La notion d'année typique en météorologie

Marie Boutigny

AGROCAMPUS OUEST

☐ **CFR Angers**

☑ **CFR Rennes**

Année universitaire : 2016 - 2017 Spécialité : Data Science Spécialisation (et option éventuelle) : ……	**Mémoire de fin d'études** ☑ d'Ingénieur de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage ☐ de Master de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage ☐ d'un autre établissement (étudiant arrivé en M2)

La notion d'année typique en météorologie

Par : Marie BOUTIGNY

Soutenu à Rennes ***le 05/09/2017***

Devant le jury composé de :

Président :

Maître de stage : Valérie Monbet

Enseignant référent : François Husson

Autres membres du jury (Nom, Qualité)

Fiche de confidentialité et de diffusion du mémoire

Confidentialité

☐ Non ☐ Oui si oui : ☐ 1 an ☐ 5 ans ☐ 10 ans

Pendant toute la durée de confidentialité, aucune diffusion du mémoire n'est possible [1].

Date et signature du maître de stage [2] :

A la fin de la période de confidentialité, sa diffusion est soumise aux règles ci-dessous (droits d'auteur et autorisation de diffusion par l'enseignant à renseigner).

Droits d'auteur

L'auteur[3] **Nom Prénom -Boutigny Marie**
autorise la diffusion de son travail (immédiatement ou à la fin de la période de confidentialité)

☑ Oui ☐ Non

Si oui, il autorise

☑ la diffusion papier du mémoire uniquement(4)

☑ la diffusion papier du mémoire et la diffusion électronique du résumé

☑ la diffusion papier et électronique du mémoire (joindre dans ce cas la fiche de conformité du mémoire numérique et le contrat de diffusion)

(Facultatif) ☑ accepte de placer son mémoire sous licence Creative commons CC-By-Nc-Nd (voir Guide du mémoire Chap 1.4 page 6)

Date et signature de l'auteur :

Autorisation de diffusion par le responsable de spécialisation ou son représentant

L'enseignant juge le mémoire de qualité suffisante pour être diffusé (immédiatement ou à la fin de la période de confidentialité)

☐ Oui ☐ Non

Si non, seul le titre du mémoire apparaîtra dans les bases de données.
Si oui, il autorise

☐ la diffusion papier du mémoire uniquement(4)

☐ la diffusion papier du mémoire et la diffusion électronique du résumé

☐ la diffusion papier et électronique du mémoire

Date et signature de l'enseignant :

(1) L'administration, les enseignants et les différents services de documentation d'AGROCAMPUS OUEST s'engagent à respecter cette confidentialité.
(2) Signature et cachet de l'organisme
(3).Auteur = étudiant qui réalise son mémoire de fin d'études
(4) La référence bibliographique (= Nom de l'auteur, titre du mémoire, année de soutenance, diplôme, spécialité et spécialisation/Option)) sera signalée dans les bases de données documentaires sans le résumé

RAPPORT DE STAGE

Marie Boutigny

5 septembre 2017

La notion d'Année Typique en Météorologie

Je tiens à remercier Valérie Monbet, Anne Cuzol et Pierre Aillot pour leur accueil, leur aide et leur pédagogie.

Table des matières

1 Introduction : Sujet et contexte **4**

2 Méthodologies de construction et de sélection d'années typiques **5**
2.1 Etat de l'art : les méthodes "classiques" 5
2.1.1 Comparaison de moyennes mensuelles 5
2.1.2 Méthode Sandia 6
2.1.3 Conclusion sur les méthodes classiques 6
2.2 Construction des séries courtes 6
2.3 Sélection de la série la plus typique : quel sens donner au mot "typique" ? . 7

3 Application et validation des méthodologies **11**
3.1 Projet DESIRES : dimensionnement d'une usine de désalinisation 11
3.1.1 Présentation du projet et des données 11
3.1.2 Etude de la variabilité de la réponse 12
3.1.3 Résultats 15
3.1.3.1 Méthodes classiques : comparaison de moyennes mensuelles et méthode Sandia 15
3.1.3.2 Années réelles et sélection par les distances proposées . . . 15
3.2 Projet MEDISA : méthodologie de dimensionnement de systèmes d'assainissement 17
3.2.1 Présentation du projet et des données 17
3.2.2 Etude de la variabilité de la réponse 19
3.2.3 Résultats 19
3.2.3.1 Extraction et génération des candidats 19
3.2.3.2 Sélection de la série la plus typique 20

4 Conclusion : Discussion et perspectives **23**

1 Introduction : Sujet et contexte

Les données météorologiques sont souvent des chroniques très longues : des séries de 20-40 ans sont facilement disponibles avec un pas de temps descendant souvent jusqu'à l'horaire. Certains modèles qui utilisent ce genre de données en entrée peuvent être assez lourds en termes de temps de calcul et il n'est pas toujours possible de les faire tourner sur des séries aussi longues (exemple : réseaux de neurones [30]). La problématique des années dites "typiques" ou "de référence" vient de ce besoin de réduire la longueur des chroniques en essayant d'obtenir la série la plus typique à partir de la série longue.

A partir de ce terme d'année typique, on peut tirer deux questionnements. Tout d'abord le terme "année" : comment extrait-on les observations qui vont constituer l'année typique ? Une année typique n'est pas nécessairement une année réelle du jeu de données. On peut envisager par exemple de prendre des séquences de longueur fixe dans la série originelle, ou encore de prendre des observations issues d'une simulation. La question est donc autrement dit de savoir comment construire les candidats à être cette année typique. Naturellement la deuxième question porte sur le mot "typique" et le sens qu'on veut lui donner. Dans le cas par exemple d'une loi de Gauss pour laquelle on dispose de 10 observations i.i.d., comment extrait-on 1 (ou $k < 10$) valeurs pour constituer un sous-échantillon typique ? On peut penser à prendre les quantiles, ou à vouloir se rapprocher de l'espérance de la loi par exemple. On semble en tout cas vouloir reproduire la distribution des valeurs de la variable. Comment peut-on résumer cette distribution ? Plusieurs outils existent : moyenne, écart-type, fonction de répartition cumulée, etc. Il faut noter que dans le cas de séries temporelles on ne connaît pas la distribution à long terme, seulement la loi empirique issue des observations dont on dispose. La question est là d'avoir un outil permettant de choisir un candidat parmi les autres, qu'on qualifiera de "typique".

La méthodologie utilisée pour construire et sélectionner une année typique peut être validée de plusieurs façons. S'intéresse-t-on seulement à l'entrée de l'application visée, c'est-à-dire à la série météorologique, ou veut-on prendre en compte la sortie du modèle ? Dans le deuxième cas, qui est celui qui sera considéré dans ce rapport, se pose aussi la question de la façon de prendre en compte cette sortie dans la validation.

Dans le contexte de ce stage deux applications seront abordées. Le projet européen DESIRES, porté par ERANET/MED, vise à créer un outil de dimensionnement d'usine de désalinisation en mer Méditerranée. L'usine fonctionnera avec deux sources d'énergie : l'éolien et le solaire. L'algorithme de dimensionnement ne peut pas prendre en entrée la série entière de données (voir ci-après). La deuxième application est le projet MEDISA (Méthodologie de Dimensionnement des Systèmes d'Assainissement), porté par Eau du Ponant, et a pour but de dimensionner les systèmes de gestion des eaux de pluie pour limiter les déversements en milieu naturel. La législation donne plusieurs critères, à calculer sur cinq ans.

Deux jeux de données seront donc utilisés. Le premier, celui du projet DESIRES, consiste en des séries de 43 ans (1958-2001) mesurées avec un pas de temps de 6 heures, au point de latitude $35,25°N$ et de longitude $23,25°W$ (proche de la Crète). Ces données sont issues de la base de données de réanalyse ERA 40 - daily d'ECMWF [27], de laquelle on a extrait les variables suivantes : R la radiation solaire en Wh/m^2, t2m la température à deux mètres en K (Kelvin) et WI l'intensité du vent en m/s. Le second jeu de données, celui du projet MEDISA, consiste en une série de 15 ans (1997-2011) de précipitations relevées à Guipavas (Brest) par Météo France. Ces données sont horaires et exprimées en

mm/h. Un extrait des deux jeux de données est présenté en table 1.

time	R	t2m	WI
$1958-01-01\ 00:00$	0	288.04	8.63
$1958-01-01\ 06:00$	0.0365	287.23	4.00
...	...	...	...

time	Rain
$1997-01-01\ 00:00$	0
$1997-01-01\ 01:00$	0
...	...

TABLE 1 – Jeux de données bruts

Dans ce rapport, seront d'abord présentées plusieurs méthodologies de construction d'années typiques (section 2). En section 3 on détaillera les résultats obtenus pour les deux applications présentées précédemment dans le but de valider les méthodes proposées en section 2. La section 4 donnera les perspectives et la conclusion.

2 Méthodologies de construction et de sélection d'années typiques

2.1 Etat de l'art : les méthodes "classiques"

On trouve dans la littérature de nombreux domaines où la problématique des années typiques est abordée. De nombreuses méthodes ont été développées, on commencera par présenter les deux plus répandues, qui sont considérées comme classiques. Dans ces méthodes, on travaille mois par mois afin de s'affranchir de la saisonnalité. Une année typique est alors en fait un ensemble de mois typiques.

2.1.1 Comparaison de moyennes mensuelles

L'idée de cette méthode, présentée par exemple dans [5] et [16], est de comparer les moyennes mensuelles pour la variable d'intérêt. Pour le mois de janvier de chaque année par exemple, on calcule sa moyenne mensuelle et on la compare à la moyenne de tous les mois de janvier. En notant X la variable d'intérêt, cela donne :

$$\min_y \{(X_{ym} - X_m)^2\},$$

où X_{ym} est la moyenne du mois m de l'année y et X_m la moyenne du mois m, calculée sur toutes les années.

Cette méthode peut être étendue à des cas multivariés en minimisant une somme pondérée des écarts au long terme de chaque variable X_p :

$$\min_y \{\sum_p w_p (X_{p,ym} - X_{p,m})^2\},$$

où $X_{p,ym}$ est la moyenne pour la variable X_p du mois m de l'année y et $X_{p,m}$ la moyenne pour la variable X_p du mois m, calculée sur toutes les années. w_p est le poids accordé à la variable X_p, avec $\sum_p w_p = 1$. Dans le cas multivarié, un point clé de la méthode est le choix de ces poids, qui dépend de l'application visée.

2.1.2 Méthode Sandia

La méthode Sandia, décrite dans [14] et utilisée par exemple dans [16] et [7], est basée sur la comparaison des fonctions de distribution cumulées (CDF). Par mois, on calcule la statistique de Finkelstein-Schafer [12] de la façon suivante :

$$FS_{ym} = \sum_j (CDF_{ym}[j] - CDF_m[j])^2,$$

où CDF_{ym} est la CDF du mois m de l'année y, CDF_m la CDF du mois m sur toutes les années, et où j parcourt le domaine des CDF.

Le mois sélectionné est alors le mois avec le FS le plus bas.

De la même façon que pour la méthode par comparaison des moyennes mensuelles, la méthode Sandia a été développée pour des cas multivariés. On choisit le mois minimisant la somme pondérée des FS (qui sont donc calculés par variable).

$$\min_y \{\sum_p w_p FS_{p,ym}\} = \min_y \{\sum_p w_p \sum_j (CDF_{p,ym}[j] - CDF_{p,m}[j])^2\},$$

où $CDF_{p,ym}$ est la CDF de la variable X_p du mois m de l'année y, $CDF_{p,m}$ la CDF de la variable X_p du mois m sur toutes les années, et où j parcourt le domaine des CDF de chaque variable. w_p est le poids accordé à la variable X_p, avec $\sum_p w_p = 1$.

2.1.3 Conclusion sur les méthodes classiques

Si les deux méthodes présentées précédemment sont les plus utilisées, de nombreuses autres existent, on peut notamment citer la méthode danoise [3] qui propose une analyse des résidus via leurs moyenne et écart-type, ou la méthode Festa-Ratto [11] qui construit une distance qui est une combinaison linéaire de trois distances basées sur la moyenne, l'écart-type et la distribution.

On distingue dans ces méthodes deux grandes étapes dans la construction d'une année typique.

Tout d'abord se pose la question de l'extraction de la série courte. Dans les méthodes évoquées précédemment, la série considérée n'est pas l'année entière mais chaque mois un par un. On a donc comme série longue l'ensemble des mois de janvier par exemple, et les séries courtes sont alors les mois de janvier de chaque année. On a autant de séries courtes qu'on a d'années. L'extraction des séries courtes consiste dans ce cas à prendre simplement les années réelles.

Une fois qu'on a extrait des séries courtes à partir de la série longue, reste la question de chercher la plus typique d'entre elles. On cherche alors à mesurer une dissimilarité entre les deux séries temporelles. Dans le cas de la comparaison de moyennes mensuelles cette mesure est la distance euclidienne entre les moyennes des séries, et dans le cas de la méthode Sandia, c'est la statistique de Finkelstein-Schafer qui sert de distance.

Ainsi une année typique sera en fait un assemblage de mois réels, typiques au sens d'une certaine distance au long terme. Les deux étapes identifiées seront abordées dans les deux parties suivantes afin de proposer d'autres méthodologies.

