Étude de la variabilité, Stationnarisation et corrélation spatio-temporelle

Dans ce chapitre, nous présentons dans un premier temps les différents jeux de données et leurs caractéristiques. Nous présentons ensuite une analyse de la variabilité de la production des centrales PV. La méthode de stationnarisation des séries de production est ensuite détaillée. Cette méthode a été présentée dans le premier article publié qui est fourni en annexe. Les séries stationnarisées par la méthode proposée sont évaluées par des tests de stationnarité et les performances de prévision. Enfin, nous mettons en évidence l'intérêt de la modélisation spatio-temporelle par le calcul des corrélations spatio-temporelles entre les mesures de production stationnarisées. La limite temporelle et spatiale de la propagation de ces corrélations est aussi évaluée.

2.1 Présentation des cas d'étude

Plusieurs données ont été exploitées dans le cadre de cette thèse. Nous présentons ici deux principaux jeux de données de centrales PV qui correspondent à des localisations, des puissances installées, des conditions climatiques et des répartitions géographiques différentes.

Les données "Coruscant"
Le premier jeu de données nommé d_1 dans toute la suite est constitué de séries temporelles de mesure de la production PV de 9 centrales situées dans le sud de la France (sauf une en région parisienne). Les centrales sont représentées sur la figure 2.1. Les puissances crête des centrales varient de 45 kWc à 5 MWc. La distance entre les centrales varie de 5 km à 783 km. Le tableau 2.1 présente les distances entre les différentes centrales labellisées $P_i, i = 1 \dots 9$. Les données couvrent une période de 20 mois à partir de juillet 2013 avec une résolution allant de 6 min à 15 min en fonction du site. Un contrôle de la qualité des données a été effectué pour enlever les valeurs aberrantes et procéder à une imputation des données manquantes. La production a été supposée nulle entre 22h et 5h du matin. Les données ont été ensuite interpolées à un pas de temps de 15 min pour la suite des travaux.

FIGURE 2.1 – Les centrales du jeu de données d_1. Les centrales sont situées dans le sud-est de la France (sauf P_3).

Tableau 2.1 – Distances (en km) entre centrales du jeu de données d_1

Distance (km)	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
P_1	0							
P_2	186	0						
P_3	680	783	0					
P_4	280	465	638	0				
P_5	10	190	670	280	0			
P_6	55	235	670	230	58	0		
P_7	6	185	690	280	15	55	0	
P_8	183	20	800	460	188	229	179	0
P_9	185	17	801	460	188	230	180	3

Les données "Hespul"

Le second jeu de données est un bon exemple d'un cas d'étude comportant un nombre important de centrales avec une répartition géographique très dense. Ce cas d'étude comporte les données de 905 onduleurs installés dans la région centre-ouest de la France. Les puissances crêtes varient de 3.2 kWc à 58 kWc. Ces onduleurs correspondent à 185 centrales photovoltaïques différentes. Les centrales sont présentées sur la figure 2.2. La distance entre les centrales varie de 1 km à 230 km. Les données de production couvrent la période de novembre 2014 à mars 2016. Le même traitement des données que celui du premier cas d'étude a été effectué sur ce second cas d'étude et les données ont été interpolées à une résolution commune de 15 min. Après traitement des données, 136 centrales ont été retenues.

2.2 Analyse des données de production

La production PV est caractérisée par une variabilité importante. Les principales sources de cette variabilité sont la saisonnalité et les conditions climatiques. Si les variations saisonnières sont assez bien prévisibles car liées à la course du soleil, celles liées au climat notamment à la couverture nuageuse, le sont beaucoup moins.

Impact de la saisonnalité

On peut caractériser l'impact de la saisonnalité sur la production PV par :

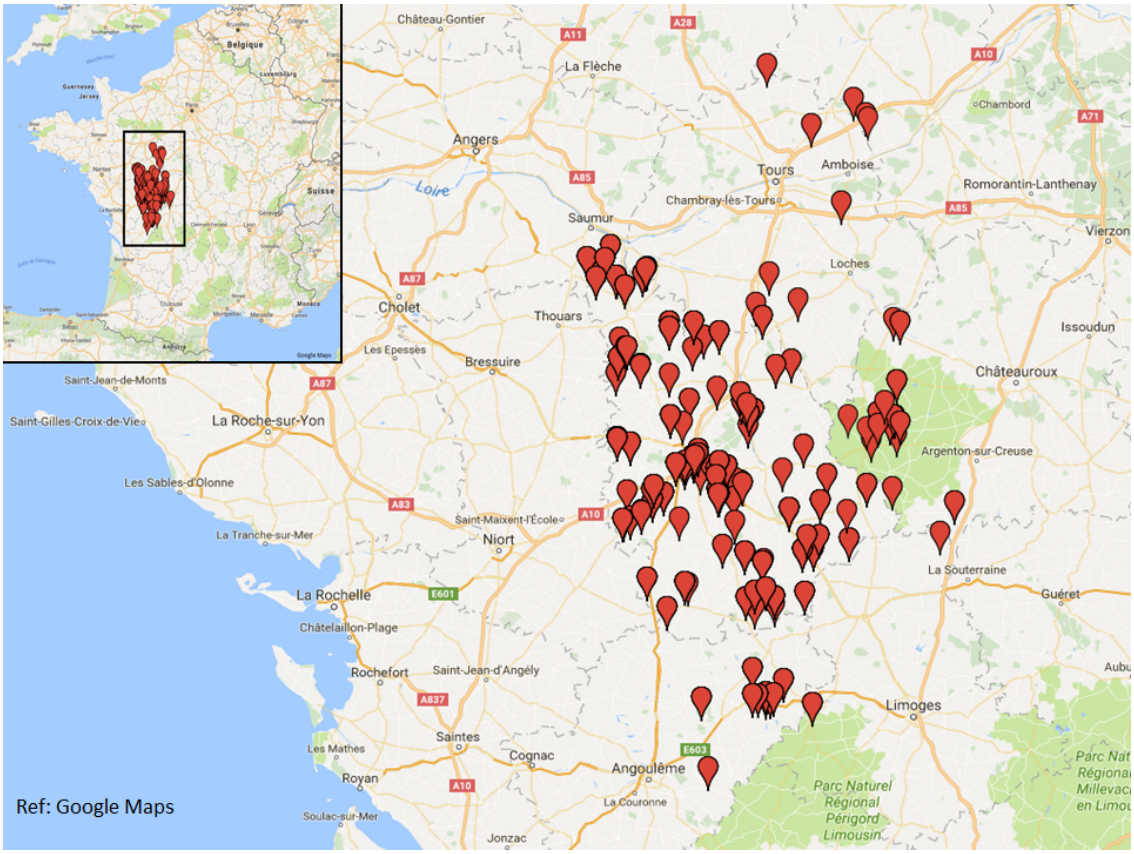


FIGURE 2.2 – Les centrales du jeu de données d_2. La distance entre les centrales varie de 1 km à 230 km. Les centrales sont situées dans le centre-ouest de la France.

— une production nulle la nuit ;

— un cycle journalier avec une pointe de production aux environs de midi (voir figure 2.4) ;

— un cycle annuel avec des niveaux de production élevés l'été et qui baissent significativement l'hiver (voir figure 2.3).

Les différentes variations que l'on peut observer sur la production journalière d'une centrale (voir figure 2.4) sont essentiellement dues aux variations météorologiques.

Variabilité de la production

La variabilité de la production d'une installation PV ou d'un réseau de plusieurs installations PV peut être analysée par différents critères. Dans [100], on retrouve un outil d'analyse de cette variabilité appelé "variabilité relative de la production". On définit la variabilité d'un ensemble de N installations PV par :

$$\sigma_{\Delta t}^{\Sigma N} = \left(\frac{1}{C^{Fleet}}\right)\sqrt{Var\left[\sum_{n=1}^{N}\Delta P_{\Delta t}^{n}\right]} \tag{2.1}$$

où C^{Fleet} est la capacité totale installée de l'ensemble des N installations PV. $\Delta P_{\Delta t}^{n}$ est la série temporelle des évolutions de la production pour la centrale n :

$$\Delta P_{\Delta t}^{n} = \left\{(t_1, \Delta P_{t_1,\Delta t}^{n}), (t_2, \Delta P_{t_2,\Delta t}^{n}), \ldots, (t_T, \Delta P_{t_T,\Delta t}^{n})\right\} \tag{2.2}$$

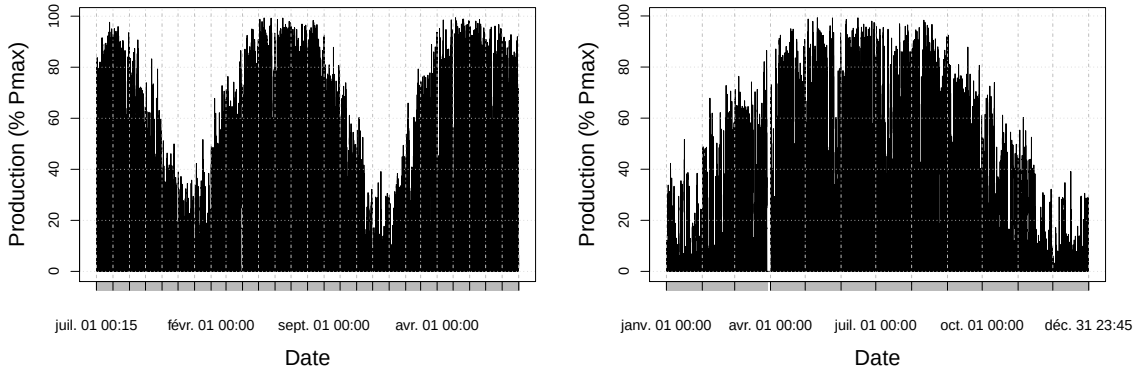


FIGURE 2.3 – Production photovoltaïque d'une centrale PV de juillet 2013 à août 2015 (gauche) et sur l'année 2014 (droite)

