession	Saint André	Saint Leu	Saint Pierre
-20.92, 55.48	-20.93, 55.33	-20.96, 55.62	-21.20, 55.30	-21.31, 55.45

TAB. 1 – *Coordonnées GPS (latitude, longitude) des sites où les données ont été enregistrées.*

selon t. ARIMA permet de combiner et d'intégrer ces processus, et ainsi d'assouplir la condition de stationarité.

Les modèles GAM (*Generalized Additive Models*) (Hastie et Tibshirani, 1990) proposent plutôt une modélisation additive de fonctions continues. Ils ont déjà été utilisés dans le contexte de l'analyse de la pollution de l'air (Dominici et al., 2002), que l'on peut juger proche de notre travail. En théorie, l'optimisation est alors réalisée selon tous les horizons de prédiction possibles. Toutefois, GAM ne fournit pas de solution pour la sélection de variables dans un contexte multivarié. En effet, la plupart des modèles d'apprentissage automatique sont basés sur un vecteur d'entrée numérique et de taille fixe.

Diverses stratégies peuvent être utilisées pour adapter des réseaux de neurones à notre tâche de prédiction multivariée. Une approche de type *divide-and-conquer* basée sur des modèles MLP (*Multi-Layer Perceptrons*) a été proposée dans (Bruneau et al., 2012). Tandis que des méthodologies de sélection de variables sont bien établies pour les modèles ARIMA (e.g. Box-Jenkins (Anderson, 1976; Box et al., 2015), Hyndman and Khandakar (Hyndman et Khandakar, 2007), parmi d'autres), aucune méthodologie équivalente ne s'est imposée pour les réseaux de neurones. Une méthode de sélection bayésienne pour les MLP est décrite dans (Bruneau et Boudet, 2012), mais elle requiert l'apprentissage avec un vecteur délibérément redondant, dans lequel la sélection est réalisée *a posteriori*.

3 Approche Proposée

3.1 Données

Les D séries temporelles définies pour le *Défi EGC 2018* sont les suivantes :
— I_D : le rayonnement solaire diffus ($W.m^{-2}$)
— I_G : le rayonnement solaire global ($W.m^{-2}$)
— *Patm* : la pression atmosphérique (hPa)
— *RH* : le taux d'humdité relatif ($\%$)
— *Text* : la température extérieure ($^\circ C$)
— *WD* : la direction du vent ($^\circ$)
— *WS* : la vitess du vent ($m.s^{-1}$)

Ces données sont fournies avec un pas temporel d'une minute, tous les horodatages spécifiant une minute exacte (i.e. secondes respectives à 0). Elles sont enregistrées pour 2 années complètes (2014 et 2015) et 5 sites géographiques, indiqués en table 1.

De manière à gérer la variété implicite aux angles dans *WD* (i.e. un angle de 350° est plus similaire à un angle de 10° que de 120°), nous prenons le cosinus et le sinus de cette dernière, formant ainsi respectivement les variables *UnitX* et *UnitY*. La prise en compte de la forte tendance journalière et saisonnière du rayonnement solaire est assurée grâce à l'utilisation du quotient entre rayonnement direct et global k_b (utilisé e.g. dans (Kylling et al., 2000)).

Cette variable construite a une interprétation intuitive (i.e. 0 pour un temps nuageux, 1 pour un temps ensoleillé), pertinente dans le contexte de panneaux photovoltaiques, notamment (Tapakis et al., 2016) :

$$k_b = \frac{I_G - I_D}{I_G} = 1 - \frac{I_D}{I_G} \tag{3}$$

Le rayonnement solaire peut également être normalisé par un modèle de rayonnement maximal théorique (Reno et al., 2012, Section 2.3). Pour une variable météorologique quelconque, des ensembles mensuels-horaires peuvent être calculés pour estimer des moyennes et variances spécifiques depuis un ensemble de données, utilisées ensuite pour la normalisation (Bruneau et al., 2012).