2.2 Construction des séries courtes

Dans les méthodes les plus utilisées, les séries courtes correspondent à une réalité physique, l'année, et sont simplement les séries réelles. Ces séries sont donc limitées par

le nombre d'années disponibles. Dans le cas de variables comme le vent par exemple, on a une grande variabilité inter annuelle, qu'on ne retrouvera pas en sélectionnant une année. De plus dans certains cas, on peut se retrouver à travailler avec des bases de données relativement courtes. Une solution qui a été envisagée est de générer des séries artificiellement.

L'idée est ici de considérer une série météorologique comme une succession d'événements, et de classer ces évènements en types de temps. On voit donc la série comme une émission de valeurs dont la loi est conditionnelle à un état sous-jacent, ce qui correspond à une chaîne de Markov cachée (HMM) [33] [34].

Le principe est le suivant :

La variable observée est une variable de classe $X \sim \mathcal{M}(\pi_1^c, ..., \pi_L^c)$ où L est le nombre de classes émises, et π_l^c la probabilité de X d'appartenir à la classe l. On suppose que cette variable d'émission X suit une chaîne de Markov d'ordre 1 conditionnellement à un état, c. Cet état sous-jacent est une réalisation d'une variable cachée de K classes : $S \in \{1, ..., K\}$. On suppose que S suit une chaîne de Markov d'ordre 1. On fait donc les hypothèses suivantes :

$$P[X_t|X_1, ..., X_{t-1}, S_1, ..., S_{t-1}] = P[X_t|X_{t-1}, S_{t-1}]$$

$$P[S_t|S_1, ..., S_{t-1}] = P[S_t|S_{t-1}]$$

On a donc les dépendances suivantes :

$$\begin{array}{ccccccccc} S_1 & \rightarrow & S_2 & \rightarrow & ... & \rightarrow & S_t & & \textit{(états)} \\ \downarrow & & \downarrow & & ... & & \downarrow & & \\ X_1 & \rightarrow & X_2 & \rightarrow & ... & \rightarrow & X_t & & \textit{(émissions)} \end{array}$$

La chaîne cachée d'états S du HMM permet de voir la série longue comme une succession de blocs d'états qui représentent les évènements, avec pour chaque bloc une émission de valeurs conditionnée par l'état. L'idée est alors d'utiliser un générateur par blocs de classes pour créer des séries artificielles, autrement dit de mélanger les évènements en respectant les probabilités de transition d'un type d'évènement à l'autre.

Pour ce faire on procède de la façon suivante : à partir de la série longue d'états on calcule les probabilités de transition d'un bloc d'état à l'autre. On commence par tirer un bloc au hasard, puis on répète les étapes suivantes :

- Avec les probabilités de transitions, tirer l'état du bloc suivant, c
- Tirer un bloc parmi ceux dont l'état est c

Contrairement à une chaîne de Markov sur les états, faire une chaîne sur les blocs d'états permet de s'assurer que la longueur des blocs sera bien respectée. Autrement dit, on relaxe l'hypothèse Markovienne et on peut donc simuler des blocs plus longs (ou plus courts) que ceux qui seraient obtenus en simulant directement selon le modèle MSAR [2].

2.3 Sélection de la série la plus typique : quel sens donner au mot "typique" ?

Une fois les séries courtes extraites, la question de chercher la série la plus typique revient à mesurer une dissimilarité entre deux séries temporelles de tailles différentes. De nombreuses mesures de distances ont été listées, principalement issues du package `TSdist` [24], elles sont listées en table 2.

Les distances CCor (cross correlation), ACF (autocorrelation function) et PACF (partial autocorrelation function) sont basées sur la corrélation. La distance SAX [22] consiste à transformer la série en une variable de classe avec des seuils et à utiliser une distance

ccor	mindist.sax	edr	ar.lpc.ceps	wav	cvm.ex	sandia
pacf	ncd	erp	ar.mah.statistic	spec.glk	cvm	mean.manhattan
acf	cdm	lcss	ar.pic	energy	ks	mean.euclidean
	pdc	dtw	pred		density.kl	
		frechet				

TABLE 2 – Listes des distances étudiées

entre chaînes de caractères. Les distances NCD, CDM et PDC sont basées sur la compression des données [17]. DTW (dynamic time warping [6]), LCSS (plus longue séquence commune [28]), EDR (edit distance [29]), ERP (edit real penality [8]) et Fréchet [13] visent à aligner au mieux les séries avant de mesurer une distance en pénalisant les zones non alignées. Les trois distances "ar..." ajustent un modèle auto-régressif sur les données avant de calculer une distance sur les coefficients ou autre. Pred est une distance basée sur la prédiction (modèle autorégressif). SpecGLK (Spectral global generalized likelihood [10]) et Wav (décomposition en ondelettes [32]) appliquent une transformation avant de calculer respectivement une généralisation de la vraisemblance et une distance sur les coefficients. La distance Energy [26] ressemble au théorème de Huygens : la distance mesurée entre les deux séries (inter) est la distance entre tous les couples de points des deux séries (global) moins la distance entre les couples de points de chaque série prise à part (intra), la distance entre deux points étant la norme euclidienne. Les distances CVM (Cramer Von Mises [9]), CVM.ex (Anderson Darling [4]), KS (Kolmogorov Smirnov [9]) et density.KL (Kulbach Leibler [21]) sont basées sur des mesures de distances entre des fonctions de distribution (cumulées) multivariées, dans lesquelles on peut faire intervenir plusieurs variables différentes et/ou une variable et son lag (X_t et $X_{t-1},\ t \in 2, ..., n$). Enfin les distances mean.euclidean, mean.manhattan et Sandia correspondent aux méthodes présentées en sections 2.1.1 et 2.1.2.

Dans un cas univarié, la série courte typique est celle minimisant une des distances citées ci-dessus. Dans un cas multivarié, la distance choisie est calculée pour chaque variable, et on minimise la somme des distances pondérées par les poids accordés aux variables.

Afin de comparer les 24 dissimilarités à notre disposition, on se propose de travailler sur une série artificielle afin d'en extraire des séries courtes de façon faussée, c'est-à-dire de façon à sélectionner à chaque fois une caractéristique bien particulière de la série.

a) Simulation des séries longues et extraction des séries courtes

On choisit de travailler sur des séries d'intensité de vent à 10 mètres de ERA 40 daily [27], qui vont de 1958 à 2001 (données du projet DESIRES).

On modélise cette série de la façon suivante :

$$X(t) = a.\cos\frac{2\pi t}{n} + b.\sin\frac{2\pi t}{n} + \omega(t)$$
$$\omega(t) = a_0 + a_1 X_{t-1} + ... + a_p X_{t-p} + \epsilon(t)$$
$$\epsilon \sim \mathcal{N}(\mu, \sigma)$$

Où $n = 1460$ est le nombre d'observations par an et $X(t)$ est la série d'intensité de vent observée.

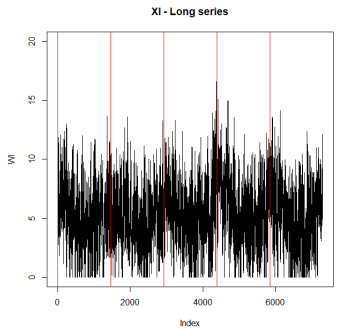

FIGURE 1 – Intensité du vent générée sur 5 ans

La saisonnalité est modélisée par les premières composantes de la décomposition en série de Fourier. Un modèle auto-régressif sur les résidus permet de modéliser la dépendance temporelle. Les résidus de ce modèle sont supposés suivre une loi normale centrée. On simule avec ces modèles 5 ans de données fictives, qui sont représentées en figure 1.

On s'attaque maintenant à l'extraction des séries courtes. Six façons de sélectionner la série courte sont proposées et représentées en figure 2.

- `random` : On choisit n observations au hasard dans la série longue. On obtient ainsi une série qui ne conserva pas la structure temporelle mais qui garde bien la distribution des valeurs de la série.
- `frag1n` : On choisit au hasard un fragment de taille n. La structure temporelle est conservée mais on ne représente pas la variabilité interannuelle, on peut donc avoir un biais dans la distribution des valeurs.
- `regstep` : On prend des observations de la série longue avec un pas de temps régulier de façon à avoir en sortie n observations. C'est comme si le temps était condensé, la structure temporelle est cassée mais la variabilité inter-annuelle est représentée.
- `<10` : On prend toutes les valeurs inférieures à 10. On garde plutôt bien la structure temporelle malgré quelques coupures, mais les valeurs fortes ne sont pas bien représentées.
- `<5` : On prend toutes les valeurs inférieures à 5. On a les mêmes caractéristiques que la série précédente, mais la structure temporelle est beaucoup moins bien respectée.
- `/10` : On prend la série entière mais on divise les valeurs par 10. La structure temporelle est parfaitement respectée, mais les valeurs sont complètement faussées.

Les trois dernières méthodes de sélection ne donnent pas n observations. Or si on veut réduire ces séries à n observations, il va falloir encore utiliser des méthodes de sélection comme `random`, `regstep` ou `frag1n`. On se propose donc de mixer les trois premiers et les trois derniers points. Ainsi on pourra vérifier l'effet de la taille. On rajoute donc les neufs séries suivantes :

- `<10 + random` : On ne prend que les valeurs inférieures à 10, puis on choisit n observations au hasard.
- `<10 + frag1n` : On ne prend que les valeurs inférieures à 10, puis on choisit un fragment.
- ...
- `/10 + regstep`

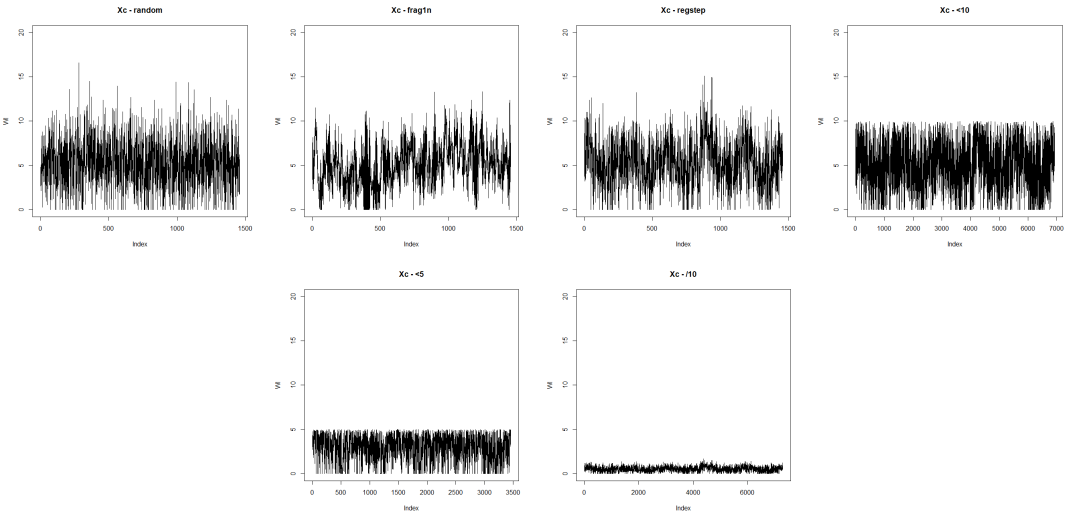

FIGURE 2 – Séries courtes extraites de la figure 1

b) Comparaison des différentes mesures de distance

Les 24 dissimilarités sont calculées sur nos séries simulées. On obtient un tableau de 15 lignes par 24 colonnes, cf. table 3, avec en lignes les séries courtes et en colonnes les distances. En entrée on a la distance entre la série courte et la série longue.

Afin de résumer l'information contenue dans ce tableau on réalise une ACP, dont les sorties sont présentées en figure 3, qui amène à s'intéresser aux trois premiers axes (>90% de variance expliquée, valeurs propres >1).

	ccor	acf	...	pred
random	1,351	3,056	...	0,116
...	...	...	...	...
/10-regstep	1,345	2,92	...	1,87

TABLE 3 – Distances entre la série simulée entière et les différentes séries courtes

Le premier plan, qui explique plus de 80% de variance, permet de distinguer deux groupes de variables. Du côté des individus on observe que les séries s'organisent selon deux diagonales. Sur la diagonale principale on a, en allant d'en haut à droite vers en bas à gauche, les séries dont les valeurs on été divisées par 10, les inférieur à 5 et enfin les inférieur à 10. Sur la seconde diagonale on a en haut à gauche les séries sélectionnées aléatoirement (`...-random`) ou avec un pas de temps régulier (`...-regstep`) et en bas à droite les fragments (`...-frag1n`). On a deux groupes de variables sur le premier plan. Le premier donne de fortes valeurs aux séries dont les valeurs sont fortement biaisées (`/10-...`) et de faibles valeurs à celles moins biaisées (`<10-...`, ou `random` et `regstep` seuls). Ce groupe ne semble pas particulièrement avoir de préférence pour le mode de sélection qui détermine si la structure temporelle est conservée ou non. Dans ce groupe on retrouve les distances basées sur les distributions (KL, CVM, Sandia, etc.) et la méthode de comparaison de moyenne. Pour le deuxième groupe on retrouve des distances basées sur des modèles auto-régressifs, l'auto-corrélation et la compression. Ces distances pénalisent les séries ne conservant pas la structure temporelle (`...-random`, `...-regstep`) et favorisent au

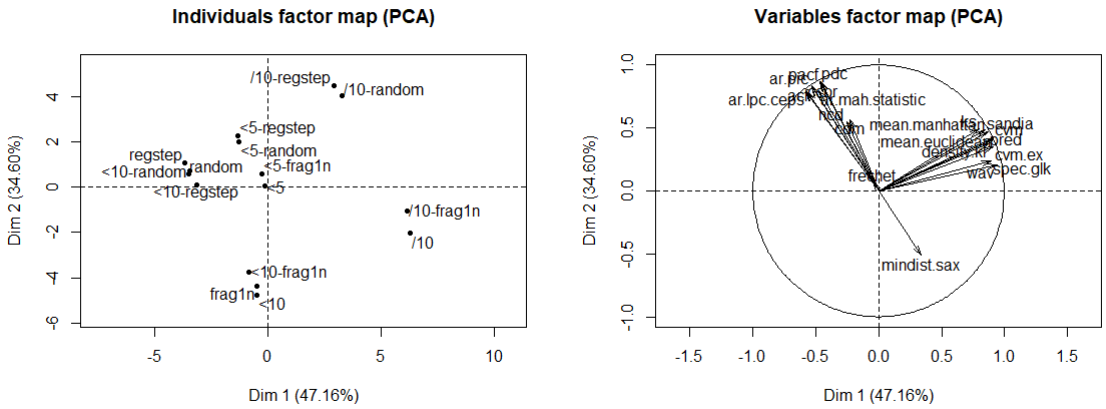

FIGURE 3 – Sorties d'ACP sur le tableau 3

contraire les fragments et les séries entières malgré le biais (`frag1n`, `...-frag1n`, `/10-...` et `<10-...`). Le troisième axe permet de repérer certaines distances (Fréchet, CDM, NCD) pour lesquelles la taille de la série courte a joué : les séries `<10`, `/10` et `<5` contribuent fortement à cet axe.

