Data Science M2 MIAGE

Évaluation - Sujet n°11 - du 28/01/2022 - UPJV - France

Consignes: Ce contrôle est à réaliser avec un ordinateur équipé du logiciel SCILAB. Tout support de cous est autorisé.

Contexte

On se propose de travailler des données météorologiques. Ces données sont issues d'une plateforme expérimentale équipée d'une station météo qui mesure:

- l'irradiation solaire (IRR) en watt-heures par mètre carré $[Wh/m^2]$;
- la puissance photovoltaïque (PUI) en watt $[W]$;
- la température (TEM) en degrés Celcius $[^oC]$.

Le fichier de données (*.sod*) qui vous est confié pour cette évaluation ne considère que l'une de ces 3 grandeurs, à savoir IRR, PUI ou TEM. Plus exactement, vous disposez de deux vecteurs colonnes:

- IRR, PUI ou TEM de 43200 lignes;
- IRR31, PUI31 ou TEM31 de 1440 lignes.

Dans la suite de l'énoncé, étant donné que vous ne travaillez pas sur les mêmes données, j'associerai DAT aux noms des fichiers IRR, PUI et TEM. D'une manière générale, on supposera que vous disposez de deux vecteurs colonnes que j'appellerai DAT et DAT31.

Les données ont été enregistrées lors du mois de juillet 2020 avec une fréquence d'acquisition d'une minute. Nous disposons donc d'une mesure par minute.

1. DAT correspond à l'ensemble des données enregistrées lors des 30 premiers jours (du 1er au 30 juillet);
2. DAT31 correspond aux données enregistrées le 31 juillet.

Traitement des données

1. À partir du vecteur DAT, générer une matrice de données, notée MAT, de type individus/variables, pour laquelle les variables seraient associées aux 30 premiers jours du mois de juillet. Ainsi la première colonne de MAT sera associée aux données du 1er juillet, la seconde à celles du 2 juillet, la $j^{ème}$ à celle du j juillet,.., la 30ème à celle du 30/07.

2. Calculer la matrice de corrélation $COR30 \in \mathbb{R}^{30\times30}$ relative à la matrice MAT. **Vérifier que les coefficients présents sur la diagonale de $COR30$ sont tous égaux à "1"**.

3. A partir de la matrice $COR30$, **donner les deux variables (donc les deux jours) les plus corrélées**.

jerome.bosche@u-picardie.fr

4. Vérifier le résultat précédent graphiquement. Conclusion.

5. Concernant le fichier $DAT31$, une partie des données n'a pas été enregistrée. En effet, le 31 juillet le dispositif d'enregistrement est tombé en panne durant 6 heures consécutives et, pour cette période, les mesures enregistrées correspondent à une valeur nulle. **Identifier l'heure de début et de fin de cette période. Expliquer.**

Estimation

Dans cette partie, on ne s'intéresse qu'aux données mesurées chaque jour entre 6h01 et 18h00.

6. Pour la période **6h01 – 18h00**, préciser la période correspondant aux valeurs de $DAT31$ non nulles. **On notera p cette période.**

7. Donner parmi les 30 premiers jours du mois de juillet, celui dont le vecteur de mesures associé à la période p est le plus corrélé à $DAT31$.

8. Vérifier le résultat précédent graphiquement. Conclusion.

9. À partir de la question précédente, proposer une méthode permettant d'estimer les données manquantes du 31 juillet. Les fonctions **polyfit** et **polyval** pourront être utilisées. **Expliquer la démarche.**

10. Tracer l'ensemble des données du 31 juillet en y intégrant les valeurs estimées dans la question précédente.

jerome.bosche@u-picardie.fr