Dans cet article, nous nous intéressons à la prédiction horaire, i.e. prédire le rayonnement solaire une heure après un temps donné. Les séries temporelles étant fournies selon un pas de temps d'une minute, d'après la terminologie introduite en section 2, nous traitons le problème de prédire x_{t+59}.

3.2 Pré-traitement

Avant de s'intéresser à la tâche de prédiction elle-même, nous avons exploré les données fournies. Les fichiers ont été pré-traités grâce aux libraries R *zoo* (Zeileis et Grothendieck, 2005) et *lubridate* (Grolemund et Wickham, 2011), qui fournissent des outils adéquats au traitement des séries temporelles.

Nous avons tout d'abord vérifié la présence d'horodatages manquants ou erronés, et de valeurs manquantes. Tous les horodatages ont été correctement identifiés (i.e. bon format et secondes respectives à 0), mais des valeurs et horodatages manquants ont été trouvés. Par exemple, dans les données de *Moufia*, il n'y a pas de données pour les horodatages allant de `2014-01-22 08:33:00` à `2014-01-22 08:57:00`. Quand l'horodatage est présent, toutes les variables sont presque toujours renseignées, à l'exception notable des données *Saint André*, où seule I_G manque entre `2014-01-05 08:59:00` et `2014-01-05 14:59:00`.

Les fichiers de données ont été complétés avec les horodatages manquants, et les valeurs manquantes ont été interpolées linéairement. Quand la plage d'interpolation est trop grande, des artéfacts visuels indésirables résultent de cette procédure (i.e. lignes droites, ou sinusoïdes dans le cas de *UnitX* and *UnitY*). Les plages temporelles associées doivent alors être exclues de l'analyse pour le site respectif. Toutefois, quand une seule valeur manque, cet artéfact est imperceptible. Afin d'implémenter cette identification visuelle, nous avons adapté un outil de visualisation de séries temporelles ((Anderson, 2012), voir Figure 1), et identifié les périodes d'exclusion visuellement. Ces dernières sont indiquées en table 2.

Toutes les périodes d'exclusion résultent de valeurs et horodatages manquants, sauf le début des données *Moufia*. Le processus stochastique que suivent ces dernières semble ainsi différer significativement des processus de rayonnement habituels (voir figure 1a). L'exploration visuelle nous a amené à constater qu'au contraire de toutes les autres variables, les séries de rayonnement ont un fort *a priori* la nuit. Les valeurs de rayonnement nocturne sont ainsi très proches de 0, et l'incertitude des capteurs à ce niveau engendre des valeurs de k_b aberrantes. Nous corrigeons ce problème en fixant :