Avec ces simulations on a pu constater quelles distances prêtaient plus attention à la structure temporelle et quelles distances se concentraient sur la répartition des valeurs. Il est difficile de caractériser les distances autrement que par ces deux points avec l'analyse réalisée, ce qui est probablement dû aux séries courtes proposées.

3 Application et validation des méthodologies

3.1 Projet DESIRES : dimensionnement d'une usine de désalinisation

3.1.1 Présentation du projet et des données

Dans la suite des travaux [18] et [19], ce projet a pour but de dimensionner une usine de désalinisation d'eau de mer. La station doit remplir un réservoir d'eau pour un ensemble d'habitations, le volume du réservoir est fixé a priori et le réservoir ne doit jamais être vide. L'énergie nécessaire est produite par une combinaison de panneaux solaires et d'éoliennes. Des batteries permettent de stocker l'excédent de production d'énergie. Les conditions initiales considèrent que le réservoir d'eau est vide et que les batteries sont partiellement chargées. Le problème de dimensionnement consiste à déterminer quel est le nombre de panneaux, éoliennes et batteries nécessaires. En pratique, on choisira la combinaison la moins coûteuse parmi celles qui permettent de garantir une production d'eau douce suffisante. Le coût de l'installation comprend l'investissement initial et l'entretien sur 20 ans. L'algorithme d'optimisation et la simulation sont décrits en annexe 1, la simulation étant tirée de [18].

Les données d'intensité de vent, de radiation solaire et de température utilisées sont celles de la base de données ERA 40 daily [27], elles couvrent la période 1958-2001. Les trois premiers mois de 2000 sont représentés en figure 4, à gauche. On dispose de mesures toutes les six heures (minuit, 6h, 12h et 18h), pour la radiation toutes les mesures à minuit

valent donc zéro. Afin d'avoir une idée de ce qui se passe globalement on peut regarder les tendances : on fait la moyenne sur toutes les années qu'on lisse sur deux jours, et on obtient les graphiques présentés en figure 4, à droite. On constate que la radiation augmente nettement et de façon assez lisse, que la température est à son minimum en février (effet quadratique), et que le vent, même s'il est plus variable, a tendance à baisser au cours des trois mois. Tous les effets mentionnés ont été testés et se sont avérés significatifs.

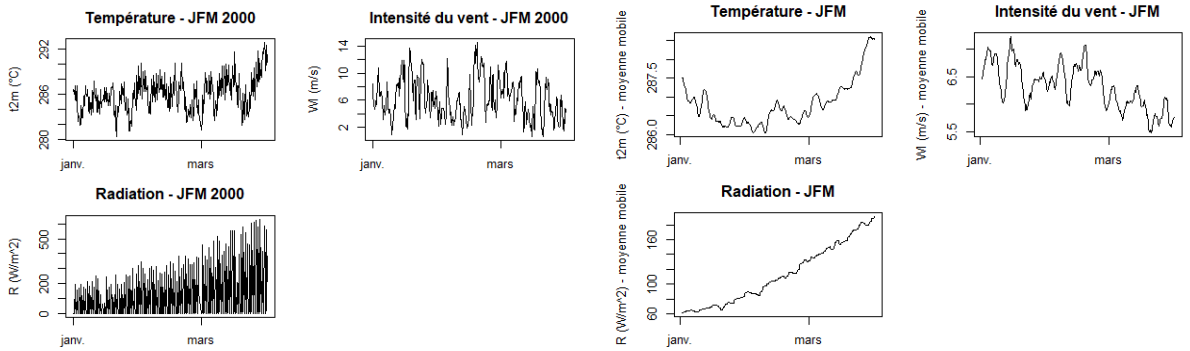

FIGURE 4 – Séries brutes (gauche) et tendances (droite) - 2000 (Janvier Février Mars)

Le but est d'extraire une année typique pour ces trois variables, l'algorithme d'optimisation utilisé dans [18] ne pouvant pas prendre 43 ans en entrée. La question du poids des variables ne sera pas abordée, on leur donnera à toutes le même poids. D'autre part on se concentrera sur les trois premiers mois de l'année, le coût numérique du calcul des distances étant important.

Les méthodes testées pour sélectionner les trois mois typiques seront pour ce qui est de construire les séries candidates une simple extraction des mois*année (on a donc 43 candidats pour chaque mois), et pour ce qui est de la sélection toutes les distances présentées en section 2.3 seront testées.

On souhaite valider les différentes méthodes de construction d'années typiques avec la sortie du système, pour pouvoir apporter une réponse adaptée au projet. On commencera donc par étudier la variabilité de la réponse du modèle, qui est le coût optimal permettant de satisfaire la demande, avant de s'attaquer aux résultats des différentes méthodologies de construction d'années typiques.

3.1.2 Etude de la variabilité de la réponse

Dans cette partie on cherche à étudier la sensibilité de la réponse. En particulier, on se demande si les séries dont on dispose sont assez longues pour que l'estimation du coût soit stable, et on cherche à comprendre ce qui dimensionne le coût du système.

On commence par créer artificiellement de la variabilité en enlevant les années une par une. On peut ainsi étudier la stabilité du coût optimal. On obtient la figure 5 (premier histogramme "1 year out").

La plupart des coûts optimaux se situent autour de 50000 euros, ce qui correspond à une configuration avec 11-12 PV, 3-4 W/G et 4-5 batteries. Mais on a quelques cas où il prend une valeur bien plus élevée : 90000 euros. On parlera alors d'une série "difficile", au sens où il faut un système avec beaucoup de panneaux, éoliennes et/ou batteries pour

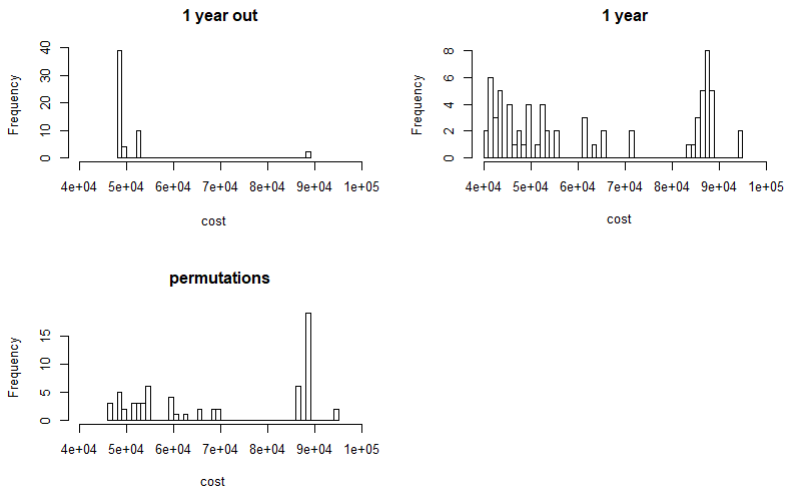

FIGURE 5 – Distribution du coût optimal (1 year out, 1 year, permutations)

produire suffisamment d'énergie. Ces coûts correspondent au cas où on enlève l'année 1958, et il faut alors 28 batteries pour satisfaire la demande. On aurait donc tendance à considérer notre série de données comme stable, en-dehors de ce cas particulier.

Afin de mieux comprendre ce qui se passe, on réalise le même histogramme mais en prenant les années une par une. On n'a alors qu'une seule année dans la simulation. On se permet de comparer les sorties aux précédentes car on constate que les sorties ne changent pas en répétant l'année 43 fois. L'histogramme des coûts optimaux obtenus avec les années une par une est présenté en figure 5 (histogramme "1 year"). On s'attendrait à avoir plus de variabilité que dans le "1 year out", mais qu'elle soit uniquement due à des années avec des coûts plus faibles. En effet le système devant satisfaire la demande, on pourrait penser qu'il est dimensionné par la pire période de la série, et qu'ainsi une année seule peut soit contenir cet événement soit mener à un coût plus faible. On a en effet beaucoup plus de variabilité, et on constate deux choses : premièrement on a beaucoup plus de cas où on doit mettre plus de 20 batteries pour satisfaire la demande (autour de 90000 euros). Ensuite, autour du cas "classique", on a beaucoup plus de variabilité, mais avec aussi bien des années plus difficiles (coût plus élevé) que des années plus faciles (coût plus faible). Le coût ne semble pas dimensionné par le pire événement de la série.

Les cas avec un nombre d'années entre 1 et 42 ont aussi été étudiés et on a pu constater qu'on allait vers quelque chose qui ressemblait à l'histogramme "1 year" au fur et à mesure qu'on enlevait des années, c'est-à-dire plus de variabilité autour de 50000 euros et de plus en plus de cas autour de 90000 euros.

Les cas avec des coûts plus élevés étant plus fréquents lorsqu'on enlève des années, on se demande si ce n'est pas dû aux coupures que l'on fait dans la série. On fait l'histogramme du coût optimal trouvé sur des séries de 43 ans mais dont les années ont été permutées (histogramme "permutation" de la figure 5). On a alors une distribution avec beaucoup de cas à plus de 20 batteries, et avec plus de variabilité que l'histogramme "1 year out" autour de 50000 euros, cette variabilité étant principalement due à des cas avec un coût plus élevé.

Une hypothèse qui expliquerait ces résultats serait que le système soit principalement

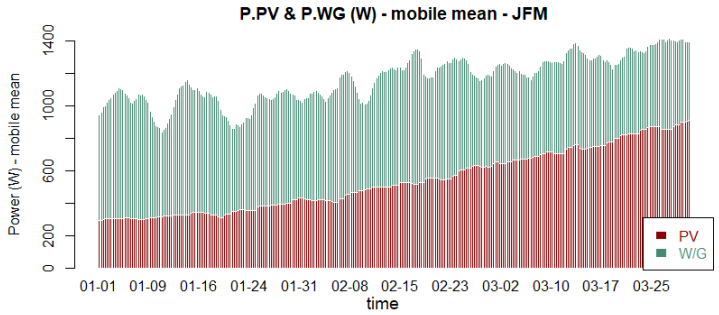

FIGURE 6 – Puissance fournie par les ressources de janvier à mars (moyenne mobile) et sa répartition entre les différentes sources (panneaux solaires PV et éoliennes W/G)

dimensionné par le début de la série. En effet, quand on enlève une seule année, le seul cas qui donne un coût anormalement élevé est celui où on change la première année (on enlève 1958). Le caractère dimensionnant du début de l'année, du mois de janvier donc, pourrait s'expliquer par deux phénomènes. Premièrement on voit peut-être l'effet du temps nécessaire au tank pour se remplir, qui serait la période la plus difficile à passer. La deuxième raison serait que janvier soit le mois le plus dur de l'année, que pendant cette période la production d'énergie par les ressources soit la plus faible.

Pour vérifier ces hypothèses, on récupère la production de puissance par les ressources (P_{RE}) lors de la simulation, décomposée en ce qui vient du solaire et ce qui vient de l'éolien. On fait alors une moyenne mesure par mesure, sur toutes les années, qu'on lisse sur deux jours, afin d'avoir la tendance des trois mois. On obtient la figure 6.

Sur les trois mois, la part de puissance fournie par le vent a tendance à diminuer, mais celle fournie par le soleil augmente largement, faisant du mois de janvier le plus difficile en termes de production d'énergie.

Si on s'intéresse aux périodes sans énergie, on doit revenir aux années une par une. On peut regarder d'une part l'année la plus difficile, 1959 (*i.e.* coût optimal maximum : 88500 euros), et d'autre part l'année la plus favorable, 1966 (*i.e.* coût optimal minimum : 40369 euros). La puissance produite pour 1959 pour une configuration "classique" (8 PV, 2 W/G, 5 batteries) est en figure 7, celle pour 1966 est en figure 8.