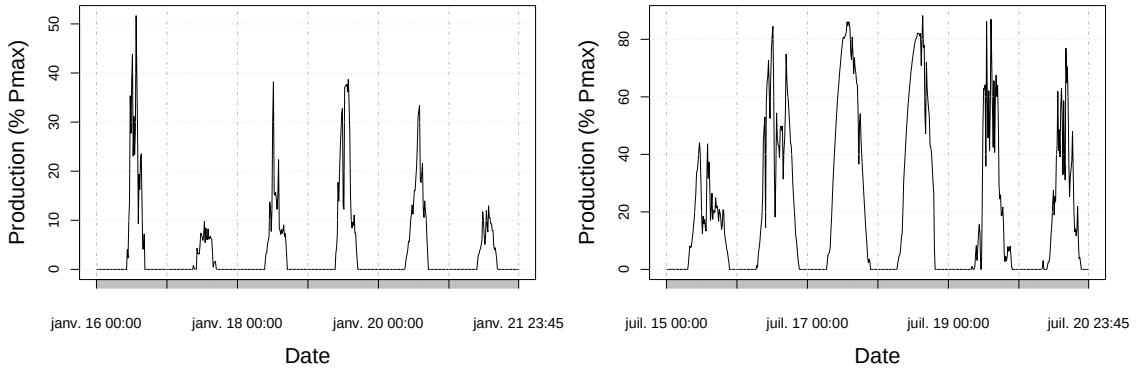


FIGURE 2.4 – Production photovoltaïque pour des semaines d'hiver (gauche) et d'été (droite)

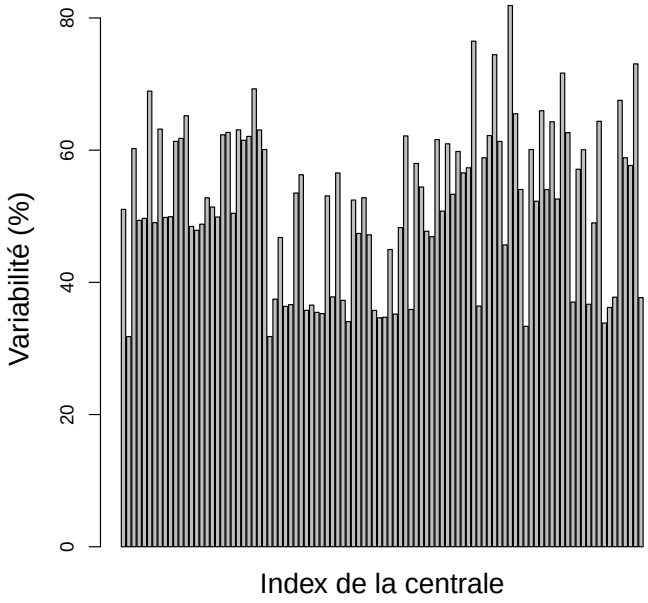
et $\Delta P^n_{t,\Delta t} = P^n_t - P^n_{t+\Delta t}$. La variabilité relative ou ROV (Relative Output Variability) d'une centrale i d'un ensemble de N centrales est définie comme :

$$ROV_i = \frac{\sigma^{\Sigma N}_{\Delta t}}{\sigma^{\Sigma i}_{\Delta t}}. \tag{2.3}$$

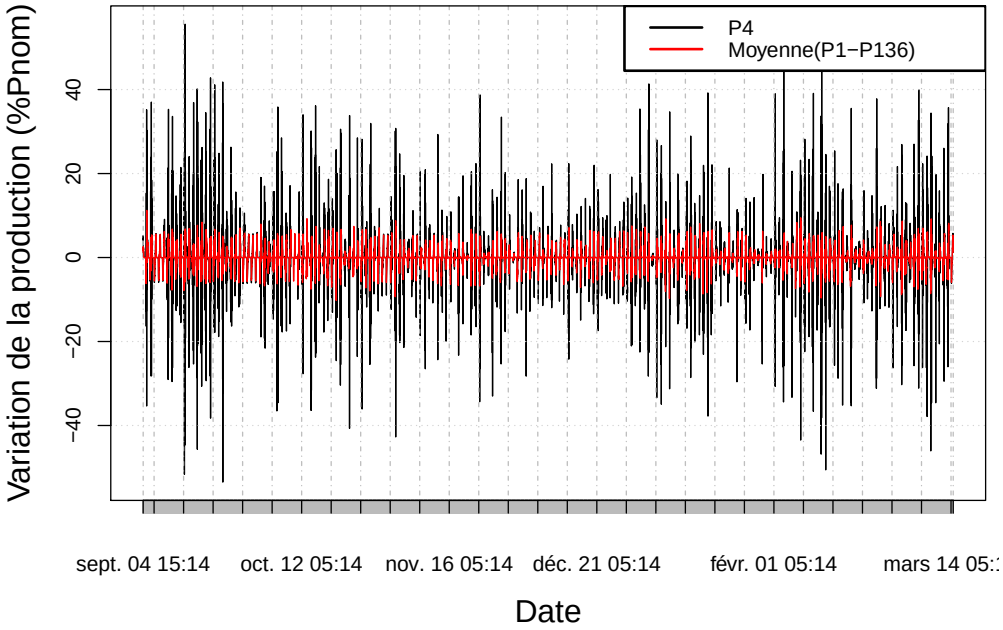
La figure 2.5a représente pour 100 centrales du jeu de données d_2 la variabilité relative en fonction de la centrale. Les valeurs de variabilité observées varient entre 30% et 80% traduisant ainsi une forte variabilité de la production PV. La variabilité peut être aussi analysée visuellement en s'inspirant de la série des différences proposée dans [100]. On définit donc une nouvelle série des différences P^d_t qui permet de visualiser les variations de la production PV au pas de temps k, $P^d_t = P_t - P_{t+k}$. La figure 2.5b représente la série des différences ainsi définie pour la centrale P_1 et l'ensemble (moyenné) des centrales du jeu de données d_2.

2.3 Méthode de stationnarisation proposée

Le développement des moyens de production d'énergie renouvelable at entrainé de nouveaux besoins en termes de prévision. Les méthodes de prévision basées sur les modèles



(a) Variabilité relative de 100 centrales du jeu de données d_2 sur la base d'un pas de temps de 15 min



(b) Variation (série P_t^d) de la production PV pour les centrales du jeu de données d_2. Le pas de temps est 15 min.

FIGURE 2.5 – Analyse de la variabilité de la production PV

numériques ne permettent pas de répondre efficacement à la maitrise de la variabilité de la production notamment à très court-terme (quelques minutes à quelques heures). Hormis la variabilité due à la course du soleil, les erreurs observées dans le cadre de la prévision de production des énergies renouvelables sont liées à la faible performance des modèles pour l'anticipation des perturbations météorologiques. Ces perturbations sont la plupart du temps des phénomènes qui se propagent spatialement. La prédictibilité de

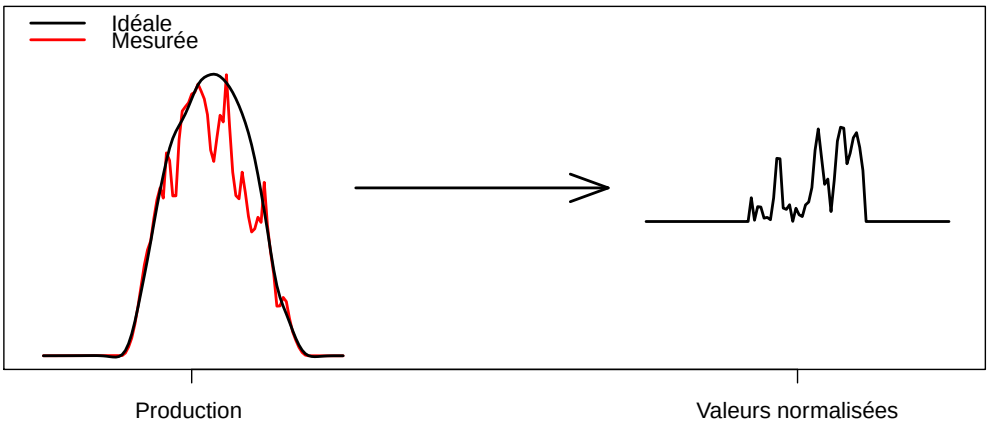


FIGURE 2.6 – Illustration du principe pratique de la stationnarisation

la production d'énergie renouvelable peut donc être améliorée par la prise en compte de ces perturbations spatiales en complément de l'analyse de la variabilité temporelle. C'est dans ce cadre que nous nous proposons de développer un modèle spatio-temporel qui exploite les informations spatiales et temporelles entre les mesures de production de différents sites de production PV. L'utilisation des mesures d'un ensemble de site de production PV pour mieux anticiper les phénomènes météorologiques nécessite au préalable de s'affranchir de la variabilité due à la course du soleil pour se concentrer sur l'effet des autres sources de variabilité de la production, notamment les nuages. L'extraction de la variabilité due à la course du soleil passe par la stationnarisation des séries de production. Cette stationnarisation permettra de répondre à la question de l'existence de corrélations spatio-temporelles non liées à la course du soleil entre les mesures de production et ainsi de l'opportunité de la mise en œuvre d'un modèle spatio-temporel. De plus, cette stationnarisation est utile pour la mise en oeuvre de certaines méthodes de prévisions classiques (non spatio-temporelles) qui font des hypothèses de stationnarité sur les données d'entrée.