$$k_b = 0.5 + e, \text{ with } e \sim \mathcal{N}(0, \sigma^2) \tag{4}$$

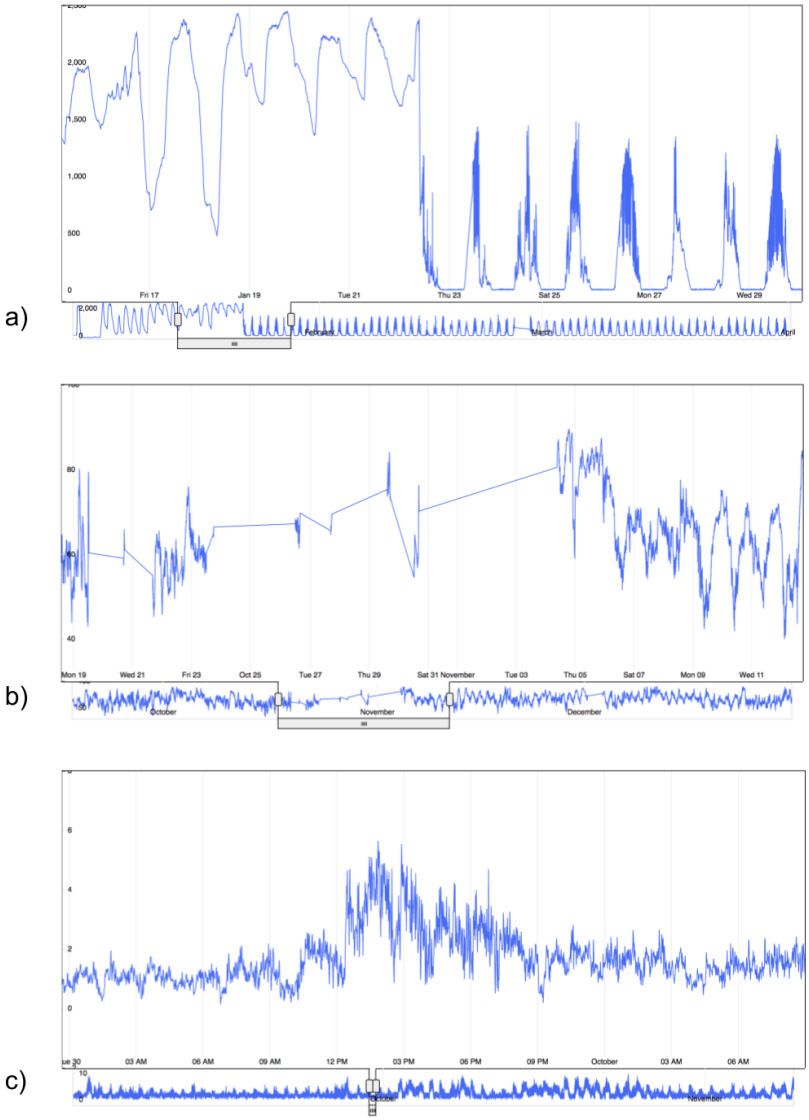

FIG. 1 – *Notre outil* focus+context *d'inspection de séries temporelles.* a) *Le début des séries* I_D *et* I_G *pour* Moufia *suivent un processus stochastique atypique.* b) *Les lignes droites d'interpolation facilient l'identification visuelle des périodes à exclure, ici de la variable* RH *pour* Moufia. c) *3 horodatages manquants le* `2014-09-30` *à* Possession *sont imperceptibles visuellement.*

Site	Périodes d'exclusion
Moufia	2014-01-01 à 2014-01-23 2014-02-25 à 2014-02-28 2015-10-19 à 2015-11-05 2015-12-01 à 2015-12-04
Saint André	2014-01-01 à 2014-03-29

TAB. 2 – *Périodes d'exclusion pour chaque site. Par simplicité et sécurité, les périodes ont été arrondies au jour englobant le plus proche. Les sites absents de la table n'ont pas de période d'exclusion.*

en période nocturne, controllant ainsi la variance des valeurs nocturnes, et limitant ainsi leur influence sur la modélisation. Les horodatages nocturnes sont déterminés en utilisant des éphémérides disponibles publiquement pour l'île de la Réunion (Bruneau et Pinheiro, 2017).

3.3 Sélection de Variables

Selon les définitions en section 3.1, les prédictions horaires dans le contexte de données ayant un pas de temps d'une minute revient à prédire $x_{\delta,t+59}$ avec la connaissance des séries temporelles x_d jusqu'à l'horodatage $t-1$. Avant même de discuter de la fonction de prédiction utilisée, un dilemme se présente au sujet de son espace d'entrée :

— utiliser seulement les valeurs instantanées des D séries temporelles à $t-1$,

— en cas d'ajout de créneaux temporels $t' < t-1$, choisir une taille de vecteur d'entrée.

L'importance de ce dilemme tient à ce que la taille du vecteur d'entrée influence fortement le temps de calcul et les besoins en termes de taille d'ensemble d'apprentissage. Dans le cas du MLP, une étape d'algorithme d'apprentissage est quadratique selon le nombre de dimensions du vecteur d'entrée (Bruneau et al., 2012), au mieux linéaire si un algorithme SGD (*Stochastic Gradient Descent*) est utilisé (Zhang, 2004). En d'autres termes, le vecteur d'entrée doit contenir suffisamment d'information de manière à assurer une bonne prédiction, mais être aussi parcimonieux que possible afin d'éviter une consommation de mémoire et un temps de calcul excessifs.