Le début de 1959 semble particulièrement difficile : on a dans les quatre premiers jours (16 premières mesures) une grosse période sans vent, on a donc très peu de production d'énergie, voire pas de production du tout pendant la nuit. La configuration utilisée ne permet pas de répondre à la demande et rajouter des panneaux solaires et des éoliennes ne changerait rien, le vent étant trop faible pour que les éoliennes fonctionnent et l'énergie solaire étant très basse. La solution vient alors des batteries, qui sont en partie chargées au début de la simulation. L'année plus facile, 1966, a un début d'année plus riche en énergie produite : on a un vent très fort sur les premiers jours. Augmenter les nombres d'éoliennes permet de constituer des réserves suffisantes pour répondre à la demande pendant de futures périodes difficiles.

Pour conclure, le début de la série semble particulièrement dimensionnant pour l'optimisation de la configuration du système. Sans permettre d'expliquer toute la variabilité (l'histogramme "1 year" montrant plus de variabilité que celui des permutations), les premières mesures de la série peuvent donner des cas très extrêmes si les conditions ne sont pas favorables. Il pourrait alors être conseillé de commencer avec un tank plein, ou de

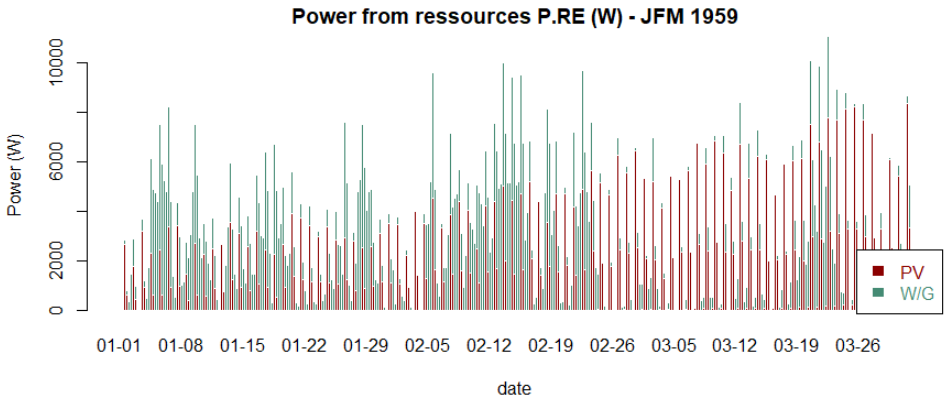

FIGURE 7 – Sorties de simulation pour JFM 1959

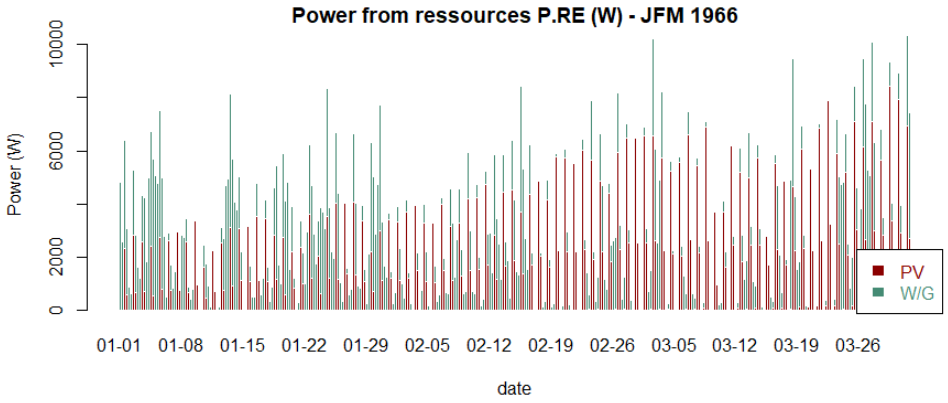

FIGURE 8 – Sorties de simulation pour JFM 1966

laisser le système fonctionner sans fournir d'eau le temps que le tank se remplisse. Le cas où on considère qu'on commence avec un tank plein a été en partie traité en annexe 2.

3.1.3 Résultats

3.1.3.1 Méthodes classiques : comparaison de moyennes mensuelles et méthode Sandia

Les méthodes dites classiques, présentées en section 2.1, ont été testées sur les données du projet DESIRES, sur toute l'année, et quelques sorties sont présentées en figure 9. On y trouve les moyennes mensuelles de chaque année, celles de toutes les années (le long terme) et enfin les moyennes mensuelles de l'année typique sélectionnée.

On constate qu'avec la comparaison de moyennes on a bien une année typique dont les moyennes mensuelles collent à celles du long terme, alors qu'avec la méthode Sandia, on s'éloigne parfois de la moyenne mensuelle, mais la distribution est plus proche de celle du long terme.

3.1.3.2 Années réelles et sélection par les distances proposées

Pour chaque distance, on mesure la distance entre toutes les séries courtes (un mois * une année) et la série longue, qui est en fait tous les mois * année mis bouts à bouts. La série courte sélectionnée est alors celle minimisant la distance concernée.

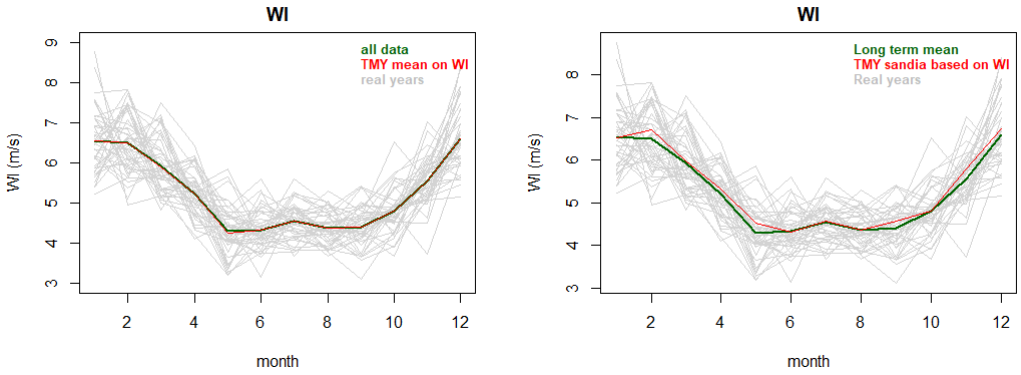

FIGURE 9 – Année typique sélectionnée par comparaison de moyennes mensuelles (à gauche) et par la méthode Sandia (à droite) pour l'intensité du vent.

On récupère ainsi les trois mois typiques sélectionnés par chaque mesure de distance, sur lesquels on fait tourner la simulation. On obtient les coûts et configurations optimales présentés en table 4 avec l'algorithme d'optimisation.

N.PV	N.s	N.WG	N.BAT	cost	meth	TMM
10	4	4	3	47921.757	ks	1960 01 - 1958 02 - 1981 03
...	...	...	...	...	...	...

TABLE 4 – Tableau récapitulatif des configurations et coûts issus de l'optimisation du dimensionnement sur les mois de janvier, février et mars sélectionnés par les différentes distances

Sur la figure 10, on trouve deux choses. Tout d'abord on a l'histogramme du coût optimal obtenu par des simulations faites sur les 43 années permutées du jeu de données originel, en ne prenant que les 3 premiers mois de l'année. On a choisi de travailler avec cet histogramme car il permet de représenter une variabilité de la réponse qui n'est pas seulement due au début de l'année. Les points correspondent au coût optimal obtenu par simulation sur les 3 premiers mois de l'année qui ont été sélectionnés par chaque distance.

Dans la partie précédente on a vu que la zone autour de 50000 euros correspond à la situation la plus "classique", alors que le pic à 87000 euros est dû à un début de série sans production d'énergie.

Il est difficile de statuer pour les distances sélectionnant des séries donnant un coût autour de 87000 euros : on sait qu'aucune distance ne favorise le début de l'année par rapport au reste, ces distances-là ont donc eu simplement la mauvaise idée de sélectionner un mois de janvier avec un début difficile.

La méthode de comparaison de moyenne, les méthodes de compression, la méthode basée sur la prédiction, et les distances PACF et EDR ont toutes largement sous estimé le coût optimal.

Dans les distances qui restent, on a d'une part celles qui sélectionnent des années donnant un coût proche de ce qui se passe dans la plupart des cas (si on exclut les cas avec début d'année difficile), c'est-à-dire des distances comme KS, Sandia, Energy, CVM ou ACF. D'autre part on a des distances plus conservatrices au sens où elles sélectionnent des années plus difficiles, qui donnent des coûts plus élevés, comme la distance de Maha-

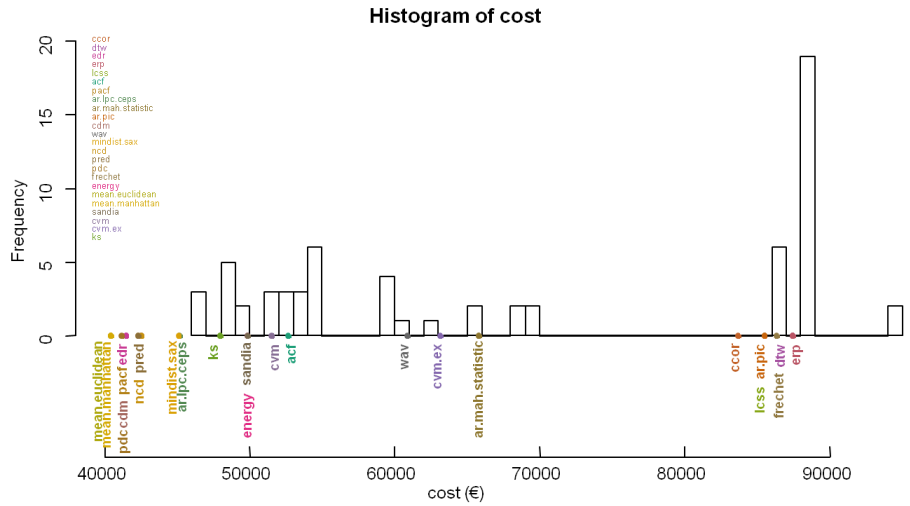

FIGURE 10 – Histogramme du coût estimé par permutation des années, avec les valeurs des coûts données par les simulation faites sur les 3 mois sélectionnés par chaque distance

raj, Anderson-Darling (CVM.ex) et la distance basée sur la décomposition en ondelettes (Wav).

Conclusion sur le projet DESIRES Avec ce projet on a voulu se centrer sur la question du choix de la mesure de distance. En effet on se trouvait face à une série de 43 ans, donc avec suffisamment de candidats, et les méthodes de géneration par blocs de classes paraissaient moins appropriées que dans le cas du projet MEDISA, donnant des résultats peu satisfaisants notamment en raison de la difficulté d'interprétation des classes. Au vu des résultats on semble pouvoir avoir un premier avis sur l'utilisation des différentes distances pour ce projet : certaines permettraient de sélectionner des années classiques, d'autres des années conservatrices, et enfin on repère des distances qui sous estiment le coût du système. On note que la méthode Sandia donne des résultats très proches de ce qui se passe avec la série entière, et sa rapidité de calcul s'avère être un avantage indéniable.

3.2 Projet MEDISA : méthodologie de dimensionnement de systèmes d'assainissement

3.2.1 Présentation du projet et des données

Le projet MEDISA vise à dimensionner des réseaux de récupération d'eau de pluie. On se place dans un cas simplifié où on considère le schéma présenté en figure 11, avec un bassin versant rendant toute la pluie reçue. L'eau en provennance de ce bassin Q_e est la seule entrée du bassin de rétention, qui a deux sorties : une pompe qui mène un flux constant Q_s à la station de traitement de l'eau, et un trop-plein qui déverse en milieu naturel le flux Q_{DEV}. Le système est simulé par le logiciel EPA SWMM [1], mais afin de pouvoir faire tourner le système en boucle, un code simplifié a été écrit, qui donne des sorties jugées suffisamment proches de celles du logiciel.

Le fonctionnement de la simulation est décrit ci-dessous.

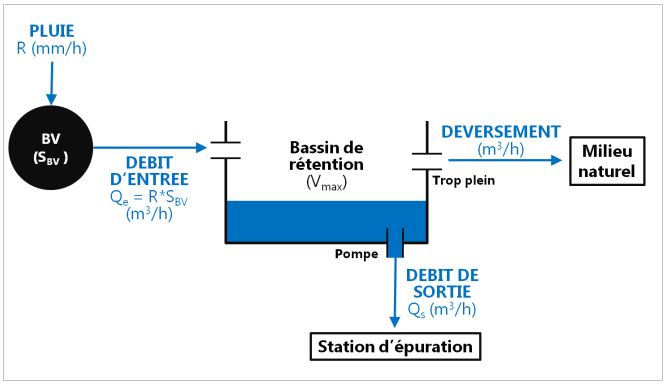

FIGURE 11 – Schéma du fonctionnement d'un réseau simplifié de récupération d'eau de pluie

A l'heure t, on peut écrire le volume potentiellement présent dans le bassin (*i.e.* si on ne tient pas compte du déversement), $V(t)$, et le volume réellement présent dans le bassin, $V_{reel}(t)$:

$$V(t) = V(t-1) \ + \ Q_e(t) \ - \ Q_s$$

$$V_{reel}(t) = min\left\{V(t), V_{max}\right\}$$

On a donc le déversement suivant :

$$Q_{DEV}(t) = V(t) - V_{reel}(t)$$

Les valeurs utilisées pour la simulation sont les suivantes :
$Q_s = 150\ l/s$
$S_{BV} = 6.10^5\ m^2$
$V_{max} = 9.10^3\ m^3$
$Q_e(t) = 10^{-3}.R(t).S_{BV}\ (m^3/h)$
R est la chronique de pluie en mm/h

Ce sont les déversements dans le trop-plein qui nous intéressent, la législation proposant de prendre comme critère le nombre de jours de déversement par an. Un jour est compté comme un jour de déversement si on a au moins lors d'une des mesures de la journée un déversement non nul.