En pratique, la stationnarisation consiste à "aplatir" la cloche (due à la course du soleil) dans les séries journalières de production pour ne garder que les autres variations qui sont liées aux perturbations météorologiques. La figure 2.6 montre l'objectif pratique visé par la sationnarisation. Sur la gauche de la figure, on voit deux courbes de production PV normalisées : celle idéale sans perturbation météorologique (courbe en noir) et celle réalisée (courbe en rouge). La production mesurée (réalisée) présente des variations dues aux perturbations météorologiques que l'on veut isoler par rapport à la courbe idéale. On veut donc obtenir la courbe de droite qui comporte uniquement les variations qui ne sont pas dues à la course du soleil.

2.3.1 Présentation de la méthode

Nous proposons ici une nouvelle méthode de stationnarisation des séries de production PV. Nous n'utilisons pas ici la différenciation des séries de production car elle nécessite une intégration a posteriori des erreurs. Cette intégration entraine une explosion des intervalles de confiance. Nous présentons dans la suite le cheminement qui a conduit à la méthode de stationnarisation proposée ainsi que les différentes analyses menées pour évaluer l'efficacité de ladite méthode.

Définition d'un indice de ciel clair pour la production PV
La méthode que nous proposons s'inspire de l'indice du ciel clair pour le rayonnement solaire [27, 22, 101]. Cet indice représente la façon dont l'atmosphère atténue la lumière d'une heure à l'autre ou au jour le jour en fonction du mouvement de la Terre autour du Soleil. Nous le définissons ici comme le rapport entre les mesures d'irradiation I_t^{meas} et celles estimées par modèle de ciel clair I_t^{sim} à l'instant t :

$$k_t^{irr} = \frac{I_t^{meas}}{I_t^{sim}}. \tag{2.4}$$

Dans le même esprit, nous définissons un indice de ciel clair pour la production photovoltaïque k_t^{pv} tel que :

$$k_t^{pv} = \frac{P_t^{meas}}{P_t^{sim}}, \tag{2.5}$$

avec P_t^{meas} la production PV mesurée à l'instant t, et P_t^{sim} la production sous hypothèses de ciel clair obtenue par transformation de l'irradiation fournie par un modèle ciel-clair en production pour le temps t. En pratique, P_t^{sim} est construit comme le produit du facteur d'efficacité η du système PV et de l'irradiation simulée I_t^{sim} sous des conditions ciel-clair :

$$P_t^{sim} = I_t^{sim} \times \eta. \tag{2.6}$$

Le paramètre η intègre le rendement du module, la surface active et les pertes.

L'indice ciel-clair pour la production k_t^{pv} ayant été défini, notre première idée pour l'appliquer a été d'utiliser les séries d'irradiation ToA (Top of Atmosphere) pour la stationnarisation. Cette irradiation est celle reçue par une surface horizontale à l'extérieur de l'atmosphère. Le choix de cette série d'irradiation est porté par le fait que l'irradiation ToA n'est affectée par aucun effet atmosphérique. Elle porte donc exclusivement la variabilité due à la course du soleil si on fait l'hypothèse de négliger les effets extra-terrestres. Les valeurs attendues sont en théorie comprises entre 0 et 1. La figure 2.7 montre la série résultant de la normalisation par le ToA pour l'année 2014 pour une centrale dans le sud de la France. En pratique, on peut observer une part non négligeable de valeurs plus grandes que 1, surtout pour les périodes avec les niveaux de production les plus faibles. Le traitement des périodes à faible niveau d'irradiation n'est pas satisfaisant. L'examen des autocorrélations de la série présentées dans la figure 2.8 montre une décroissance rapide qui confirme la non-stationnarité des séries.

L'utilisation des séries d'irradiation ToA ne permettant pas d'obtenir des séries stationnaires, il a été envisagé d'utiliser les séries d'irradiation ciel-clair. Le modèle ciel-clair retenu est le modèle ESRA [22]. Ce modèle, plus complet, permet de prendre en compte les paramètres de l'atmosphère à savoir l'effet des aérosols et de l'absorption des gaz. La figure 2.9 illustre les différences entre les irradiations ToA et ESRA.

Les limites de la normalisation par le modèle ciel-clair
Comme présenté dans l'état de l'art sur les modèles de ciel clair (partie 1.3), les méthodes de stationnarisation des séries de production PV (pas celle d'irradiation) par modèle de ciel-clair proposées dans la littérature sont très peu détaillées, ce qui les rend difficiles à reproduire. De plus, dans la littérature on ne retrouve pas d'analyse formelle sur l'efficacité des méthodes de stationnarisation, surtout avec des critères de vérification de stationnarisation. La normalisation par l'irradiation ciel-clair présente des limitations. La figure 2.10 présente pour une centrale située dans le sud-est de la France, les séries de production

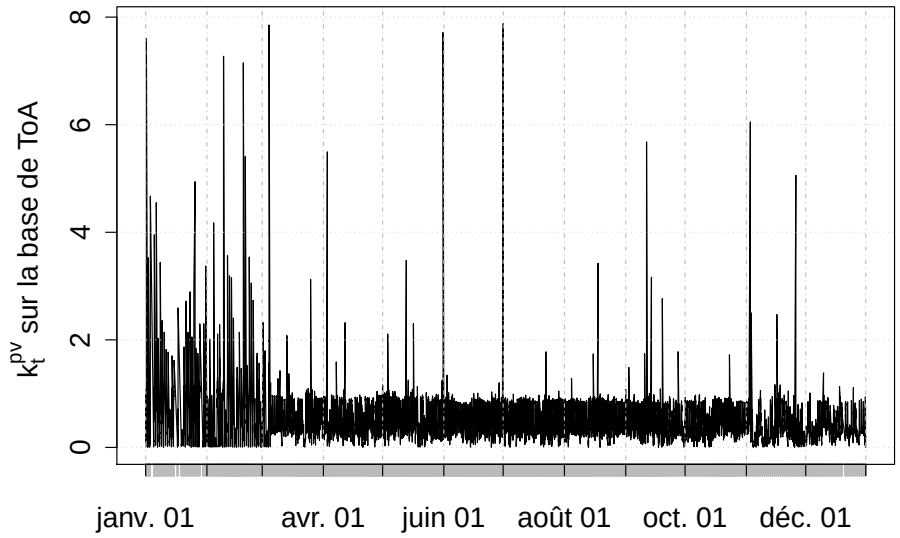


FIGURE 2.7 – Série normalisée sur la base de l'irradiation ToA d'une centrale située dans le sud-est de la France pour l'année 2014.

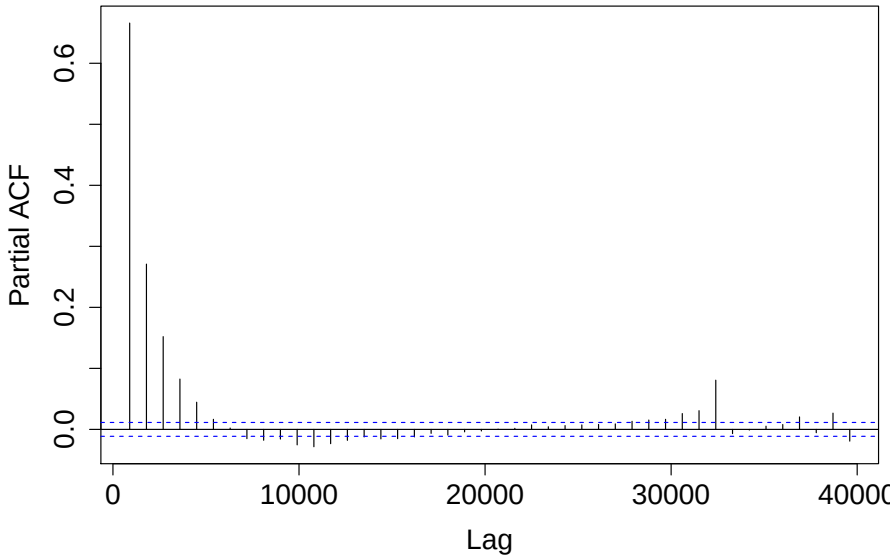


FIGURE 2.8 – PACF de la série normalisée sur la base de l'irradiation ToA d'une centrale située dans le sud-est de la France pour l'année 2014.

normalisées par la puissance maximale et celles normalisées par une série d'irradiation ciel-clair obtenue par le modèle ESRA. La simulation de l'irradiation a été effectuée sous l'hypothèse d'une surface horizontale en admettant que la variation de la production en sortie due au niveau d'inclinaison est assimilée par le facteur d'efficacité η. La principale limitation de cette stationnarisation concerne les faibles valeurs de production observées en début et fin de journée qui ne sont pas traitées. Ce constat est le même pour la méthode basée uniquement sur l'apprentissage statistique proposée par Bacher [26] qui exclut tout simplement les faibles valeurs d'irradiation.