Comme indiqué dans la section 2, des méthodologies sont bien établies pour la sélection de variables dans les modèles ARIMA. Nous choisissons d'effectuer cette sélection de variables sur chacune des variables météorologiques. Nous utilisons la méthode itérative proposée par (Hyndman et Khandakar, 2007), et implémentée dans la librairie R *forecast*, conduisant à sélectionner un ensemble de créneaux passés pertinents pour la prédiction à l'horizon 1. En principe, cette sélection n'est pas adaptée à notre tâche de prédiction à l'horizon 60. Nous formons cependant l'hypothèse que cette sélection univariée fournit un vecteur d'entrée utile malgré tout à l'horizon 60 pour une méthode de régression générique. Implicitement, la modélisation de covariances entre les variables météorogologiques est déléguée à la procédure d'apprentissage. Nous réalisons cette sélection grâce aux données de *Possession* pour l'année 2014. Ce choix, qui demeure abritraire, est motivé par l'absence d'intervalle d'exclusion pour ce site. Alternativement, il aurait été possible de réaliser une sélection spécifique à chaque site, mais

l'application d'un modèle utilisant cette sélection aux données d'un autre site créerait alors un nouveau problème méthodologique.

En pratique, l'estimation de modèles ARIMA d'ordre supérieur à 30 (i.e. valeur de p ou q des équations (1) et (2)) est coûteuse en temps de calcul. Les variables météorologiques ont une saisonnalité de 24 heures *a priori* très marquée : une approche naïve requerrait alors des modèles ARIMA d'ordre au moins égal à 1440, ce qui est impossible en pratique. Pour contourner ce problème, nous avons extrait des séries à pas de temps horaire des données originales (i.e. horodatages ayant leur minutes et secondes à 0). Nous avons estimé des modèles ARIMA sur ces séries construites, alors optimisés pour la prédiction de x_{t+59} avec la connaissance de x_{t-1}, x_{t-61}, etc... Par commodité nous définissons un index horaire T lié à la périodicité horaire, t étant alors lié à une période d'une minute. La conversion de ces index est illustrée en figure 2. Nous soulignons qu'avec ce système, x_{t-1} désigne la même valeur que x_{T-1}, et que l'horizon à une heure est indiqué de manière équivalente par x_{t+59} et x_T.

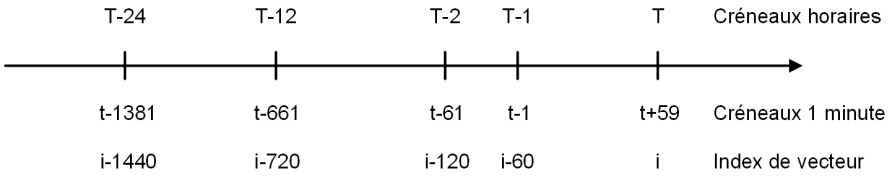

FIG. 2 – *Conversion entre horodatages ayant pour période une minute ou une heure. Comme dans un contexte programmatique l'horizon de prédiction est souvent choisi comme horodatage de référence, la conversion avec un index de tableau est également indiquée.*

Les sélections automatiques réalisées selon (Hyndman et Khandakar, 2007) sont vérifiées à l'aide de graphes d'auto-correlation. Des créneaux passés ont été ajoutés à la sélection initiale jusqu'à avoir un graphe d'auto-correlation satisfaisant. Supposant que cette procédure engendre un modèle ARIMA(p, d, q), nous inférons la sélection de $\max(p, q)$ créneaux temporels (i.e. $x_{t-1} \dots x_{t-\max(p,q)}$ sont concaténés au vecteur d'entrée pour l'apprentissage). Pour limiter le temps de calcul, $\max(p, q)$ est limité à 20. La sélection pour le pas de temps d'une minute, d'une heure, ainsi que les termes saisonniers sont concaténés. Les résultats pour chaque variable météorologique sont résumés en table 3. Nous notons que toutes les sélections sont relativement parcimonieuses, sauf pour *Patm*. Le vecteur résultant est de taille 70. Tous les modèles ARIMA ont utilisé une version différenciée des séries temporelles. Cela signifie que les données originales ne vérifiaient pas l'hypothèse de stationarité (i.e. moyenne est variance constante). Une manière simple d'établir la stationarité faible est de normaliser les séries temporelles selon des moyennes mensuelles-horaires (Bruneau et al., 2012).