La chronique de pluie est représentée en figure 12. C'est une variable particulière puisqu'on a deux états, présence et absence, avec dans l'un des deux une distribution d'intensité de pluie. On peut voir la chronique comme une succession dévénements, ce qui est souvent fait lorsqu'on traite des données de pluie [23] [20], la série se prête donc bien à une classification.

On se propose donc de tester les méthodes de construction d'années typiques suivantes :

Etape 1) Extraction de séries courtes :

- Années réelles
- Générateur par blocs de classes

Etape 2) Sélection de la série typique :

- Distances listées en section 2

Les méthodes seront testées uniquement sur les trois premiers mois de l'année pour des raisons de temps de calcul.

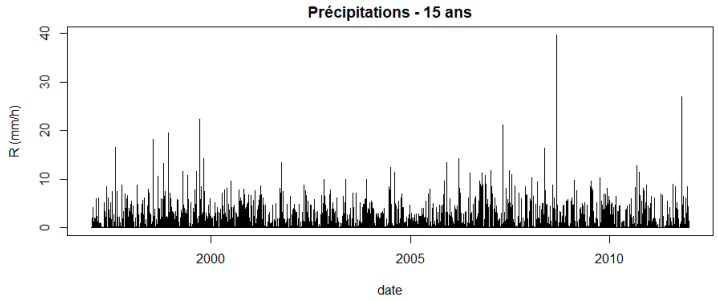

FIGURE 12 – Chronique de pluie en mm/h, 1997-2011

3.2.2 Etude de la variabilité de la réponse

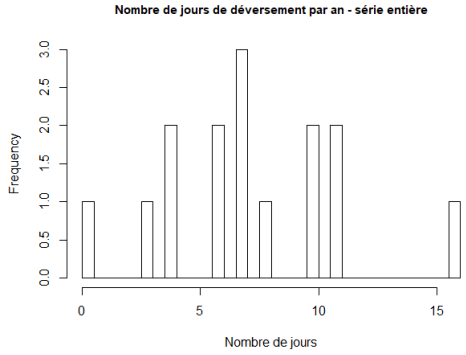

FIGURE 13 – Histogramme du nombre de jours de déversement par an sur 3 mois (JFM) sur la série entière

Dans le cas de ce projet, la sortie qui nous permettra de valider ou d'invalider la série typique sélectionnée est le nombre de jours de déversement par an. En faisant tourner la simulation sur la série entière, on constate qu'on a en moyenne 7,33 jours de déversement par an. De cette simulation on tire non pas une valeur mais une distribution (une valeur par an), cf. figure 13. Si on fait tourner la simulation sur les années une par une, on trouve exactement le même nombre de jours de déversement par an que si on simule sur la série entière. Il n'y a donc pas d'influence d'une année sur l'autre.

Sur la figure 13, on constate la présence de deux années extrêmes (0 et 16 jours de déversement). Les autres années se répartissent autour de la valeur moyenne, on y distingue trois groupes : les années moyennes (autour de 7 jours), les difficiles (10 jours) et les plus faciles (4 jours).

3.2.3 Résultats

3.2.3.1 Extraction et génération des candidats

Pour ce projet, pour ce qui est de l'extraction des séries courtes, on a d'une part les séries réelles et des séries générées d'autre part.

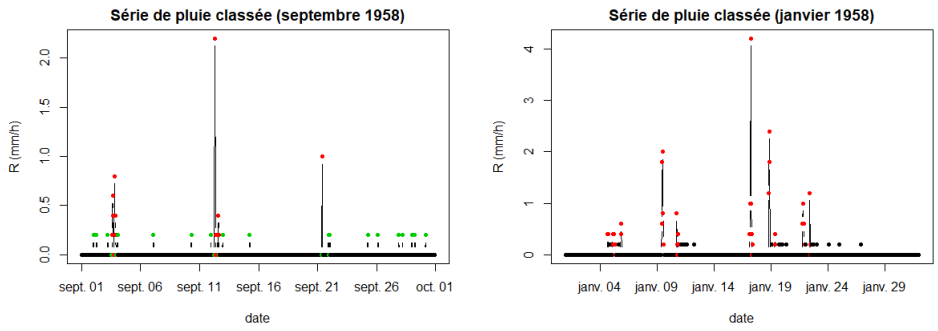

FIGURE 14 – Série d'états de pluie : septembre à gauche, janvier à droite

Dans le cas des séries extraites les candidats sont simplement les années réelles, comme dans le cas du projet DESIRES.

Les séries générées l'ont été avec la méthodologie proposée en section 2.2.1. La série de pluie a d'abord été stratifiée afin d'avoir une variable de classe. Les classes d'observations ainsi définies étaient les suivantes : une classe pour le temps sec (<0.2 mm/h), et par temps de pluie 4 classes correspondant aux quantiles 0.25, 0.5, 0.75 et 1. Sur ces observations on ajuste un modèle HMM par mois, et on en récupère une série d'états. Le nombre d'états cachés a été choisi par minimisation du critère BIC (on teste de 2 à 5 états). Le mois de septembre par exemple donne 3 états. Un extrait de la série coloré en fonction de l'état est présenté en figure 14, à gauche. Les états sont les suivants : d'une part le temps sec, d'autre part le temps de pluie séparé entre les évènements brefs de faible intensité et les évènements longs ou de forte intensité. Beaucoup de mois, comme janvier (figure 14 à droite), n'ont que deux états : temps sec vs. temps pluvieux. Il faut noter que les évènements brefs de faible intensité sont alors considérés comme du temps sec.

30 séries ont été générées par mois à partir de ces classifications en états.

Un exemple de série ainsi générée avec la classification HMM du mois de janvier est présenté en figure 16, à droite. On retrouve une série semblable aux séries réelles, avec des périodes de temps sec et de temps pluvieux, et on retrouve bien des évènements des différentes intensités. Si on fait tourner la simulation sur les années générées, on trouve des nombres de jours de déversement dont la variance n'est pas significativement différente de celle des années réelles. Les histogrammes des nombres de jours de déversement des années réelles et des séries générées sont représentés en figure 15, et on constate que la variabilité est bien représentée, même si les deux extrêmes qu'on trouvait dans les années réelles ne se retrouvent pas dans les séries générées. Il est possible que le générateur utilisé ait du mal à donner des séries aussi particulières. Il en effet est connu que les générateurs stochastiques de condition météo ont tendance à sous-estimer la variabilitÃ© inter-annuelle.

En figure 16, on trouve deux exemples (qui donnent le même nombre de jours de déversement) de séries, une extraite à gauche et une générée à droite, pour le mois de janvier. Les points en rouge correspondent aux mesures où il y a déversement. Ces séries seront candidates à être des mois typiques.

3.2.3.2 Sélection de la série la plus typique

Pour ce qui est de la sélection de la série, les distances présentées en section 2.2.2 ont été testées sur les deux types de séries courtes (réelles et générées).

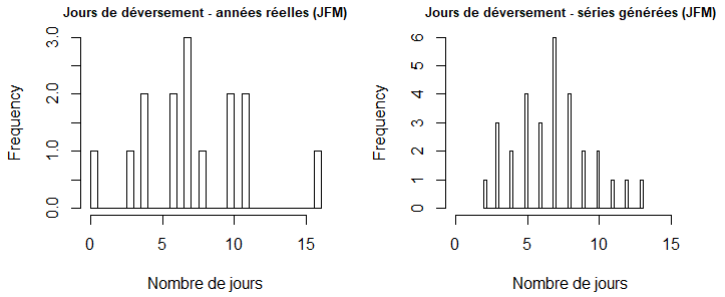

FIGURE 15 – Histogramme du nombre de jours de déversement pour les années réelles à gauche et pour les séries générées à droite

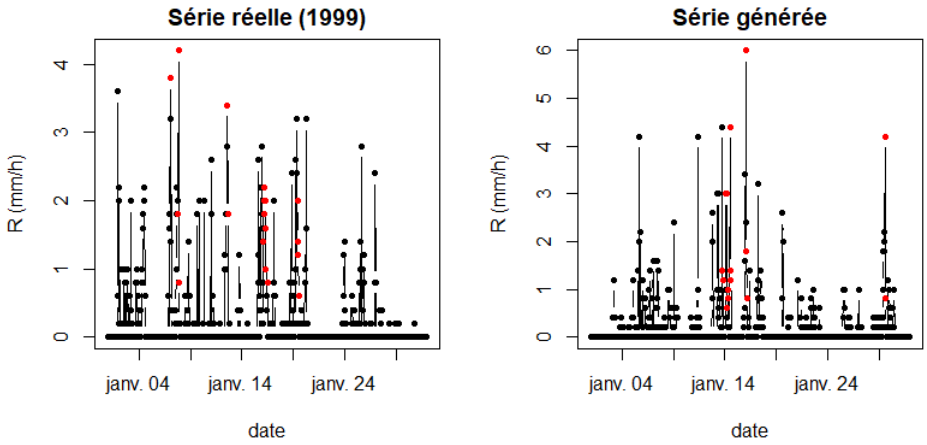

FIGURE 16 – Exemples de séries candidates pour le mois de janvier : à gauche une année réelle, à droite une série générée

Afin d'économiser du temps de calcul, les 24 distances présentées précédemment n'ont pas toutes été testées. Certaines ont été tout de suite éliminées car elle se basaient sur des modèles linéaires, qui ne peuvent fonctionner correctement sur de la pluie, qui est une variable de type présence absence avec une distribution en cas de présence. Ainsi les distances basées sur des modèles auto-régressifs ont été abandonnées. On a ensuite choisi de se baser sur les résultats obtenus avec l'application précédente. Les distances suivantes sont ainsi exclues : mean.euclidean, mean.manhattan, PCD, CDM, PCAF et EDR. Enfin les distances LCSS, Energy et Fréchet n'ont pas pu être utilisées car trop coûteuses en mémoire.

Les distances restantes (ACF, CCor, SAX, NCD, KS, Wav, DTW, ERP, CVM, AD, KL et Sandia) ont été testées d'abord sur les séries réelles, puis sur les séries générées. Sur les mois ainsi sélectionnés la simulation a été réalisée, on en récupère un nombre de jours de déversement. De la même façon que dans le projet DESIRES, ces nombres ont été représentés sur l'histogramme du nombre de jours de déversement par an réalisé sur la série entière (celui de la figure 13). Le graphique est en figure 17, on y trouve en bas les points correspondant aux nombres de jours de déversement trouvés avec les mois sélectionnés par chaque distance parmi les séries réelles, et en haut la même chose mais pour les mois sélectionnés parmi les séries générées.

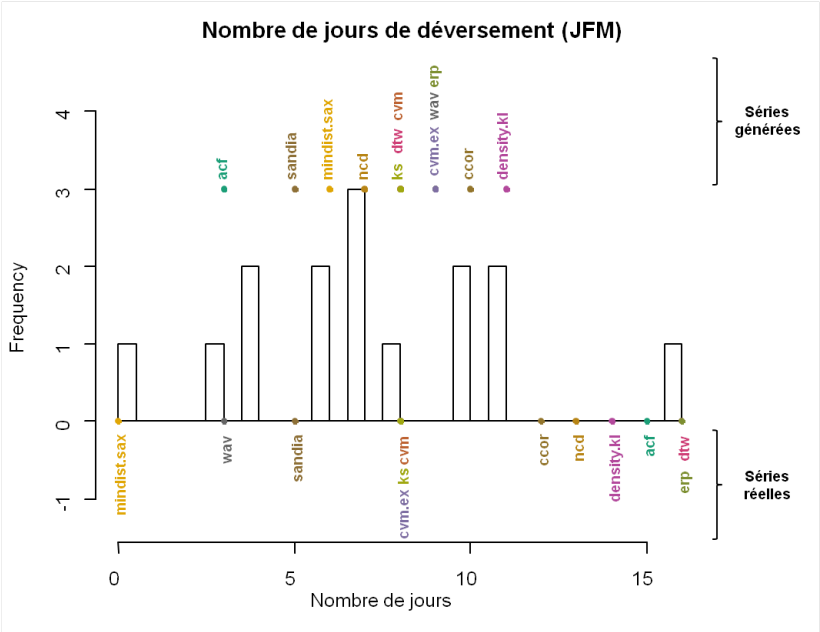

FIGURE 17 – Histogramme du nombre de jours de déversement et points des valeurs pour les mois typiques sélectionnés par les différentes distances parmi les années réelles (en bas) et les années générées (en haut)

Si on s'intéresse d'abord aux séries réelles, on constate que beaucoup de distances sélectionnent des mois donnant un nombre de jours de déversement assez élevé : CCor, NCD, ACF, ERP, DTW, KL. Dans les distances proches de la moyenne, on retrouve des distances basées sur la distribution : Sandia, CVM, Anderson Darling et KS. La distance basée sur la représentation de SAX donne une sortie très faible.