Normalisation par une fonction de l'irradiation ciel-clair

La figure 2.11 présente la relation entre les valeurs normalisées de la production et l'irradia-

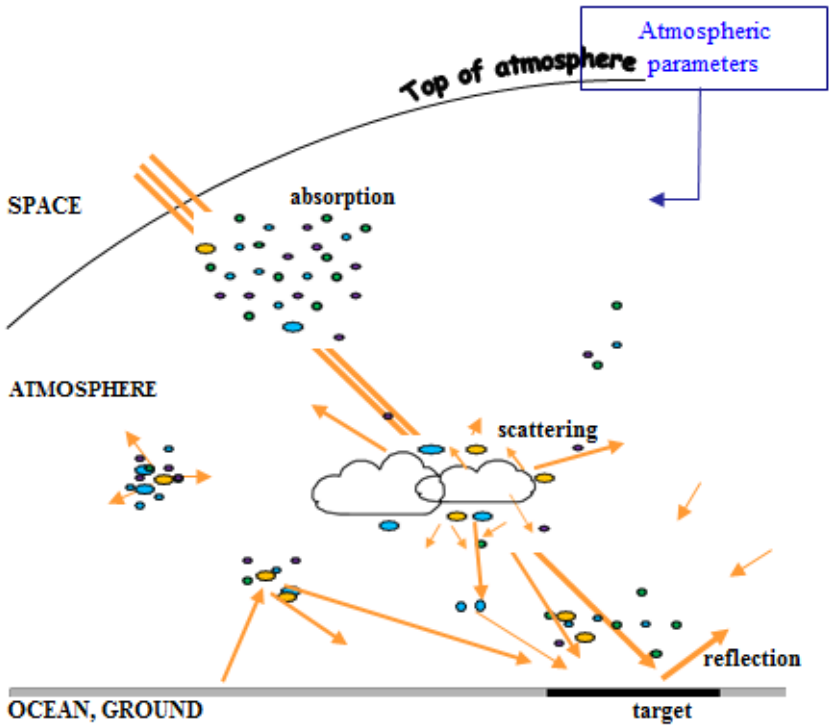


FIGURE 2.9 – Différences entre les irradiations ToA et le modèle ciel-clair ESRA. Le modèle ciel-clair ESRA prend en compte les différentes interactions qui ont lieu dans l'atmosphère.

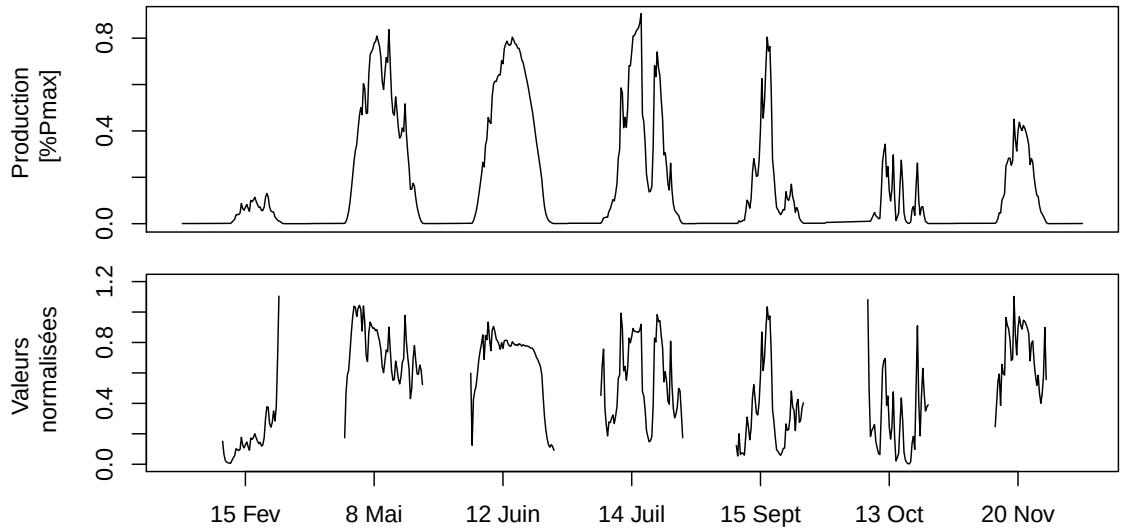


FIGURE 2.10 – Exemple de séries de production normalisées par la puissance maximale et de séries normalisées sur la base de l'irradiation ciel-clair ESRA pour différents jours de l'année 2014. La centrale est située dans le sud-est de la France.

tion ESRA pour deux périodes temporelles définies comme avant et après le midi solaire. Les deux classes de valeurs correspondent aux phases de croissance de la production (du

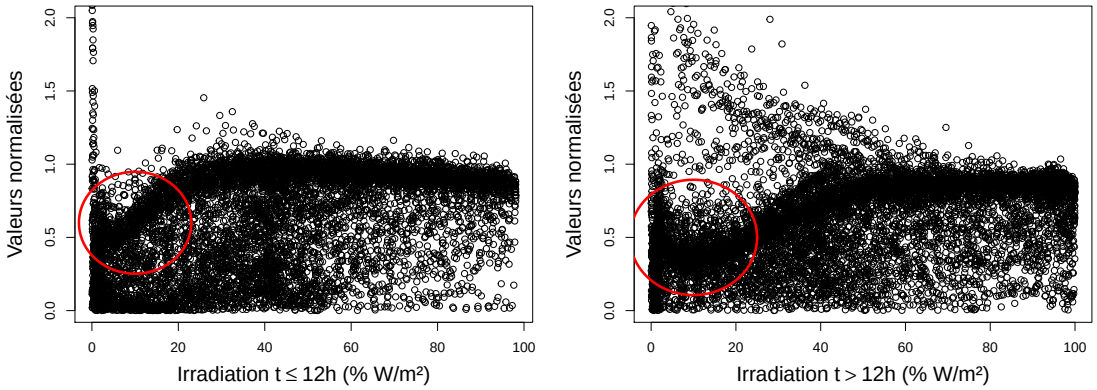


FIGURE 2.11 – Relation entre valeurs normalisées de production et irradiation ESRA selon le moment de la journée pour l'année 2014. La centrale est située dans le sud-est de la France.

lever du soleil au midi solaire) et de décroissance, c'est-à-dire du midi solaire au coucher du soleil. L'analyse de la figure montre (cercle rouge) l'inefficacité de la normalisation par l'irradiation ESRA pour les faibles valeurs de production.

Pour améliorer la qualité des séries normalisées et résoudre les problèmes spécifiques aux périodes de faible production comme les débuts et fins de journée, nous proposons une nouvelle relation entre la production réelle et P_t^{sim} en utilisant une fonction de normalisation f. Cette fonction sert à modéliser le lien entre les deux productions (réelle et simulée) dans le but de produire une nouvelle série u_t plus stationnaire définie pour les heures auxquelles P_t^{sim} est non nulle par :

$$u_t = P_t / f(P_t^{sim}). \tag{2.7}$$

Plusieurs formes ont été envisagées pour la fonction de normalisation f à savoir linéaire, linéaire par morceaux ou quadratique. Le choix de la fonction de normalisation appropriée a été effectué en utilisant un critère quantitatif basé sur l'évolution de l'écart-type journalier de la série u_t. En pratique, pour chaque type de fonction de normalisation f précédemment énuméré, nous construisons la série des écarts-types journaliers de u_t. La fonction f retenue est celle pour laquelle la série u_t présente une moyenne indépendante du temps et des écart-types journaliers de plus faible variabilité intra-journalière.

La fonction de normalisation que nous avons retenue pour corriger les défauts de la simple normalisation par l'irradiation est linéaire par morceaux et dépend du sens de l'évolution quotidienne de la production (soit une augmentation au début de la journée, soit une diminution après le midi solaire). L'idée est de corriger les faibles valeurs d'irradiation dans la normalisation par l'ajout d'une pénalisation. Cette pénalisation permet aussi de corriger le nombre non négligeable de valeurs de la série normalisée qui excèdent la valeur théorique de la borne maximale qui est 1. La fonction retenue s'écrit :

$$f(P_t^{sim}) = P_t^{sim} + f_a(P_t^{sim}) + f_b(P_t^{sim}) \tag{2.8}$$

avec f_a définie du lever du soleil au midi solaire et f_b du midi solaire au coucher du soleil. Les fonctions f_a et f_b participent à améliorer la stationnarisation particulièrement pour les faibles niveaux de production en début et fin de journée. Elles sont définies pour chaque jour par :

$$\left\{\begin{array}{rcl} f_a(0) & = & \alpha_a \\ f_a\left(\beta_a \frac{P_{max}^{sim}}{2}\right) & = & 0 \\ f_a(P_{max}^{sim}) & = & \gamma \end{array}\right. \quad \left\{\begin{array}{rcl} f_b\left(P_{max}^{sim}\right) & = & \gamma \\ f_b\left(\beta_b \frac{P_{max}^{sim}}{2}\right) & = & 0 \\ f_b(0) & = & \alpha_b \end{array}\right. \tag{2.9}$$

où P_{max}^{sim} représente la production maximale simulée pour la journée. La continuité en le midi solaire (ou maximum journalier d'irradiation sous hypothèses ciel-clair) est assurée par le coefficient γ. Les valeurs des coefficients $\alpha_a, \alpha_b, \beta_a, \beta_b, \gamma$ sont obtenues grâce à un processus d'optimisation qui vise à minimiser le critère d'écart-type. L'optimisation est faite sous les contraintes $\beta_{a,b} \in (0, 2)$. Les coefficients sont initialisés de manière aléatoire, une fenêtre d'un mois de valeurs de la série d'irradiation ESRA est choisie avant le jour d'intérêt. Les coefficients sont choisis sur la fenêtre de valeurs d'irradiation en optimisant l'écart-type de la série normalisée.