4 Expériences

4.1 Protocole

Comme décrit en section 3.3 et résumé par la figure 2, pour un temps présent $T - 1$ notre protocole se concentre sur la prédiction de k_b au temps T. En particulier, nous allons tester les

Variable	Créneaux sélectionnés
k_b	$t-1, t-2, T-2, T-3, T-24$
Patm	$t-1, \ldots, t-20, T-2, \ldots, T-6, T-12, T-24$
RH	$t-1, \ldots, t-5, T-24$
Text	$t-1, t-2, t-3, T-12, T-24$
WS_Mean	$t-1, t-2, t-3, T-2, T-3, T-12, T-24$
UnitX	$t-1, \ldots, t-5, T-2, T-3, T-4, T-12, T-24$
UnitY	$t-1, \ldots, t-6, T-2, T-3, T-12, T-24$

TAB. 3 – *Créneaux temporels sélectionnés par ARIMA pour chaque variable météorologique. NB : comme* $T-1$ *et* $t-1$ *sont équivalents, seul* $t-1$ *est indiqué quand opportun.*

hypothèses suivantes :

— si la sélection de variables parmi les créneaux $t^* < t-1$ offre une meilleure performance que l'utilisation des valeurs à $t-1$ seules,

— si un modèle appris sur les données d'un site réalise de meilleures performance sur ses données de test respectives qu'un modèle appris sur les données d'un autre site.

Nous désignons désormais le vecteur formé des valeurs des 7 séries temporelles à $t-1$ comme le vecteur *instantané*, et le vecteur résultant de la procédure décrite en section 3.3 comme le vecteur *arima*. Pour tester nos hypothèses, pour chaque site nous utilisons les données de 2014 pour l'apprentissage, et celles de 2015 pour le test. De ces ensembles, nous excluons les éléments associés à la prédiction, triviale, d'un créneau nocturne. Nous utilisons les deux modèles de régression suivants, avec leurs algorithmes d'appentissage associés :

— *Xgboost* : ce modèle est basé sur le *Gradient Boosting* (Friedman, 2001) et les *Generalized Boosted Models* (Ridgeway, 2007). Nous l'avons implémenté grâce à la librairie R *Xgboost*, qui réalise des ensembles d'arbres de régression. L'algorithme utilise plusieurs hyper-paramètres, tels que la vitesse d'apprentissage η, ou la profondeur maximale d'un arbre. Ces paramètres ont été réglés au moyen de la librarie R *caret* avec une validation croisée à 3 ensembles. La totalité de l'ensemble d'apprentissage est utilisée de manière séquentielle par *Xgboost*. Aucune normalisation de données n'a été utilisée dans les résultats présentés pour *Xgboost*, car empiriquement une meilleure performance est alors obtenue.

— *MLP* : ce type de réseau de neurones utilise une seule couche cachée. Nous l'avons implémenté grâce à la librairie R *mxnet* (Chen et al., 2015). La complexité (i.e. le nombre de neurones de la couche cachée) des modèles spécifiques aux vecteurs *instantané* et *arima* est sélectionnée grâce à une validation croisée à 10 ensembles sur les données du site *Possession*. Les meilleures tailles de couches cachées ont été déterminées à 10 et 30, respectivement pour les vecteurs *instantané* et *arima*. Des MLP utilisant ces tailles ont ensuite été optimisés pour chaque site, toujours avec une validation croisée à 10 ensembles. Les réseaux de neurones étant sensibles à la normalisation des données en entrée, les séries temporelles utilisées pour la construction des vecteurs *instantané* et *arima* ont été normalisées grâces aux ensembles mensuels-horaires décrits dans (Bruneau et al., 2012). Les prédictions sont ensuite réalisées en moyennant le résultat des 5 meilleurs modèles en termes d'erreur de validation.