Quand on passe aux séries générées, une chose saute aux yeux : on a moins de variabilité dans les nombres de jours de déversement, ce qui pourrait être dû au fait que dans les années réelles, le fait d'avoir peu de choix faisait qu'on se retrouvait avec des mois sélectionnés ayant des sorties plus extrêmes. On a tout de même des distances qui sélectionnent des séries qui sous ou sur estiment la sortie.

On remarque que certaines distances choisissent des mois qui donnent le même nombre de jours de déversement que lors de la sélection parmi les années réelles : Sandia, CVM, KS, Anderson Darling. Ce sont les distances dont les mois typiques donnaient des sorties proches de la moyenne. Pour les autres distances, le nombre de jours de déversement change quand on passe des années réelles aux séries générées, ce qui paraît problématique car cela laisse à penser que la sortie n'est pas liée au mode de sélection de ces deux distances. Toutefois il est difficile de savoir si c'est en effet dû à une indépendance entre le mode de sélection de la distance et la simulation ou si c'est simplement dû à un plus grand choix dans le cas des séries générées. On distingue les distances pour lesquelles le nombre de jours de déversement change assez peu et va dans le sens de se rapprocher de la moyenne quand on passe aux séries générées et donc à plus de choix (CCor, density.kl), et celles dont le changement est plus radical (ACF notamment).

De ces résultats on peut conclure plusieurs choses. D'une part on constate que le générateur par blocs de classes semble satisfaisant puisque les séries qui en sortent donnent des sorties de simulation semblables à celles obtenues avec les séries réelles. L'avantage des

séries générées est double : il est facile d'obtenir de nombreuses séries, et elles contiennent à la fois la variabilité intra-annuelle et la variabilité inter-annuelle.

Les distances étudiées sélectionnent des séries variables en terme de sortie de simulation, et on repère ainsi des séries sélectionnées plus conservatrices que d'autres, au sens où le nombre de jours de déversement est plus élevé. On peut retenir la méthode Sandia et les distances CVM et KS qui donnent un nombre de jours de déversement proche de ce qui se passe dans la plupart des cas et qui donnent de plus les mêmes sorties sur es deux types de séries (réelles et générées).

4 Conclusion : Discussion et perspectives

Discussion sur les applications

L'intérêt de l'utilisation d'un générateur par blocs de classes dans le projet MEDISA reposait principalement sur la connaissance du fait que la forme des évènements pluivieux ait un impact important sur le déversement. On retrouve dans cette l'idée le principe de considérer les chroniques de pluie comme une succession d'évènements. Toutefois on avait dans notre cas un pas de temps d'une heure, et un évènement durant généralement entre 1 et 2 heures à Brest, il est impossible de voir la forme des évènements pluivieux. La classification fonctionnait tout de même assez bien car on séparait en temps sec et pluie avec éventuellement une différenciation selon l'intensité. On se retrouvait à mélanger les évènements, ce qui marchait plutôt bien. Mais avec une chronique plus fine, on pourrait déterminer quels sont les évènements dimensionnants, ceux auxquels il faudrait prêter attention.

Pour le projet DESIRES, on a pu mettre en évidence l'importance du début de l'année dans le dimensionnement du système. On se demande alors si dans le cadre du projet il serait envisageable de commencer avec un tank plein. On pourrait alors mieux voir ce qui est dimensionnant dans les chroniques en dehors du début d'année.

Manques, choses non testées

Plusieurs choses n'ont pas pu être traitées par manque de temps, et souvent à cause de problèmes de mémoire et de temps de calcul. Tout d'abord pour les deux applications on s'est restreint aux trois premiers mois de l'année. Ensuite toutes les distances relevées n'ont pas pu être testées, et plusieurs tests de sensibilité du processus de sélection d'année typique n'ont pas pu être réalisés.

On aurait notamment pu s'intéresser à la sensibilité des résultats aux poids des variables. Pour le projet DESIRES on était dans un cadre multivarié, et on a toujours accordé les mêmes poids aux trois variables. Plusieurs méthodes d'attribution de poids ont été envisagées : la littérature propose des valeurs [14], souvent adaptées à une base de données en particulier [15], mais il est aussi possible d'en attribuer automatiquement, par exemple avec un algorithme génétique [7] ou avec une ACP en utilisant les composantes principales comme variables, pondérées par les valeurs propres [31] (cette méthode a été testée avec la sélection par méthode Sandia et par comparaison de moyenne sur des données de vent et vagues).

Dans le cas des séries générées par blocs de classes, il faudrait tester la consistance du choix des distances en échantillonnant les séries générées et en regardant la stabilité de la sortie des séries sélectionnées.

Autres méthodologies envisagées

Une méthode qui a été envisagée est celle présentée dans [25]. Elle consiste à classer les séries courtes et à choisir les séries les plus proches des centres des classes. On a ainsi autant d'années typiques que de classes. Cette méthode n'a pas été traitée en priorité d'une part à cause de sa spécificité qui consiste à sélectionner plusieurs années typiques, mais aussi car la distance préconisée est un mélange de DTW et d'une distance euclidienne calculée sur la sortie du modèle. Or inclure le modèle dans la construction de l'année typique n'est pas réalisable dans un cas réel.

Autres questions non traitées

Deux questions seraient à traiter dans le futur.

Premièrement il y a certains cas où on peut ou veut prendre plus d'une année, par exemple dans le projet MEDISA la législation indique que le dimensionnement devrait se baser sur 5 années. Comment doit-on alors choisir ces 5 années ? Veut-on les 5 années les plus typiques, où souhaite-t-on inclure des années plus extrêmes ? On aurait notamment voulu adapter la méthode présentée dans [25] dont on a parlé à l'instant.

La deuxième question est celle de la spatialisation. Dans les deux projets c'est toute une zone qui est considérée, il faudrait alors étudier la façon d'intégrer cette zone et non pas un seul point.

Conclusion générale

A partir d'un état de l'art, on a pu tirer les grandes lignes de la construction d'années typiques. Des changements ont été proposés aux deux étapes ainsi identifiées : l'extraction et la sélection des séries courtes.

Pour ce qui est de la construction des séries courtes candidates, se pose la question de savoir si on veut extraire ou générer les séries. La notion de types de temps ou d'évènements entre en jeu, on va chercher à savoir si la chronique est adaptée à une classification. La longueur de la série, qui détermine le nombre de candidats dans le cas d'une simple extraction, est aussi à considérer.

La question de la sélection de la série la plus typique revient à un choix de mesure de dissimilarité. La première chose à étudier est la (les) variable(s) qui va (vont) servir dans la sélection : quelles sont-elles, quel poids leur accorder, ont-elles des particularités à prendre en compte ? L'application qui utilisera ces données est aussi à prendre en compte : le modèle est-il sensible à la structure temporelle, à la distribution des valeurs, à des évènements particuliers ? Les dissimilarités proposées peuvent être classées comme suit :

- Distances basées sur la structure de la série
  - Modèles auto-régressifs : *ar.pic, ar.lcp.ceps, ar.mah.statistic, pred*
  - Mesures élastiques (warping) : *DTW, Fréchet, EDR, ERP, LCSS*
  - Autre caractéristique : *NCD, CDM, PDC, ACF, PACF, CCor*
- Distances basées sur les valeurs de la série
  - Distribution : *KL, CVM, Sandia, KS, AD*
  - Autre caractéristique : *Comparaison de moyenne, ondelettes, spectralGLK*

Avec ce qui a été testé dans les deux applications, on a ainsi pu identifier les questions à se poser lorsqu'on souhaite construire une année typique, et vers quel type de méthodologie se diriger.

Références

[1] United States Environmental Protection Agency. Storm water management model SWMM. *https://www.epa.gov/water-research/storm-water-management-model-swmmresources*.

[2] Pierre Ailliot and Valérie Monbet. Markov-switching autoregressive models for wind time series. *Environmental Modelling & Software*, 30 :92–101, 2012.

[3] Bo Andersen, S Eidorff, H Lund, E Pedersen, S Rosenorn, and O Valbjorn. Meteorological data for design of building and installation : a reference year (extract). Technical report, report, 1977.

[4] Theodore W Anderson and Donald A Darling. Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. *The annals of mathematical statistics*, pages 193–212, 1952.

[5] RF Benseman and FW Cook. Solar radiation in new zealand-standard year and radiation on inclined slopes. *New Zealand Journal of Science*, 12(4) :696, 1969.

[6] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994.

[7] ALS Chan. Generation of typical meteorological years using genetic algorithm for different energy systems. *Renewable Energy*, 90 :1–13, 2016.

[8] Lei Chen and Raymond Ng. On the marriage of lp-norms and edit distance. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 792–803. VLDB Endowment, 2004.

[9] Donald A Darling. The kolmogorov-smirnov, cramer-von mises tests. *The Annals of Mathematical Statistics*, 28(4) :823–838, 1957.

[10] Jianqing Fan and Wenyang Zhang. Generalised likelihood ratio tests for spectral density. *Biometrika*, 91(1) :195–209, 2004.

[11] Roberto Festa and Corrado F Ratto. Proposal of a numerical procedure to select reference years. *Solar Energy*, 50(1) :9–17, 1993.

[12] Jack M Finkelstein and Ray E Schafer. Improved goodness-of-fit tests. *Biometrika*, pages 641–645, 1971.

[13] M Maurice Fréchet. Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, 22(1) :1–72, 1906.

[14] Irving J Hall, RR Prairie, HE Anderson, and EC Boes. Generation of a typical meteorological year. Technical report, Sandia Labs., Albuquerque, NM (USA), 1978.

[15] YJ Huang. International weather for energy calculations (iwec weather files) users manual, 2011.

[16] Serm Janjai and P Deeyai. Comparison of methods for generating typical meteorological year using meteorological data from a tropical environment. *Applied Energy*, 86(4) :528–537, 2009.

[17] Eamonn Keogh, Stefano Lonardi, and Chotirat Ann Ratanamahatana. Towards parameter-free data mining. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 206–215. ACM, 2004.

[18] E Koutroulis and D Kolokotsa. Design optimization of desalination systems power-supplied by PV and W/G energy sources. *Desalination*, 258(1) :171–181, 2010.

[19] Eftichios Koutroulis, Dionissia Kolokotsa, Antonis Potirakis, and Kostas Kalaitzakis. Methodology for optimal sizing of stand-alone photovoltaic/wind-generator systems using genetic algorithms. *Solar energy*, 80(9) :1072–1088, 2006.

[20] Demetris Koutsoyiannis, Demosthenes Kozonis, and Alexandros Manetas. A mathematical framework for studying rainfall intensity-duration-frequency relationships. *Journal of Hydrology*, 206(1-2) :118–135, 1998.

[21] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1) :79–86, 1951.

[22] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 2–11. ACM, 2003.

[23] María-Carmen Llasat. An objective classification of rainfall events on the basis of their convective features : application to rainfall intensity in the northeast of spain. *International Journal of Climatology*, 21(11) :1385–1400, 2001.

[24] Usue Mori, Alexander Mendiburu, and Jose A Lozano. Distance measures for time series in r : The tsdist package. *R JOURNAL*, 8(2) :451–459, 2016.

[25] Victor Picheny, Ronan Trépos, and Pierre Casadebaig. Optimization of black-box models with uncertain climatic inputsâapplication to sunflower ideotype design. *PloS one*, 12(5) :e0176815, 2017.

[26] Gábor J Székely and Maria L Rizzo. Energy statistics : A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8) :1249–1272, 2013.

[27] Sakari M Uppala, PW Kållberg, AJ Simmons, U Andrae, V d Bechtold, M Fiorino, JK Gibson, J Haseler, A Hernandez, GA Kelly, et al. The era-40 re-analysis. *Quarterly Journal of the royal meteorological society*, 131(612) :2961–3012, 2005.

[28] Michail Vlachos, George Kollios, and Dimitrios Gunopulos. Discovering similar multidimensional trajectories. In *Data Engineering, 2002. Proceedings. 18th International Conference on*, pages 673–684. IEEE, 2002.

[29] Robert A Wagner and Michael J Fischer. The string-to-string correction problem. *Journal of the ACM (JACM)*, 21(1) :168–173, 1974.

[30] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network : A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.

[31] Liu Yang, Kevin KW Wan, Danny HW Li, and Joseph C Lam. A new method to develop typical weather years in different climates for building energy use studies. *Energy*, 36(10) :6121–6129, 2011.

[32] Hui Zhang, Tu Bao Ho, Yang Zhang, and M-S Lin. Unsupervised feature extraction for time series clustering using orthogonal wavelet transform. *Informatica*, 30(3), 2006.

[33] Walter Zucchini and Peter Guttorp. A hidden markov model for space-time precipitation. *Water Resources Research*, 27(8) :1917–1923, 1991.

[34] Walter Zucchini and Iain L MacDonald. *Hidden Markov models for time series : an introduction using R*, volume 22. CRC press Boca Raton, 2009.

Annexes

Annexe I : Simulation et dimensionnement d'une usine de désalinisation (projet DESIRES)

La description du système est tirée essentiellement du papier [2]. On peut schématiser le fonctionnement du système comme c'est indiqué en figure 1.

