Projet Math & *Signal*

Lucas DAGON, Nolann WICKERS

Janvier 2026

Description du projet

Introduction

Interpréter les données situés dans le fichier `data/strasbourg_entzheim.csv` pour pouvoir lancer des analyses descriptives des données de manière scientifique.

Organisation générale du projet

Les fichiers de ce dépôt se répartissent en grandes catégories :

- Données brutes (CSV) : fichiers originaux temporels d'observation météorologique.
- Résumés statistiques (fichiers `summary_*.csv`) : sorties de scripts de synthèse.
- Scripts et modules : lecteur CSV, filtrage de signal, et fonctions de synthèse.

Types et unités Avant toute interprétation, on doit vérifier les métadonnées pour confirmer les unités:

- Températures (`tavg, tmin, tmax`) : degrés (°C).
- Précipitations (`prcp`) : millimètres (mm) par période d'observation.
- Chute de neige (`snow`) : millimètres (mm).
- Vent (`wspd`, `wpgt`) : killomètre par heure (km/h).
- Pression (`pres`) : hectoPascal (hPa) ou millibar.
- Ensoleillement (`tsun`) : minutes d'ensoleillement.

Valeurs manquantes et qualité Les colonnes peuvent contenir des valeurs manquantes (chaînes vides, `NA`, ou `NaN`). Avant l'analyse, normaliser les valeurs manquantes en `NaN` et produire un bilan par variable : comptage, pourcentage.

Fichiers de synthèse

Exemples : `summary_tavg.csv`, `summary_tsun.csv`, etc. Ces fichiers contiennent des métriques décrivant la distribution d'une variable (moyenne, écart-type, quantiles, nombre de valeurs manquantes, etc.), ces fichiers sont crées automatiquement après execution du script.

Recommandations d'interprétation spécifiques par variable

- Températures (tavg,tmin,tmax) : vérifier cohérence physique (tmin, tavg, tmax), rechercher discontinuités journalières.
- Précipitations (`prcp`) : distribution fortement asymétrique (beaucoup de zéros) — ne pas utiliser la moyenne seule pour décrire la variabilité; employer quantiles et fréquence d'occurrence (nombre de jours > seuil).
- Chute de neige (`snow`) : traiter comme précipitation mais vérifier unité et seuils de mesure; forte présence de zéros attendue.
- Vent (`wspd,wpgt`) : distinguer vitesse moyenne et rafales; pour la sécurité, analyser les maxima à court terme et la distribution des rafales.
- Pression (`pres`) : regarder dérives lentes et oscillations synoptiques; utile pour corrélation avec autres variables.
- Ensoleillement (`tsun`) : souvent mesuré en minutes par jour; vérifier la granularité temporelle et convertir en heures si nécessaire.

1. Signification des métriques descriptives

Contexte : résumé pour une colonne numérique (ex. `tsun`).

Count nombre d'observations numériques valides.

Missing nombre de valeurs manquantes / non numériques.

Mean moyenne arithmétique (sensible aux outliers).

Std écart-type (ddof=1) : dispersion autour de la moyenne.

Min / Max valeurs extrêmes observées (vérifier plausibilité).

Quantiles $q_1, q_5, q_{25}, q_{50}(=\text{médiane}), q_{75}, q_{95}, q_{99}$: position relative des valeurs.

Skew asymétrie : > 0 queue droite (valeurs élevées rares), < 0 queue gauche.

Kurtosis aplatissement : > 0 queues lourdes, < 0 distribution plus plate que la normale.

Outliers (IQR) règle IQR : seuils $[Q1 - 1.5 \cdot IQR,\ Q3 + 1.5 \cdot IQR]$; compte des valeurs hors intervalle.

PSD peak freq fréquence dominante du spectre (Welch) ; période associée $\approx 1/f_{\text{peak}}$ (en échantillons).

Conseils pratiques :

- Si beaucoup de `Missing` : imputation ou filtrage avant analyse.
- Si `mean` $\gg$ `median` et `skew`> 0 : présence d'outliers à droite.
- Utiliser PSD et autocorr pour détecter périodicités vs bruit.
- Combiner visualisations (histogramme, boxplot, autocorr, PSD) pour validation qualitative.

4. Estimation de la proportion de bruit

But : estimer la part de variance due au bruit dans une série temporelle.

- Lissage : $\tilde{x}$ obtenu par Savitzky–Golay (ou moyenne mobile).
- Résidu (bruit estimé) : $r = x - \tilde{x}$.
- Variances :
$$\sigma^2_{\text{bruit}} = \text{Var}(r), \qquad \sigma^2_{\text{tot}} = \text{Var}(x).$$
- Proportion de bruit :
$$p_{\text{bruit}} = \frac{\sigma^2_{\text{bruit}}}{\sigma^2_{\text{tot}}} \in [0, 1].$$
- SNR (en dB) :
$$\text{SNR}_{\text{dB}} = 10 \log_{10} \left(\frac{0, \sigma^2_{\text{tot}} - \sigma^2_{\text{bruit}})}{\sigma^2_{\text{bruit}}} \right).$$
- Sorties utiles : courbe lissée $\tilde{x}$, résidu r, p_{bruit}, SNR_{dB}.

Interprétation qualitative rapide :

- $p_{\text{bruit}} \to 0$: variance majoritairement due au signal (peu de bruit).
- $p_{\text{bruit}} \to 1$: variance majoritairement due au bruit.
- Règles SNR (empiriques) : $> 20\,\text{dB}$ propre, 10–20 dB acceptable, 0–10 dB bruité, $\leq 0\,\text{dB}$ bruit dominant.
- Vérifications : tracer $x, \tilde{x}, r$; histogramme et autocorr(r) (bruit blanc attendu) ; PSD pour séparer basses fréquences (signal) et hautes (bruit).
- Limites : suppose bruit additif et que le lissage capture le signal utile ; paramètres de lissage influent fortement.