			Moufia	Possession	Saint André	Saint Leu	Saint Pierre
instant	RMSE	Xgboost	0.268	0.266	0.271	0.278	0.264
		MLP	0.318	0.332	0.316	0.384	0.355
	MAE	Xgboost	0.208	0.211	0.217	0.222	0.201
		MLP	0.204	0.246	0.225	0.288	0.263
arima	RMSE	Xgboost	**0.249**	**0.245**	**0.257**	**0.255**	**0.241**
		MLP	0.283	0.283	0.332	0.315	0.286
	MAE	Xgboost	**0.193**	**0.195**	**0.208**	**0.203**	**0.183**
		MLP	0.202	0.214	0.255	0.243	0.205
persistence	RMSE		0.400	0.379	0.405	0.386	0.389
	MAE		0.279	0.268	0.296	0.276	0.270

TAB. 4 – *Erreur de test pour les vecteurs* instantané *et* arima*. Les meilleurs RMSE et MAE sont indiqués en gras pour chaque site.*

4.2 Résultats

En table 4, nous indiquons les RMSE (*Root-Mean-Square Error*) et MAE (*Mean Absolute Error*) obtenues en test pour chaque site. Le RMSE est choisi par son rapport étroit avec les métriques généralement optimisées par les algorithmers d'apprentissage. Le MAE, qui revient à l'erreur moyenne attendue d'une prédiction unique, est choisi pour son interprétabilité. À titre de comparaison, nous indiquons également le résultat obtenu en appliquant le modèle de persistence, utilisé notamment dans (Martín et al., 2010).

Les modèles appris ont tous des performances sensiblement meilleures que le modèle de persistence. L'utilisation du vecteur *arima* améliore sensiblement la qualité des modèles MLP et *Xgboost* l'utilisant, exception faite du MLP pour *Saint-André*. Le MAE atteint alors au mieux 18.3% et 20.2%, respectivement pour les modèles *Xgboost* et MLP. Cette amélioration est cependant à relativiser, car elle reste notamment limitée par rapport à l'impact d'utiliser le modèle *Xgboost* ou MLP. Par exemple, *Xgboost* utilisant le vecteur *instantané* est toujours meilleur que le MLP utilisant le vecteur *arima*.

Par la suite, nous retiendrons les modèles appris avec le vecteur *arima*. Avant d'appliquer les modèles retenus sur les données de test d'autres sites que ceux ayant servi à leur apprentissage, la table 5 indique la correlation moyenne entre sites pour les 7 variables météorologiques. Assez logiquement, les variables liées au vent telles que *WS_Mean*, *UnitX* et *UnitY* ont une corrélation inter-site faible, quasiment nulle pour les variables encodant la direction. *Text* et *Patm* ont la plus forte corrélation, de manière assez logique également car la température extérieure et la pression atmosphérique sont *a priori* assez homogènes sur un territoire de la taille de l'île de la Réunion. La variance assez élevée associée à *Patm* invite à modérer cette observation, suggérant plutôt des groupes homogènes.

Globalement, la table 5 reflète que les sites (et leurs données respectives) ne peuvent pas être agrégés *a priori*. Cette conjecture est renforcés en remarquant que k_b a une assez faible corrélation inter-site. La table 6 montre les valeurs de RMSE obtenues en appliquant les modèles spécifiques à un site donné aux données de test des différents sites. La diagonale de cette table reprend ainsi les RMSE affichés en table 4. La lecture des lignes de cette table montre les

k_b	*Patm*	*RH*	*Text*
0.44 ± 0.09	0.67 ± 0.41	0.57 ± 0.08	0.89 ± 0.02
WS_Mean	*UnitX*	*UnitY*	
0.23 ± 0.13	-0.04 ± 0.36	-0.01 ± 0.33	