En pratique, on évalue pour plusieurs valeurs de coefficients, l'écart-type de la série normalisée et on retient les coefficients qui minimisent l'écart-type. Le critère d'optimisation des coefficients est donc directement lié aux propriétés de stationnarité de la série d'intérêt. La stationnarité de la forme normalisée de u_t a été évaluée en analysant son auto-corrélogramme et aussi par des tests de racine unitaire. La figure 2.12 présente les résultats de l'optimisation des coefficients à utiliser pour le processus de stationnarisation du jeu de données d_1. La figure montre que les coefficients les plus variables sont les α_a et α_b qui permettent de corriger les débuts et fins de journée.

La procédure de stationnarisation peut être résumée pour une centrale PV par les étapes ci-après :

1. Nettoyage des données aberrantes de la série de production PV
2. Simulation de la série d'irradiation ciel clair ESRA et de la série de puissance correspondante (équation (2.6))
3. Détermination des coefficients appropriés des fonctions (f_a, f_b) en utilisant un processus d'optimisation sur un intervalle glissant des valeurs d'irradiation simulées
4. Normalisation la série mesurée P_t^{meas} pour obtenir la série u_t.

Le graphique 2.13 présente pour les centrales du jeu données d_2 les corrélations entre couples de centrales en fonction de la distance avant et après stationnarisation. Il permet d'apprécier visuellement l'effet de la stationnarisation sur les corrélations entre couples de centrales ; on constate une baisse significative de ces corrélations et une répartition moins concentrée autour des fortes valeurs.

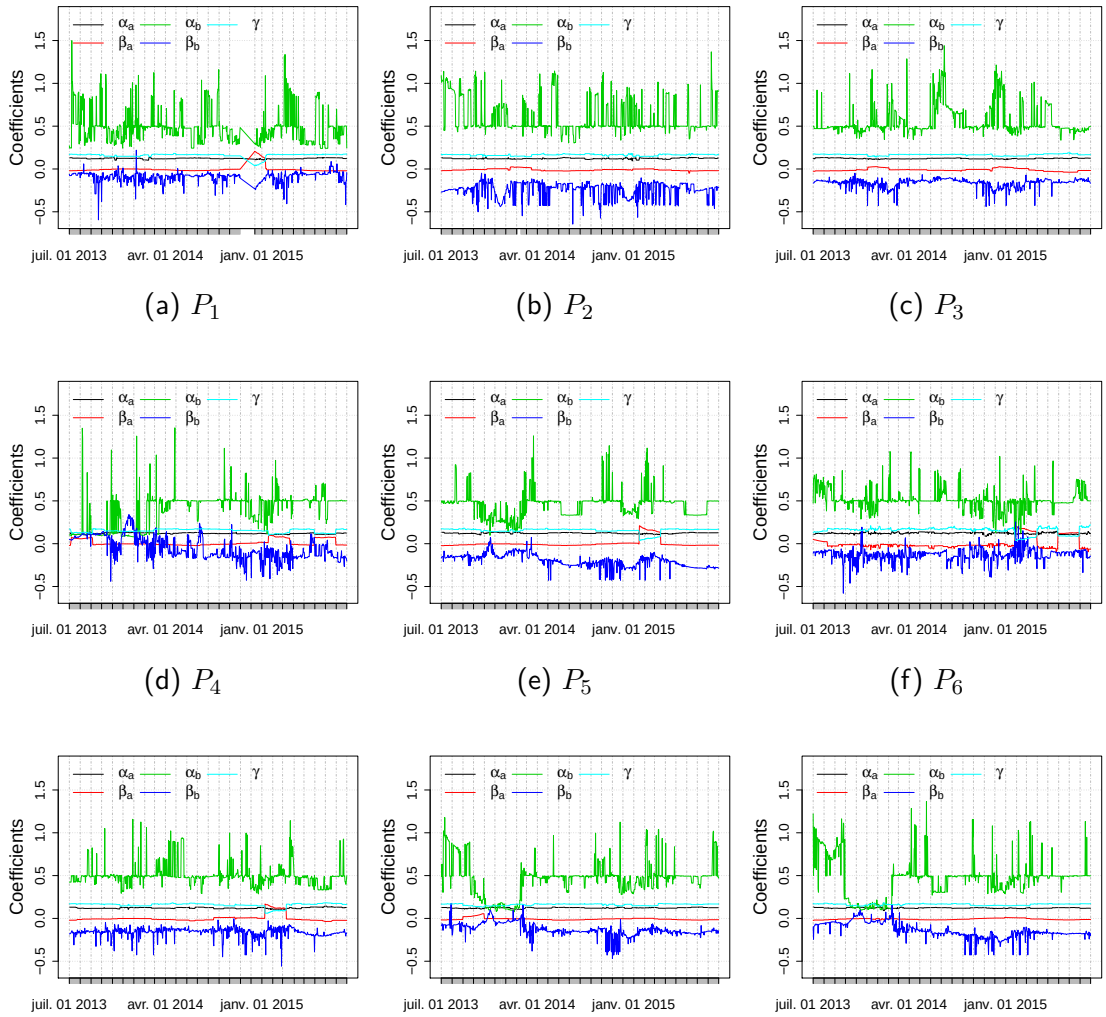
2.3.2 Étude des performances de la méthode de stationnarisation

Étude des autocorrélations

La méthode de stationnarisation proposée est comparée à la normalisation par la série d'irradiation ToA. Les figures 2.14a et 2.14b représentent les fonctions d'autocorrélations partielles (PACF) des séries stationnarisées selon les deux méthodes pour la centrale P_1 du jeu de données d_1. L'examen de ces autocorrélations montre que la méthode de stationnarisation proposée permet une meilleure correction (pour les lags élevés, la valeur de PACF est toujours contenue dans l'intervalle de confiance) des séries dans l'objectif d'appliquer un modèle autorégressif.

Tests de stationnarité

La notion de stationnarité faible, ou au second ordre se définie par l'invariance des moments d'ordre 1 et 2 au cours du temps. Il existe différentes sources de non-stationnarité.



(a) P_1 (b) P_2 (c) P_3

(d) P_4 (e) P_5 (f) P_6

(g) P_7 (h) P_8 (i) P_9

FIGURE 2.12 – Évolution journalière des coefficients des fonctions de stationnarisation pour chacune des centrales ($P_1 - P_9$) du jeu de données d_1.

La source de non stationnarité peut être déterministe (trend stationary) ou stochastique (difference stationary). En pratique, il est difficile de vérifier de manière rigoureuse l'hypothèse de stationnarité du second ordre. On utilise l'hypothèse de stationnarité intrinsèque qui est une version affaiblie de l'hypothèse de stationnarité du second ordre. Cette hypothèse intrinsèque stipule que les accroissements de la fonction aléatoire sont stationnaires au second ordre c-a-d l'espérance des accroissements est nulle et sa variance ne dépend que du vecteur séparant les deux points.

Pour évaluer l'efficacité de la méthode de stationnarisation, une première étude avec le test de Dickey-Fuller augmenté (ADF) [102, 103] a été réalisée. C'est un test de présence de racine unitaire ou test de stationnarité intrinsèque. En effet, si on considère un processus Y_t et $\phi(L)$ le polynôme de l'opérateur de retard L , si le processus $\phi(L)Y_t$ est stationnaire, si 1 est une racine du polynôme ϕ alors le processus Y_t sera non stationnaire. Cela explique pourquoi la plupart des tests de stationnarité sont des tests de détection de racine unitaire.