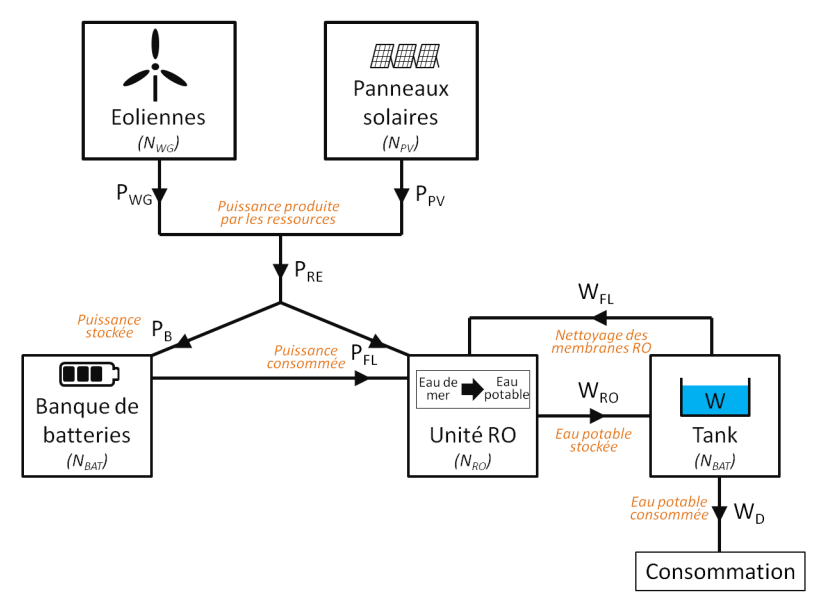

Figure 1 – Schéma simplifié du fonctionnement de l'usine de désalinisation

1. Simulation

Puissance produite par les panneaux photovoltaïques (PV)

La puissance maximale (en W) produite par un module PV peut s'écrire comme suit :

$$P_M^i(t,\beta) = N_s N_p \left[V_{OC.STC} - K_V.(T_C^i(t) - 25°C)\right] . \left[I_{SC.STC} - K_I.(T_C^i(t) - 25°C)\right] . \frac{G^i(t,\beta)}{1000} . FF^i(t)$$

Où

$$T_C^i(t) = T_A^i(t) + \frac{NOCT - 20°C}{800} . G^i(t,\beta)$$

$$FF^i(t) = \frac{P_{max}}{V_{OC.STC}.I_{SC.STC}}$$

Avec :
$FF^i(t)$ le Fill Factor défini dans [3]
$T_A^i(t)$ la température ambiante ($°C$)
$G^i(t,\beta)$ la radiation à l'heure t du jour i sur un panneau d'angle β (en W/m^2)
$N_p + N_s = N_{PV}$ le nombre de modules PV
$V_{OC.STC} = 21V$; $I_{SC.STC} = 7.22A$; $K_V = -2.3mV/°C$;$K_I = 6\nu A/°C$

La puissance transférée à la batterie par un module PV est :

$$P_{PV}^i(t,\beta) \;\; tel\ que \;\; n_S = \frac{P_{PV}^i(t,\beta)}{P_M^i(t,\beta)}$$

Où n_S est le facteur de conversion de puissance, avec dans notre cas $n_S = 0.96$.
Le nombre de chargeurs de batterie s'écrit comme suit :

$$N_{ch}^{PV} = \frac{N_{PV} P_{PV}^m}{P_{ch}^m}$$

Avec
N_{PV} le nombre de modules PV,
$P_{PV}^m = 110W$ la puissance maximale d'un module PV sous STC,
$P_{ch}^m = 300W$ la puissance du chargeur de batterie.

Puissance produite par les éoliennes

La fonction de transfert donnant l'énergie produite par l'éolienne en fonction de la vitesse du vent proposée par [2] demande de se référer à des tables que nous ne possédons pas. Il a été décidé d'utiliser les caractéristiques d'un éolienne "classique" : [1], EOL Force 1 (cf. figure 2).

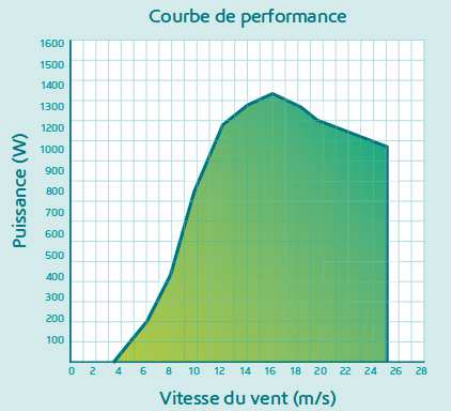

FIGURE 2 – Courbe de performance de EOL Force 1

La puissance (en W) se calcule comme suit :

Si $v^i(t,h) < v_{CI}$, $P_{WG}^i(t,h) = 0W$
Si $v_{CI} \leq v^i(t,h) \leq v_R$, $P_{WG}^i(t,h) = \frac{1}{2}\rho_A C_P A (v^i(t,h))^3$
Si $v_R < v^i(t,h) \leq v_{max}$, $P_{WG}^i(t,h) = P_R + (v^i(t,h) - v_R)\frac{P_{max} - P_R}{v_{max} - v_R}$
Si $v_{max} < v^i(t,h) \leq v_{CO}$, $P_{WG}^i(t,h) = P_m ax + (v^i(t,h) - v_{max})\frac{P_{CO} - P_{max}}{v_{CO} - v_{max}}$
Si $v^i(t,h) > v_{CO}$, $P_{WG}^i(t,h) = 0W$

Avec $v^i(t,h) = v_{ref}^i(t)\left(\frac{h}{h_{ref}}\right)^{\alpha}$, où $v_{ref}^i(t)$ et h_{ref} sont la vitesse du vent et la hauteur à laquelle elle est mesurée dans les données.

On a :
$P_R = \frac{1}{2}\rho_A C_P A (v_R)^3$ qui donne $\frac{1}{2}\rho_A C_P A$
$v_{CI} = 3.5m/s$; $v_R = 12m/s$; $P_R = 1180W$
$v_{max} = 16m/s$; $P_{max} = 1335W$; $v_{CO} = 25m/s$; $P_{CO} = 970W$
$\alpha = 0.2$; $h_{ref} = 10m$; $h = 13m$; $v_{ref}^i(t) = WI(m/s)$

Capacité de la banque de batteries

Le nombre de batteries en série est : $n_B^s = \frac{V_{BUS}}{V_B}$, avec $V_B = 12V$ et $V_{BUS} = 24V$.

La capacité maximale de la banque de batteries s'écrit : $C_n = \frac{N_{BAT}}{n_B^s} C_B \quad (Ah)$, avec N_{BAT} le nombre total de batteries et $C_B = 230Ah$ la capacité nominale de chaque batterie. La capacité minimale autorisée de la banque en découle : $C_{min} = DOD.C_n$, avec la profondeur de décharge maximale autorisée $DOD = 80\%$.

La capacité initiale de la banque de batterie est $C^1(0) = \frac{1-DOD}{2} C_n$, et la capacité de la banque de batteries disponible à la mesure t du jour i est :

$$C^i(t) = C^i(t-1) + n_B \frac{P_B^i(t)}{V_{BUS}} \Delta t$$

Avec $n_B = 80\%$ pendant le chargement et $n_B = 100\%$ pendant le déchargement, $\Delta t = 6h$, et $P_B^i(t)$ est calculé à chaque pas de la simulation.

Unités de désalinisation (RO)

Les unités de désalinisation sont alimentées par les ressources (panneaux PV et éoliennes) et les batteries.

La puissance (en W) produite par les ressources s'écrit ainsi :

$$P_{RE}^i(t) = N_{ch}^{PV} n_S P_M^i(t, \beta) + N_{WG} P_{WG}^i(t, h)$$

La puissance d'entrée des convertisseurs DC/AC nécessaire à faire tourner les RO est :

$$P_L = \frac{P_{RO}}{n_i} = \frac{N_{RO} P_u}{n_i}$$

Avec
P_{RO} la puissance totale qui doit alimenter les unités RO pour qu'elles fonctionnent
$n_i = 0.9$ l'efficacité de la conversion DC/AC
$N_{RO} = 1$ le nombre d'unités RO
$P_u = 1120W$ la consommation de puissance AC par chaque unité RO

Le volume total d'eau désalinisée produit par les unités RO entre deux mesures (en m^3) s'écrit : $W_{RO} = N_{RO} W_u \Delta t$, avec $W_u = 0.475m^3/h$ et $\Delta t = 6h$. La quantité minimale d'eau (en m^3) dans le réservoir est définie comme : $W_{min} = 0.3 W_{TANK}$, avec $W_{TANK} = 9.592m^3$.

Le volume d'eau désalinisée disponible dans le réservoir à la mesure t du jour i est $W^i(t)$ en m^3.

Nettoyage

Lorsque les unités RO ne fonctionnent pas, les membranes des RO sont nettoyées. On consomme alors : $P_{FL} = N_{RO} P_{u,FL}$ en W, et $W_{FL} = N_{RO} W_{u,FL}$ en m^3. Dans notre cas on a $P_{u,FL} = 190,4W$ et $W_{u,FL} = 79.5l$.

Simulation

On note $W_D^i(t)$, en m^3, la demande en eau à la mesure t du jour i. Les flux de puissance et d'eau dans le système sont décrits par le fonctionnement suivant :

. . . . . .

Initialisation : `success = TRUE`

A chaque pas :

Si $P^i_{RE}(t) \geq P_L$,

Les ressources permettent de faire fonctionner les RO

$P^i_B(t+1) = P^i_{RE}(t) - P_L$, calcul de $C^i(t+1)$

Si $W_{RO} \geq W^i_D(t)$,

L'eau produite par les RO couvre la demande

$W^i(t+1) = W^i(t) + W_{RO} - W^i_D(t)$

Si $W_{RO} < W^i_D(t)$,

L'eau produite ne couvre pas la demande, on prend ce qui manque dans le tank

$W^i(t+1) = W^i(t) - (W^i_D(t) - W_{RO})$

Si $W^i(t+1) < 0$,

Le tank ne suffit pas pour couvrir la demande : `success = FALSE`

$W^i(t+1) = 0$

Si $P^i_{RE} < P_L$,

Les ressources ne permettent pas de faire fonctionner les RO, on prend ce qui manque dans les batteries

$P^i_B(t+1) = P^i_{RE}(t) - P_L$, calcul de $C^i(t+1)$

Si $C^i(t+1) \geq 0$,

Il y a assez de stock dans les batteries, les RO fonctionnent

Si $W_{RO} \geq W^i_D(t)$,

L'eau produite par les RO couvre la demande

$W^i(t+1) = W^i(t) + W_{RO} - W^i_D(t)$

Si $W_{RO} < W^i_D(t)$,

L'eau produite ne couvre pas la demande, on prend ce qui manque dans le tank

$W^i(t+1) = W^i(t) - (W^i_D(t) - W_{RO})$

Si $W(t+1) < 0$,

Le tank ne suffit pas pour couvrir la demande : `success = FALSE`

$W^i(t+1) = 0$

Si $C^i(t+1) < 0$,

Il n'y a pas assez de stock de batterie, les RO ne marchent pas, les batteries sont remplies avec les ressources, l'eau sera prise dans le tank

$P^i_B(t) = P^i_{RE}(t)$, calcul de $C^i(t+1)$

Si $W^i(t) < W^i_D(t)$,

Le tank ne suffit pas pour couvrir la demande : `success = FALSE`

$W^i(t+1) = 0$

Si $W^i(t) \geq W^i_D(t)$,

Le tank permet de couvrir la demande

$W^i(t+1) = W^i(t) - W^i_D(t)$

Si $W^i(t+1) \geq W_{FL}$ et $C^i(t+1) \geq 0.8(P_{FL}/V_{BUS})\Delta t$,

Il reste encore assez d'eau pour nettoyer les membranes

$W^i(t+1) = W^i(t+1) - W_{FL}$

$C^i(t+1) = C^i(t+1) - 0.8 * \frac{P_{FL}}{V_{BUS}}\Delta t$

A la fin de chaque pas :

Si $W^i(t+1) > W_{TANK}$, $\quad W^i(t+1) = W_{TANK}$

Si $C^i(t+1) > C_{max}$, $\quad C^i(t+1) = C_{max}$

. . . . . .

On suit ainsi le volume d'eau dans le réservoir, la capacité des batteries et la satisfaction de la demande en eau au cours de la simulation.

2. Configuration : algorithme d'optimisation

Dans [2], un algorithme génétique permet d'optimiser le dimensionnement de la plateforme. Le temps de calcul d'une simulation étant court, on propose de parcourir simplement une partie de l'espace des possibles. On délimite cette partie par les valeurs trouvées comme étant la configuration optimale pour un quartier résidentiel de 15 maisons par dans [2]. En effet en faisant tourner la simulation avec ces paramètres, on se rend compte que la plateforme est sur dimensionnée : le tank et les batteries sont en permanence pleins et si on enlève quelques éoliennes ou modules PV, la demande est toujours satisfaite.

Les paramètres à optimiser sont : $X = \{N_{PV}, N_{WG}, N_{BAT}, \beta_1, \beta_2, W_{TANK}, N_{RO}\}$. Les bêtas sont les angles des panneaux solaires sur deux périodes de l'année (été et hiver), et dans notre cas on choisit de les fixer à $\beta_1 = \beta_2 = 0$. Pour le reste, on conserve comme paramètres à optimiser N_{PV}, N_{WG} et N_{BAT}, les autres étant fixés aux valeurs données comme optimales pour un quartier résidentiel. A ces trois paramètres on est obligés d'ajouter le nombre de panneaux en série N_s, on n'a pas les valeurs nécessaires pour le calculer.