TAB. 5 – *Correlation inter-site moyenne pour les variables météorologiques étudiées. L'écart-type reflète la dispersion des valeurs moyennées.*

Apprentissage \ Test		Moufia	Possession	Saint André	Saint Leu	Saint Pierre
Moufia	Xgboost	0.249	0.392	0.353	0.340	0.365
	MLP	0.283	0.297	0.291	0.317	0.288
Possession	Xgboost	0.307	0.245	0.306	0.291	0.274
	MLP	**0.272**	0.283	**0.276**	**0.303**	0.290
Saint André	Xgboost	0.318	0.439	0.257	0.302	0.398
	MLP	0.330	0.351	0.332	0.353	0.320
Saint Leu	Xgboost	0.278	0.347	0.295	0.255	0.339
	MLP	0.307	0.315	0.312	0.315	0.289
Saint Pierre	Xgboost	0.330	0.314	0.306	0.310	0.241
	MLP	0.306	0.313	0.303	0.317	0.286

TAB. 6 – *RMSE pour la prédiction inter-site utilisant le vecteur* arima*. La ligne reflète le site d'apprentissage, la colonne le site de test. Les cas où le meilleur modèle d'un site ne lui est pas spécifique sont en gras.*

performances d'un modèle donné sur tous les sites. De manière assez frappante, *XGboost* est assez médiocre sur les données des autres sites, quand les MLP sont assez robustes de ce point de vue. Ceci peut-être expliqué en partie par la normalisation des données : les ensembles mensuels-horaires permettent probablement d'encoder des propriétés climatiques générales. De manière également intéressante, la meilleure performance des MLP n'est parfois pas obtenue sur leur site d'origine (e.g. le MLP de *Sain Leu* obtient son meilleur score sur *Saint Pierre*), et le meilleur MLP pour un site donné n'est pas toujours spécifique à ce site (e.g. le MLP de *Possession* est le meilleur MLP pour 4 sites). Cette dernière observation est sans doute partiellement due à l'utilisation des données de *Possession* pour la procédure de sélection de variables et de complexité de modèle.

5 Conclusion

Nous avons répondu à la dimension prédictive du Défi EGC 2018, plus précisément la prédiction de l'indice de rayonnement solaire k_b à l'horizon d'une heure. Nous avons présenté une séquence de pré-traitements, une sélection de variables basée sur le modèle ARIMA, et des résultats expérimentaux obtenus grâce aux modèles *Xgboost* et MLP. *Xgboost* obtient alors les meilleures performances. Notre procédure de sélection de variables apporte un gain de

performance, qui reste cependant marginal au fait de substituer *Xgboost* au MLP, notamment. Une rapide étude des corrélations inter-site montre que les données des 5 sites ne peuvent pas être simplement agrégées en un seul ensemble d'apprentissage. Après avoir entrainé nos modèles sur leur site respectif, nous avons observé leur performance sur les données de test d'autres sites. La normalisation mensuelle-horaire semble alors donner un avantage au MLP. Une modélisation des dépendances entre sites plus poussée est une extension possible à nos observations préliminaires.

L'approche adoptée dans cet article fait peu de cas de la nature séquentielle des données traitées. Prendre en compte cette nature peut *a priori* mener à des améliorations qualitatives, par exemple en adaptant une architecture de réseaux de neurones convolutionnelle (Krizhevsky et al., 2012) ou récurrente (Williams et Zipser, 1989) à notre problème de prédiction. Ces modèles ont déjà été utilisés dans des problèmes séquentiels apparentés, tels que l'analyse de sentiment dans un texte (Severyn et Moschitti, 2015) ou la traduction automatique (Liu et al., 2014).

Références

Anderson, B. (2012). d3-tsline : a time-series line graph visualization using D3. `https://github.com/boorad/d3-tsline`.

Anderson, O. D. (1976). *Time series analysis and forecasting : the Box-Jenkins approach*. Butterworths.