Les hypothèses du test de Dickey-Fuller augmenté pour un processus Y_t sont les suivantes :

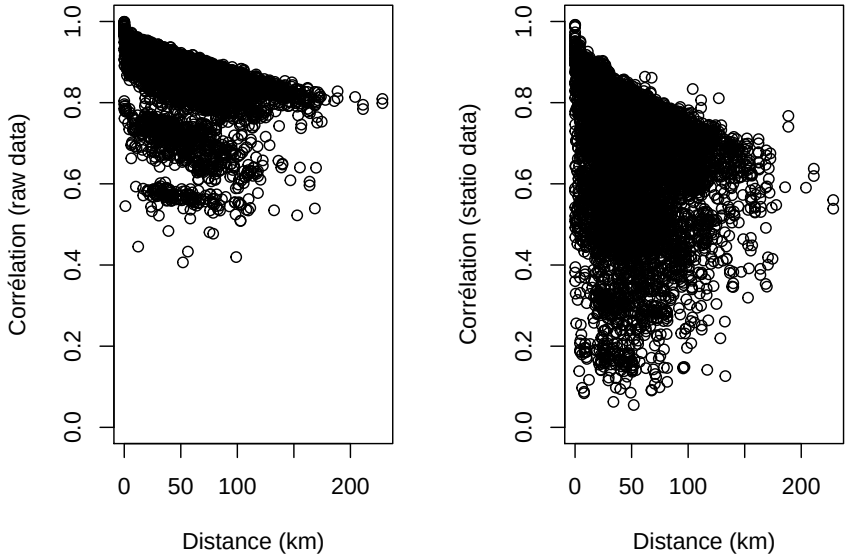
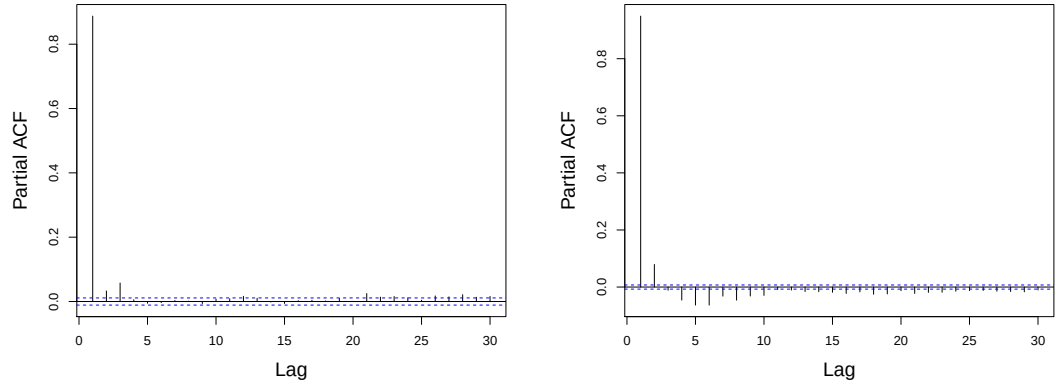


FIGURE 2.13 – Corrélation entre couples de centrales en fonction de la distance avant stationnarisation (à gauche) et après stationnarisation (à droite).



(a) PACF série normalisée par la méthode proposée

(b) PACF série normalisée par l'irradiation ToA

FIGURE 2.14 – Auto-corrélation partielle (PACF) des séries normalisées

(H0) : Il y a présence de racine unitaire donc le processus Y_t est non stationnaire.
(H1) : Il n'y a pas de racine unitaire donc le processus Y_t est stationnaire. Puisqu'on teste la stationnarité intrinsèque, on travaille sur la régression :

$$\Delta Y_t = d_t + \sum_{i=1}^{p} \gamma_i \Delta Y_{t-i} + \rho Y_{t-1} + u_t \tag{2.10}$$

avec d_t une fonction déterministe du temps et u_t les erreurs.

La statistique de Dickey-Fuller est alors définie comme :

$$\Delta D_\rho = \frac{D_\rho}{1 - \sum_{i=1}^{p} \gamma_i} \tag{2.11}$$

Tableau 2.2 – Valeur des statistiques de Dickey-Fuller pour les séries normalisées sur la base du ToA et suivant la méthode proposée pour les centrales $P_1 - P_4$

Centrales	Série normalisée par le ToA	Série normalisée par la méthode proposée
P_1	-2.00	-20.35
P_2	-2.35	-17.28
P_3	-1.14	-18.64
P_4	-2.78	-20.64

où D_ρ est la statistique de test.

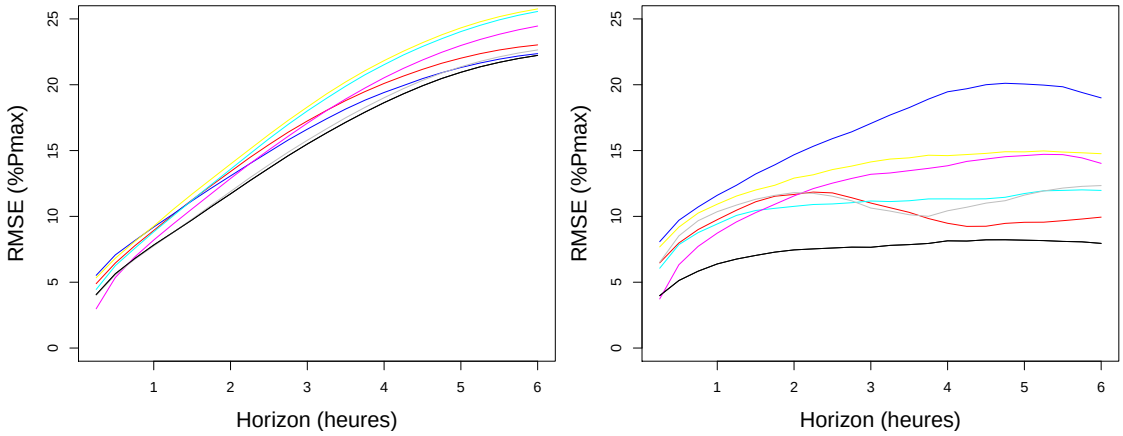
Dans le tableau 2.2, nous présentons les valeurs de statistique de Dickey-Fuller et de seuil pour les séries normalisées par l'irradiation ToA et celles stationnarisées par la méthode proposée précédemment pour quatre centrales du jeu de données d_2. On remarque que pour les séries normalisées suivant la méthode proposée, on rejette l'hypothèse de non stationnarité car les valeurs des statistique sont inférieures aux valeurs seuils à 10% et 5% qui sont respectivement -3.41 et -3.12. Le résultat est inverse pour la série normalisée par le ToA où l'hypothèse nulle de non stationnarité n'est pas rejetée. Le constat est le même pour la plupart des centrales du jeu de données. La méthode de stationnarisation proposée permet donc d'obtenir les conditions de stationnarité au sens du critère de Dickey-Fuller.

Performances des séries stationnarisées pour la prévision

L'objectif premier de la stationnarisation est d'améliorer les performances des méthodes de prévision. Au-delà des tests de stationnarité et de l'examen des autocorrélations, il convient donc d'évaluer l'efficacité de la méthode de stationnarisation proposée à travers l'examen des performances des séries stationnarisées.

Un modèle autorégressif (AR) a été utilisé pour prévoir la production PV pour les deux types de séries de chaque jeu de données à savoir les séries stationnarisées selon la procédure proposée et celles brutes (normalisées par la valeur maximale observée). Le critère du RMSE a été utilisé pour évaluer les performances. Pour l'ensemble des centrales du jeu de données d_1, l'amélioration moyenne du RMSE obtenue en utilisant les séries stationnarisées est de 7%. Les erreurs de prévision selon le critère RMSE sont représentées sur la figure 2.15 pour les séries stationnarisées ou non. Elles illustrent encore plus clairement l'apport de la stationnarisation dans le cadre de l'amélioration des performances de prévision. Les séries stationnarisées permettent donc d'obtenir de meilleures performances de prévision que la simple normalisation par les valeurs maximales. Une analyse plus précise de l'apport de la méthode de stationnarisation sur les performances en prévision pour les débuts et fins de journée a été effectuée. Les plages horaires de 6h à 9h et de 19h à 22h ont respectivement été utilisées pour calculer les critères d'évaluation. L'amélioration moyenne du MAE pour ces deux tranches horaires lorsqu'on utilise les séries stationnarisées plutôt que les données brutes est respectivement de 8% et 9%.

La même étude a été réalisée sur les centrales du jeu de données d_2. La figure 2.16 représente l'amélioration du RMSE obtenue en utilisant les séries stationnarisées. L'amélioration moyenne pour un horizon de 3 heures est de 10% et peut monter jusqu'à 15%. Cette réduction significative des erreurs de prévisions confirme l'efficacité de la méthode de stationnarisation et son intérêt à l'utiliser pour prétraiter les données avant utilisation dans les modèles de prévision.



(a) Série non stationnarisée (b) Série stationnarisée

FIGURE 2.15 – Erreurs de prévision RMSE d'un modèle AR avec des séries stationnarisées ou non pour le jeu de données d_1. Chaque ligne représente une centrale PV.

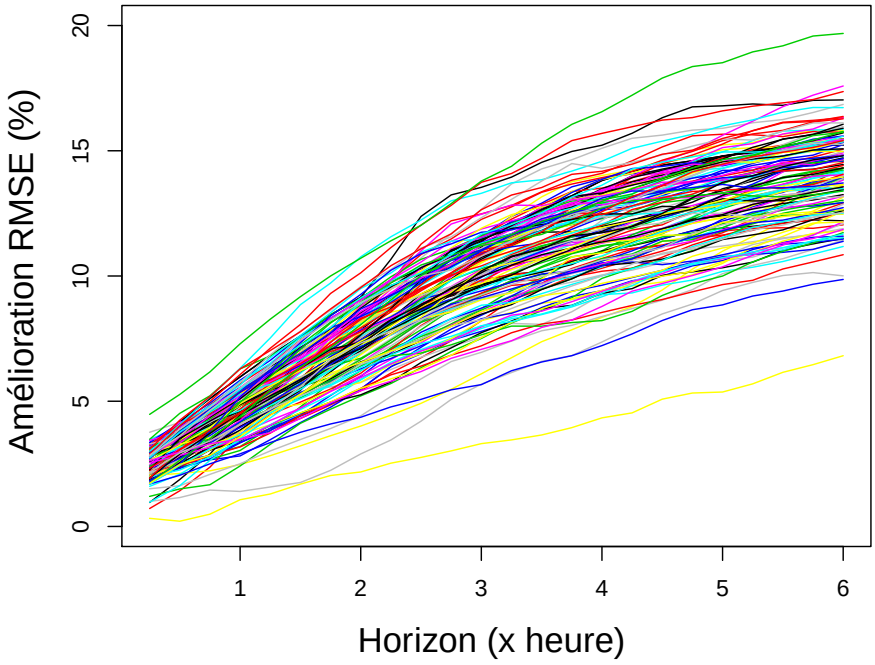


FIGURE 2.16 – Jeu de données d_2 : Amélioration du RMSE pour un modèle AR avec des séries stationnarisées par rapport à des séries non stationnarisées. Chaque ligne représente une centrale. Le pas de temps est de 15 min.