On a donc $X = \{N_{PV}, N_{WG}, N_{BAT}, N_s\}$ avec les limites suivantes : $1 \leq N_{PV} \leq 26$; $1 \leq N_{WG} \leq 15$; $1 \leq N_{BAT} \leq 12$; $1 \leq N_s \leq floor\left(\frac{N_{PV}}{2}\right)$. Toutefois, on constate que pour certaines séries on a besoin d'augmenter le nombre de batteries pour satisfaire la demande, on prend donc $1 \leq N_{BAT} \leq 30$. Pour ne pas tester toutes les combinaisons, une fois qu'on a atteint une configuration qui marche (*i.e.* le tank n'est jamais vide, `L==0`), on sort de la dernière boucle (celle sur N_{BAT}). L'algorithme est décrit ci-après :

```
.....................................
cont = TRUE
for (N.WG in 1:15)
    N.PV = 1 ; N.PV.max = 26
    while (N.PV ≤ N.PV.max)
        N.s = 1
        while (N.s ≤ (1+floor(N.PV/2)))
            N.BAT = 1
            while (cont == TRUE and N.BAT ≤ min(30, max(N.PV, N.WG)))
                simulation => L
                if (L == 0) {cont = FALSE ; N.PV.max=N.PV}
                N.BAT = N.BAT+1
            cont = TRUE ; N.s = N.s+1
        N.PV = N.PV+1
.....................................
```

On se retrouve avec un ensemble solutions qui permettront de toujours répondre à la demande en eau. La solution qui sera choisie sera la mois coûteuse sur 20 ans (coût

d'installation et 20 ans de maintenance). Le coût est calculé comme suit :

$$\begin{aligned} g(X) = & N_{PV} \times 622.94 \\ & + N_{WG} \times 2875.2 \\ & + N_{BAT} \times 1882.32 \\ & + N_{ch}^{PV} \times 128 \\ & + W_{TANK} \times 280 \\ & + N_{RO} \times 17356.17 \end{aligned}$$

On récupère ainsi une configuration optimale, au sens où elle permet de satisfaire la demande, et au sens où son coût est minimisé.

Annexe II : Résultats du projet DESIRES en démarrant la simulation avec un tank plein

1. Taille de la série

De la même façon qu'on l'a fait pour le cas avec le tank vide, on trace les histogrammes des coûts optimaux obtenus en commençant la simulation avec un tank plein, et ce d'une part pour les années une par une (1 year) et d'autre part pour les séries entières dont on enlève une année (1 year out). Ces histogrammes sont présentés en figure 3 et 4.

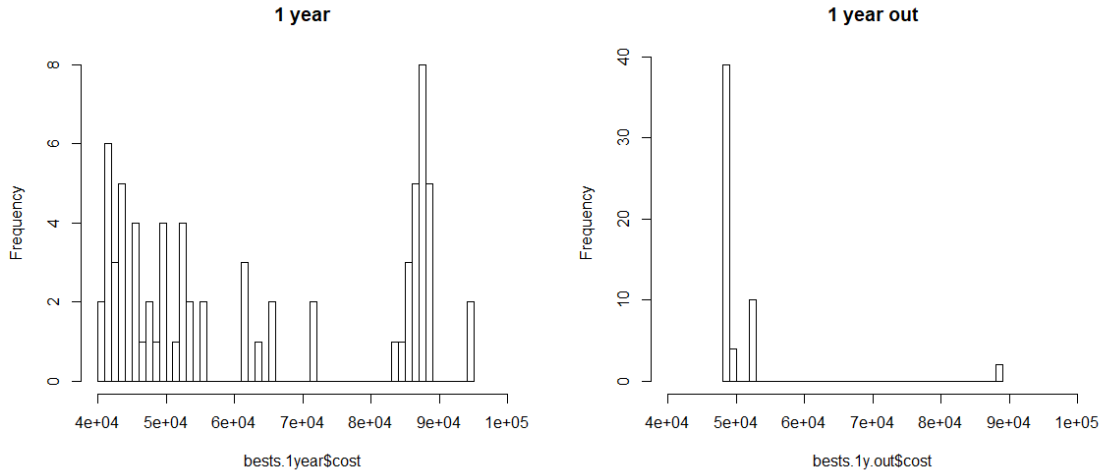

Figure 3 – Histogramme du coût pour une initialisation avec le tank vide : à gauche une année tournante, à droite une année en moins tournante

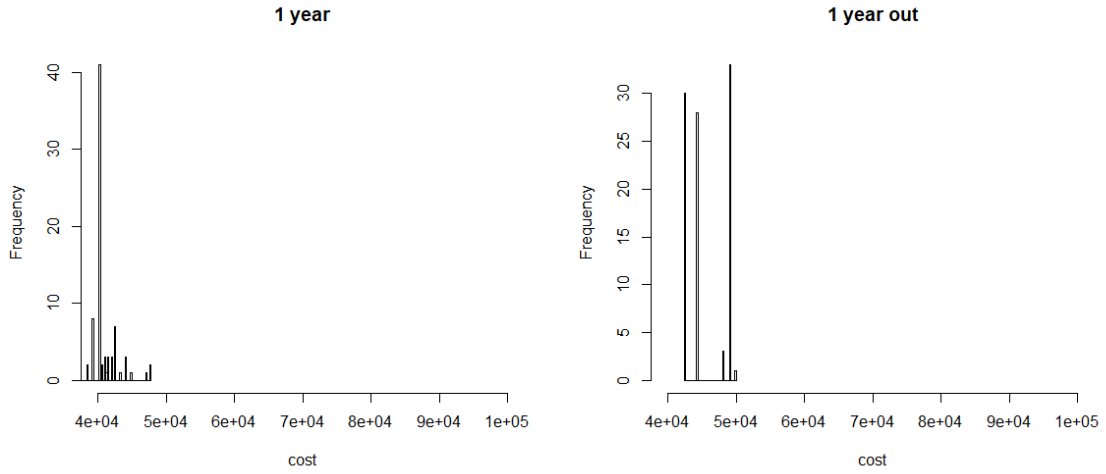

Figure 4 – Histogramme du coût pour une initialisation avec un tank plein : à gauche une année tournante, à droite une année en moins tournante

On constate alors que le pic autour de 90000 euros, qu'on avait identifié comme étant dû à des années avec un début d'année difficile, a complètement disparu dans le cas où on commence avec un tank plein. On remarque aussi que dans l'histogramme "1 year out" on a moins de variabilité, les coûts optimaux peuvent prendre autour de deux valeurs : un coût plus faible, autour de 43000 euros, et un plus élevé, autour de 49000 euros. Dans le

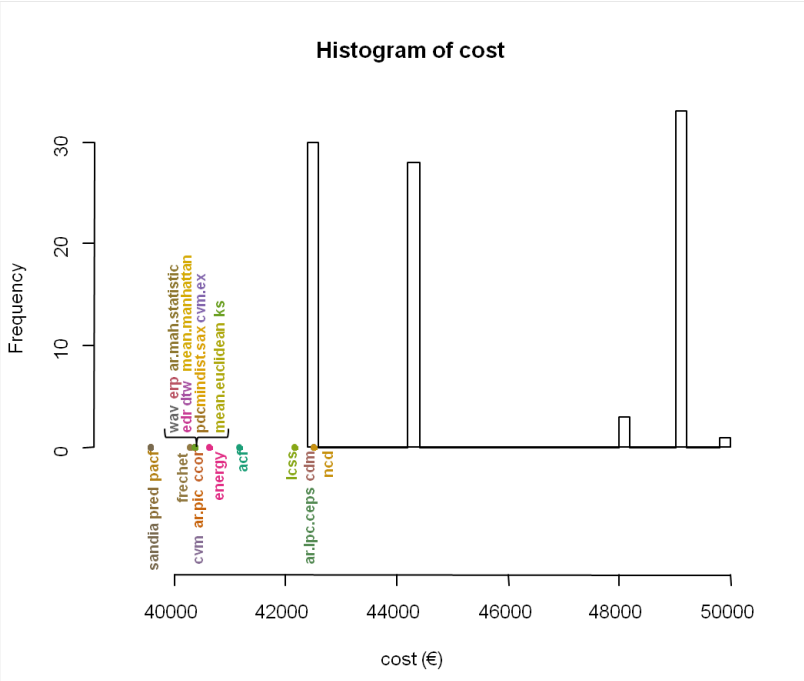

FIGURE 5 – Histogramme du coût et valeurs obtenues avec les mois sélectionnés par les différentes distances

cas où on prend une année, on a plus de variabilité mais centrée autour d'une valeur très fréquente et plus basse que le minimum de l'histogramme "1 year out". Les coûts plus faibles dans le cas où on ne prend qu'une année s'expliquent par le fait qu'on n'a pas la variabilité inter annuelle dans ces séries.

On note aussi que le coût le plus élevé pour "1 year out" ne se retrouve pas dans l'histogramme "1 year". Ceci peut s'expliquer par le fait qu'en découpant en années on peut couper au milieu d'une période difficile et donc ne pas retrouver la période la plus difficile dans les années une par une.

2. Résultats

On refait la simulation avec les mois sélectionnés par les distances (cf. section 3.1.3.2 du rapport) en commençant avec un tank plein. Dans le cas où on commençait à vide, on avait utilisé l'histogramme des coûts optimaux obtenus par permutation pour situer les distances. Ici les permutations n'ont pas pu être faites par manque de temps, on utilise donc l'histogramme obtenu en enlevant une année à la fois ("1 year out", figure 4), et on obtient ainsi la figure 5.

La majorité des distances sous-estiment largement le coût optimal. Dans les distances qui semblent plus satisfaisantes, on note LCSS, deux distances basées sur la compression de données (CDM et NCD) et une distance basée sur un modèle auto-régressif (ar.lpc.ceps).

Pour les autres distances on trouve des valeurs de l'ordre de celles obtenues en prenant les années une à une. On peut penser que c'est dû au fait que dans le cas de ce projet, le système serait principalement dimensionné par la pire période de la série. Ce point mériterait d'être vérifié, par exemple en étudiant le lien entre périodes sans production d'énergie et coût optimal.

Références

[1] Eolice. Eoliennes pour solutions d'independance energetique. *http://enerlice.fr/pdf/nos-eoliennes.pdf*.

[2] E Koutroulis and D Kolokotsa. Design optimization of desalination systems power-supplied by PV and W/G energy sources. *Desalination*, 258(1) :171–181, 2010.

[3] Tomas Markvart. *Solar electricity*, volume 6. John Wiley & Sons, 2000.

	Diplôme : Ingénieur Agronome Spécialité : Data Science Spécialisation / option : Enseignant référent : François Husson
Auteur(s) : Marie Boutigny Date de naissance* : 25/03/1994	Organisme d'accueil : IRMAR Adresse : Institut de recherche mathématique de Rennes (IRMAR – UMR CNRS 6625), Université de Rennes 1, Beaulieu - Bâtiment 22-23, 263 avenue du Général Leclerc, 35042 Rennes CEDEX, France Maître de stage : Valérie Monbet
Nb pages : 21 Annexe(s) : 8	
Année de soutenance : 2017	

Titre français : La notion d'année typique en météorologie

Titre anglais : Notion of typical year in meteorology

Résumé (1600 caractères maximum) :

La problématique des années typiques vient d'une demande des utilisateurs de données météorologiques de réduire la taille de jeux de données, trop longs pour leurs temps de calcul, en réduisant la base de données à une seule année dite typique. Grâce à un état de l'art deux étapes clefs de l'obtention d'années typiques ont été identifiées, et des modifications ont été proposées et testées au niveau de ces deux étapes afin de dégager les questions à se poser et les méthodologies vers lesquelles se tourner lorsque se pose le problème d'obtenir une année typique. La première étape consiste en la construction des candidats à l'élection de l'année typique. A la place d'une extraction d'années réelles, un générateur par blocs de classes a été proposé. La seconde étape est la sélection du candidat le plus typique. Plusieurs mesures de dissimilarités adaptées aux séries temporelles ont été proposées et étudiées.

Les différentes méthodologies ont été testées dans deux projets : le premier est le projet DESIRES, porté par ERANET/MED, qui vise à dimensionner une usine de désalinisation fonctionnant à l'énergie solaire et éolienne. Le second est le projet MEDISA, porté par Eau du Ponant, visant à dimensionner un bassin de rétention pour gérer les déversements d'eau de pluie en milieu naturel.

Abstract (1600 caractères maximum) :

The typical year problem comes from a need of meteorological data users to reduce the datasets' size because of the computational cost of their models. A state of art allowed identifying two key steps in the obtention of typical years, and changings have been suggested and tested at both levels in order to identify the questions to be asked and the methodologies to be used when it comes to obtaining typical years. The first step consists in building the candidates to the typical year election. Instead of an extraction of the real years, a generator by blocks of classes has been suggested. The second step is about chosing the most typical candidate. Several measures of dissimilarity for time series have been suggested and studied.

The different methodologies have been tested in two projects: the first one is DESIRES, supported by ERANET/MED, it aims at dimensioning a desalination system supplied by wind generators and photovoltaics. The second one is MEDISA by Eau du Ponant, which aims at dimensioning a retention basin for rainfalls.

Mots-clés : année typique, météorologie, séries temporelles

Key Words: typical year, reference year, meteorology, time series

* *Elément qui permet d'enregistrer les notices auteurs dans le catalogue des bibliothèques universitaires*

Document à intégrer au mémoire