Barbounis, T., J. Theocharis, M. Alexiadis, et P. Dokopoulos (2006). Long-term wind speed and power forecasting using local recurrent neural network models. *IEEE Transactions on Energy Conversion 21*(1), 273–284.

Box, G. E., G. M. Jenkins, G. C. Reinsel, et G. M. Ljung (2015). *Time series analysis : forecasting and control*. Wiley & Sons.

Bruneau, P. et L. Boudet (2012). Bayesian variable selection in neural networks for short-term meteorological prediction. In *Neural Information Processing*, pp. 289–296.

Bruneau, P., L. Boudet, et C. Damon (2012). Neural architectures for global solar irradiation and air temperature prediction. In *ICANN*, pp. 548–556.

Bruneau, P. et P. Pinheiro (2017). Sunrise and Sunset Times for La Réunion island. `https://github.com/pbruneau/Reunion-Sun-Rise-Set/`.

Chen, T., M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, et Z. Zhang (2015). Mxnet : A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv 1512.01274*.

Dominici, F., A. McDermott, S. Zeger, et J. Samet (2002). On the use of generalized additive models in time-series studies of air pollution and health. *American Journal of Epidemiology 156*, 193–203.

Friedman, J. (2001). Greedy function approximation : a gradient boosting machine. *Annals of Statistics 29*(5), 1189–1232.

Grolemund, G. et H. Wickham (2011). Dates and times made easy with lubridate. *Journal of Statistical Software 40*(3), 1–25.

Hastie, T. et R. Tibshirani (1990). *Generalized additive models*. Wiley & Sons.

Hyndman, R. et Y. Khandakar (2007). Automatic time series for forecasting : the forecast package for R. Technical Report 6/07, Monash University.

Krizhevsky, A., I. Sutskever, et G. Hinton (2012). Imagenet classification with deep convolutional neural networks. In *ACM SIGKDD*, pp. 1097–1105.

Kylling, A., T. Persen, B. Mayer, et T. Svenoe (2000). Determination of an effective spectral surface albedo from ground-based global and direct UV irradiance measurements. *Atmospheres 105*(D4), 4949–4959.

Liu, S., N. Yang, M. Li, et M. Zhou (2014). A recursive recurrent neural network for statistical machine translation. In *ACL*, pp. 1491–1500.

Lynch, P. (2008). The origins of computer weather prediction and climate modeling. *Journal of Computational Physics 227*(7), 3431–3444.

Martín, L., L. Zarzalejo, J. Polo, A. Navarro, R. Marchante, et M. Cony (2010). Prediction of global solar irradiance based on time series analysis : Application to solar thermal power plants energy production planning. *Solar Energy 84*(10), 1772–1781.

Reno, M., C. Hansen, et J. Stein (2012). Global horizontal irradiance clear sky models : Implementation and analysis. SANDIA Report SAND2012-2389.

Ridgeway, G. (2007). Generalized Boosted Models : A guide to the gbm package.

Severyn, A. et A. Moschitti (2015). Twitter sentiment analysis with deep convolutional neural networks. In *ACM SIGIR*, pp. 959–962.

Tapakis, R., S. Michaelides, et A. Charalambides (2016). Computations of diffuse fraction of global irradiance : Part 1–analytical modelling. *Solar Energy 139*(7), 711–722.

Williams, R. et D. Zipser (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Computation 1*(2), 270–280.

Zeileis, A. et G. Grothendieck (2005). zoo : S3 infrastructure for regular and irregular time series. *CoRR math/0505527*.

Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *ICML*, pp. 116.

Summary

This paper describes a flexible approach to short term prediction of meteorological variables. In particular, we focus on the prediction of the solar irradiance one hour ahead, a task that has high practical value when optimizing solar energy resources. As *Défi EGC 2018* provides us with time series data for 5 geographical sites from *La Réunion* island, we test the value of using recently observed data as input for prediction models, as well as the performance of models across sites. After describing our data cleaning and normalization process, we combine a variable selection step based on *AutoRegressive Integrated Moving Average* (ARIMA) models, to using general purpose regression techniques such as neural networks and regression trees.