2.4 La corrélation spatio-temporelle

La prévision spatio-temporelle de la production PV au sens de l'utilisation des mesures de centrales voisines (comme un réseau de capteurs) repose sur l'existence de lien spatio-temporel entre ces centrales. Dans cette partie, nous justifions la création d'un modèle spatio-temporel par la mise en évidence de liens spatio-temporels entre des centrales spatialement distribuées. Pour cela, nous effectuons une analyse basée sur des indicateurs spécifiques aux liens spatiaux ainsi qu'une étude des corrélations des mesures de

production des différents sites.

2.4.1 Etude du lien spatial

La dépendance de la production PV au temps est une caractéristique du phénomène de production PV (voir figures 2.3 et 2.4). Il s'agit donc d'en analyser les occurrences spatiales. La production PV est donc considérée comme une variable régionalisée avec autant d'observations que de sites de production. Le lien spatial entre ces différentes observations est étudié avec différents critères incluant l'indice de Moran et les corrélogrammes.

Indice de Moran

La statistique la plus utilisée pour tester l'autocorrélation spatiale dans une série définie spatialement est celle de Moran [104]. Cette statistique s'écrit formellement de la façon suivante pour une variable régionalisée de production Y :

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_i (Y_i - \bar{Y})^2} \tag{2.12}$$

avec Y la production PV sur un réseau discret de N sites ; W une matrice carrée de poids positifs de dimension N tel que $w_{i,j}$ quantifie les influences du site j sur le site i. La matrice de poids que nous avons retenue ici est une matrice de distance. On peut retrouver dans la littérature des matrices de contiguïté mais elles ne sont pas adaptées au cas de la production PV qui est faite sur des sites et non des régions entières, ce qui rend difficile la définition de voisinage.

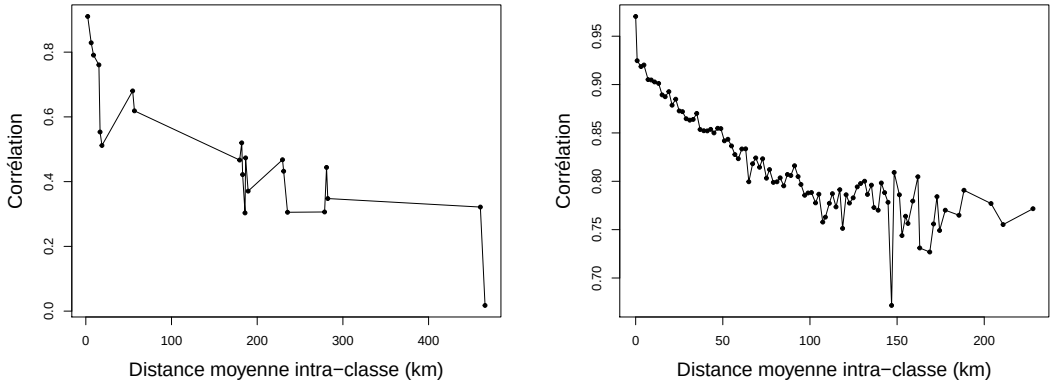
La statistique de Moran centrée et réduite suit asymptotiquement une loi normale d'espérance nulle et de variance unitaire et sert ainsi de base au test de l'autocorrélation spatiale dans une série. Lorsque I est proche de 1 en valeur absolue on a autocorrélation (positive si I l'est, négative sinon) ; une valeur proche de 0 en valeur absolue traduit l'indépendance des observations.

Le calcul de cette statistique sur les jeux de données d_1 et d_2 à un pas de temps horaire fournit des valeurs moyennes respectives de l'indice de Moran de 0.37 et 0.66. La valeur est plus faible pour le jeu de données d_1 car les centrales y sont plus distantes. Ces valeurs significatives traduisent la présence de corrélations spatiales entre les différentes centrales. La corrélation étant plus forte dans le cas du jeu de données plus dense.

Corrélogramme "modifié"

Les auto-corrélogrammes sont utilisés pour étudier l'existence de corrélation spatiale entre les mesures de production. Nous utilisons ici une version modifiée des méthodes classiques de calcul d'auto-corrélogramme [105]. Le principe de cette méthode peut être résumé par les étapes suivantes :

1. Répartir les corrélations entre les paires de centrales en classes en fonction des distances séparant les sites
2. Dans chaque classe, tester la significativité des corrélations par des tirages aléatoires de valeur de corrélation croisée de sorte à ce qu'une centrale soit utilisée une seule fois. Par exemple si la corrélation entre A et B est choisie, toutes les autres paires de combinaison intégrant A et B sont exclues de la base de tirage.
3. Répéter la procédure jusqu'à ce qu'il n'y ait aucune centrale inutilisée
4. Déterminer le nombre de coefficients de corrélation négatifs et positifs



(a) Auto-corrélogramme des centrales du jeu de données d_1 excepté la centrale P_3

(b) Auto-corrélogramme des centrales du jeu de données d_2

FIGURE 2.17 – Corrélogrammes spatiaux modifiés de la production PV pour les deux jeux de données

5. Après un nombre suffisant de tirage, la significativité peut être assurée si on a plus de valeurs positives que négatives.

Nous avons calculé pour les centrales des jeux de données d_1 et d_2 les corrélations intra-classes avec 1000 ré-échantillonnages. La figure 2.17 présente les valeurs de corrélations spatiales inter-classes obtenues pour les deux jeux de données. Les valeurs de corrélation obtenues décroissent avec la distance. De plus, ces valeurs sont supérieures à 0.65 dans le cas de d_2 où les centrales sont plus rapprochées et valent en moyenne 0.5 pour le jeu de données d_1. Ce résultat traduit la présence significative d'un lien spatial entre les productions des centrales. Dans toute la suite, nous utilisons les séries stationnarisées.

2.4.2 Calcul des corrélations spatio-temporelles

La présence de corrélation spatiale entre les mesures de production a été démontrée par l'analyse spatiale du phénomène de production. Pour compléter cette analyse, nous étudions les corrélations temporelles entre les séries de production des différentes installations. La figure 2.18 représente les fonctions de répartition empiriques des valeurs de corrélations croisées (cross-correlation) entre les séries temporellement retardées de production du jeu de données d_1. Pour un couple de centrales PV s_1, s_2 par exemple, la valeur de corrélation est obtenue en calculant la corrélation entre s_1 et les différentes séries affectées d'un retard $lag(s_2, k)$. Le retard maximal utilisé est l'horizon limite de prévision soit 6 heures dans ce cas. La même opération est répétée en intervertissant s_1 et s_2. La valeur retenue est la valeur maximale obtenue. Les fonctions de répartition sont représentées pour trois classes de distances de 0 à plus de 100 km avec des tailles de classes de 50 km. La première classe (moins de 50 km) présente les valeurs de corrélation les plus élevées. Sur l'ensemble du graphique, on observe que les valeurs de corrélation croisées sont plutôt élevées, concentrées dans l'intervalle $[0.4 - 0.8]$ et décroissantes avec la distance. L'effet de la course du soleil n'est plus présent dans les séries stationnarisées. Les valeurs de corrélations croisées observées traduisent donc la dépendance spatio-temporelle entre les séries de production. Ce transfert d'information essentiellement dû aux mouvements des nuages permet d'anticiper les perturbations météorologiques et d'améliorer la qualité des prévisions uniquement en utilisant les données de production.

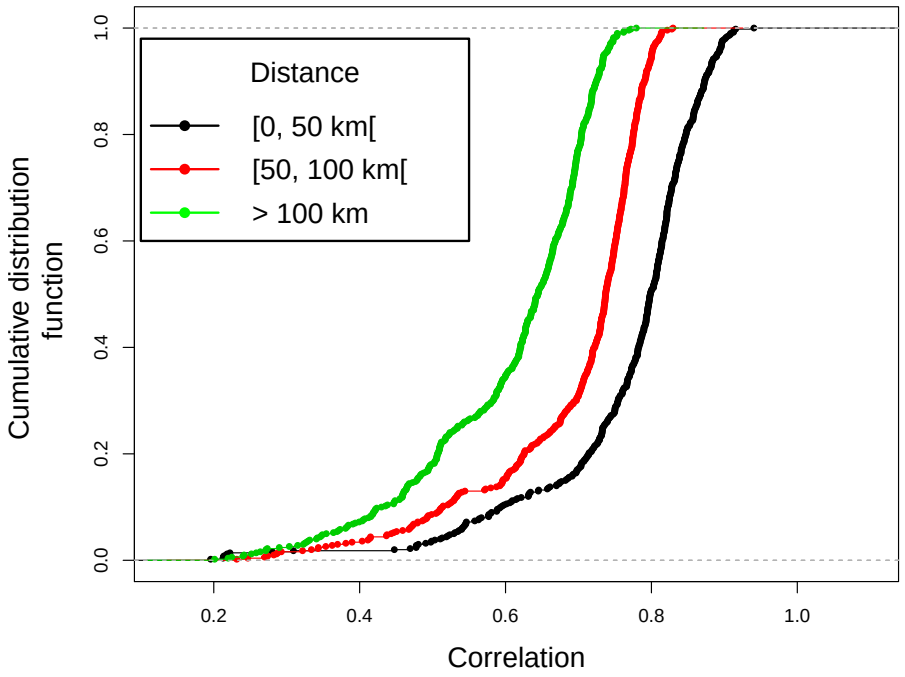


FIGURE 2.18 – Jeu de données d_2 : Fonction de répartition empirique des corrélations croisées entre les séries retardées de production. Les courbes vertes, rouges et noires correspondent respectivement aux trois classes de distance entre les centrales.

