Mini Project: Explainable ML for Urban Heat Exposure Prediction (Taipei)

Purpose & Setup

- Dataset: data_output.csv | N=1104 rows, 220 spatial locations | years 2018–2021, 4 seasons
- Target: mean_t (seasonal mean temperature, °C); Features include GVI/NDVI, SVF/BVF/TVF, building-height proxies, land-use
- Model: RandomForestRegressor + median imputation; season one-hot encoded; coordinates excluded from features for interpre

Validation (spatial leakage aware)

- Why not random KFold? Spatial autocorrelation can leak information via nearby samples, inflating metrics.
- Spatial Block CV: 0.02° grid blocks (≈2 km), 66 blocks; GroupKFold ensures whole blocks are held out.
- Sensitivity check: leave-location-out CV using location id groups.

Core Results

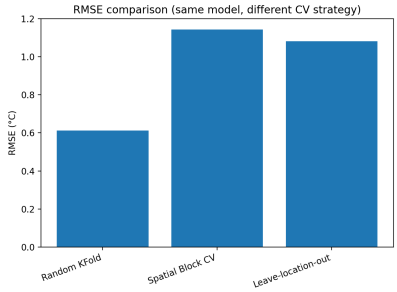
- Random 5-fold KFold: RMSE 0.61°C | MAE 0.50°C | R^2 0.96
- Spatial Block CV: RMSE 1.14°C | MAE 0.79°C | R^2 0.85
- Leave-location-out: RMSE 1.08°C | MAE 0.76°C | R^2 0.86
- Interpretation: performance drops under spatial CV → random CV was overly optimistic (evidence of spatial leakage).

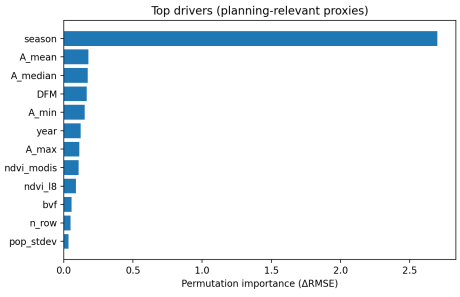
Explainability

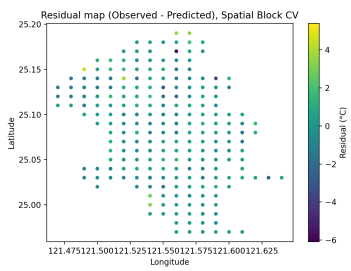
- Permutation importance used to highlight planning-relevant drivers (higher ΔRMSE = more predictive):
- season (ΔRMSE 2.70); A_mean (ΔRMSE 0.18); A_median (ΔRMSE 0.17); DFM (ΔRMSE 0.17); A_min (ΔRMSE 0.15); year (ΔR

Key Limitations

- Proxy variables & resolution: street-view and land-use proxies may not fully capture microclimate drivers.
- Temporal mismatch: features and temperatures may not be perfectly aligned in time; climate anomalies can dominate.
- External validity: model may not transfer to other cities or years without recalibration; treat as decision-support prototype.







Note: Metrics are out-of-fold estimates. Spatial block CV reduces leakage by withholding spatially contiguous blocks.