La Figure 2.19 qui présente le variogramme spatio-temporel pour les séries de production stationnarisées du jeu de données d_2 permet de compléter cette analyse. En effet, la portée est assez faible (de l'ordre de 10 km) et on n'observe pas d'effet pépite. De plus la forte montée des variogrammes traduit une forte variabilité spatiale entre les mesures de productions à courte distance.

Horizons et résolutions spatiales limites de la propagation des corrélations spatio-temporelles

La corrélation spatio-temporelle étant établie, nous allons à présent évaluer les limites temporelles et spatiales de la propagation de ces corrélations. Pour les corrélations temporelles, nous calculons le délai temporel t_{lim} à partir duquel la corrélation entre les séries de production retardées de chaque couple de centrales $(i, j), 1 \leq i, j \leq N$ est en valeur absolue inférieure ou égale à un seuil limite. Le tableau 2.3 présente les valeurs de t_{lim} pour un seuil de corrélation limite de 0.2 pour les centrales du jeu de données d_1. La valeur maximale est proche de 7 heures. La centrale la plus éloignée (P_3) présente les plus faibles valeurs de délai temporel car la propagation de corrélation est quasiment inexistante au vu de la distance. La figure 2.20 représente les valeurs de ce temps limite pour les centrales du jeu de données d_2. Le retard temporel limite dépasse les 7 heures pour ce cas d'étude. On peut conclure que la corrélation spatio-temporelle existant entre les centrales voisines est significative (> 0.20) pour des retards pouvant aller jusqu'à 7 heures. Au-delà de ces valeurs de délais temporels, il n'y a plus d'informations à exploiter de ces dépendances pour améliorer les prévisions. Ce résultat conforte les travaux de Bacher [26] sur les horizons limites d'utilisation des données historiques et justifie par la même occasion l'intérêt de la modélisation spatio-temporelle pour les horizons allant jusqu'à 6 heures.

Du point de vue spatial, le croisement des valeurs de corrélations entre les séries et les

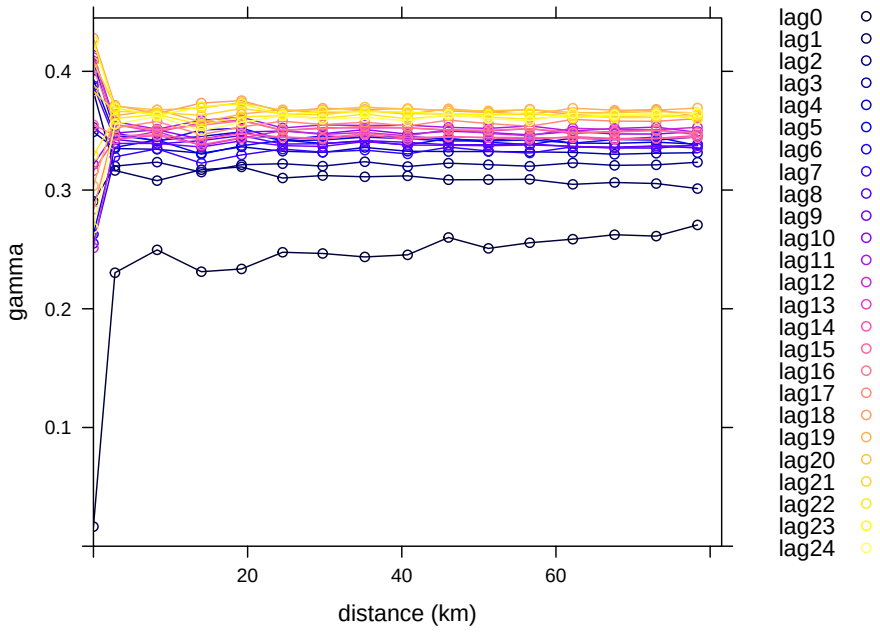


FIGURE 2.19 – Variogramme spatio-temporel des séries de production post stationnarisation

Tableau 2.3 – Jeu de données d_1 : Valeurs en heures de délai temporel à partir duquel la corrélation inter production est négligeable (seuil=0.2)

t_{lim} en heures	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
P_2	1.75							
P_3	0.75	0.25						
P_4	3.00	0.25	0.50					
P_5	5.25	4.75	0.50	0.75				
P_6	5.00	1.25	0.75	3.00	5.25			
P_7	5.25	4.00	0.50	1.75	6.75	5.25		
P_8	4.75	3.25	0.25	1.25	4.25	4.50	4.25	
P_9	4.75	4.75	0.25	1.25	5.25	4.75	5.25	4.75

distances entres sites montre qu'au-delà de 250 km les valeurs de corrélations sont très faibles. Cette distance constitue la limite au-delà de laquelle il n'y plus d'informations spatio-temporelles à exploiter.

2.5 Conclusion

La forte variabilité qui caractérise la production PV et ses différentes sources ont été présentées. La première source de variabilité est la saisonnalité. Nous avons montré dans ce chapitre la relation entre la saisonnalité et la production PV. Cette saisonnalité se traduit par une production négligeable la nuit, des cycles journaliers liés à la position du soleil et des cycles annuels dus à l'alternance des saisons. La variabilité de la production qui est due à la course du soleil a fait l'objet de nombreuses études et peut être presque complètement modélisée. Dans le cadre de cette thèse, nous nous intéressons à la modéli-

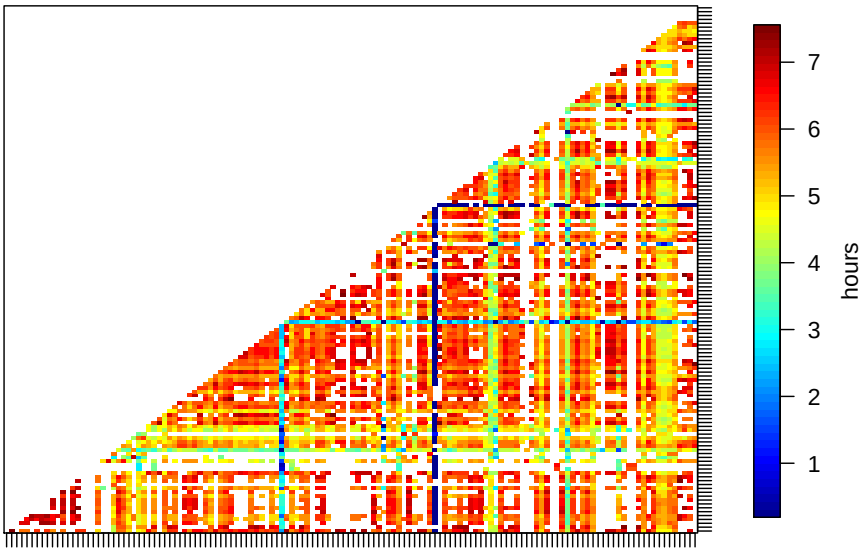


FIGURE 2.20 – Jeu de données d_2 : Valeurs en heures de délai temporel à partir duquel la corrélation inter production est négligeable (seuil=0.2). Chaque point d'axe représente une centrale PV.

sation spatio-temporelle de la production PV. Pour mettre en évidence cette corrélation et justifier l'intérêt de notre modélisation, il était nécessaire de s'affranchir des effets de la course du soleil sur les niveaux de production. En effet, un travail direct sur des séries non traitées aurait simplement été une mise en évidence de transferts de corrélation qui ne sont que la conséquence de la course du soleil qui serait à une position donnée sur certaines centrales avant d'autres. La méthode de stationnarisation que nous avons proposée dans ce chapitre nous permet d'exclure des séries de production l'effet de la course du soleil afin de nous consacrer aux autres sources de variabilité. Cette méthode de stationnarisation a montré de bonnes propriétés pour les tests de stationnarité mais a surtout donné de meilleures performances pour la prévision en comparaison avec les séries brutes. Les séries stationnarisées ont ensuite été utilisées pour mettre en évidence la corrélation spatiale et temporelle qui existe entre les centrales. Les valeurs de corrélation calculées sont significatives pour des horizons pouvant aller jusqu'à 6 heures. La mise en œuvre d'un modèle qui exploiterait ces informations serait donc prometteuse pour la réduction des erreurs de prévision. L'analyse de ces corrélations spatiales et temporelles a aussi mis en évidence les limites spatiales de la propagation de ces informations ; au-delà de 250 km de distance entre centrales, les corrélations entre séries de production sont très faibles. Dans la suite, nous allons proposer un modèle de prévision de la production PV qui exploite ces corrélations spatio-temporelles pour une meilleure prédictibilité de la production. La centrale P_3 qui est la plus distante des autres centrales du jeu de données d_1 ne sera pas utilisée dans toute la suite.