Cherry Blossom Peak Bloom Prediction (Team 5103)

Methodology, data analysis, and 2026 predictions

Team 5103

Setup

Competition context

- Predict next-year peak bloom day-of-year (DOY) for 5 sites.
- Evaluation: point accuracy (absolute error) + interval quality (coverage, then width).
- Full rules and framing are in `README.md`.

What this project does

- Reproducible R workflow in `solution.qmd`
- Independent Python check in `Solution.ipynb`
- Final outputs:
  - `cherry-predictions.csv`
  - `cherry-predictions-python.csv`
  - `cherry-predictions-final.csv`

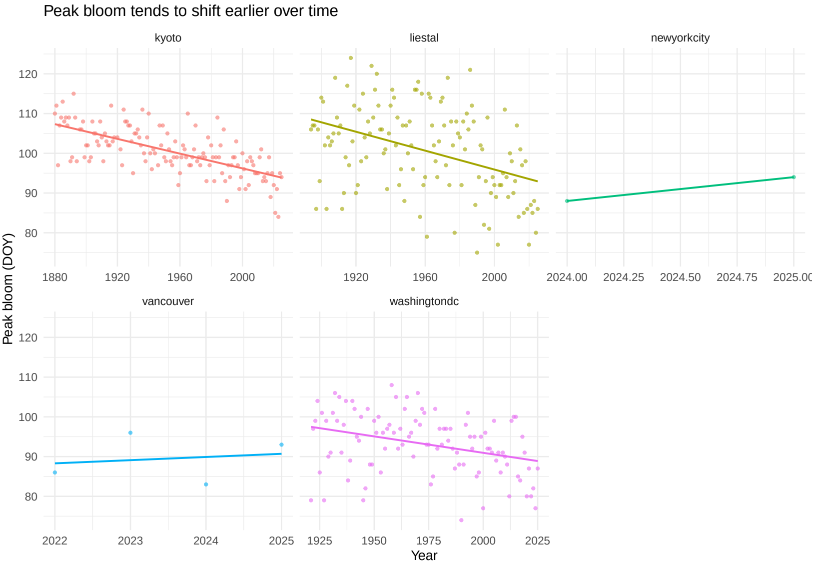
Data used

- Core competition data: Kyoto, Washington DC, Liestal, Vancouver, NYC
- Auxiliary data: Japan regional, MeteoSwiss, South Korea, USA-NPN (NYC)
- Data structure and licensing notes: `data/README.md`

Load core data

```
# A tibble: 1,080 x 7
   source      location    lat  long   alt  year bloom_doy
   <chr>       <chr>     <dbl> <dbl> <dbl> <int>     <dbl>
 1 competition kyoto      35.0  136.    44   812        92
 2 competition kyoto      35.0  136.    44   815       105
 3 competition kyoto      35.0  136.    44   831        96
 4 competition kyoto      35.0  136.    44   851       108
 5 competition kyoto      35.0  136.    44   853       104
 6 competition kyoto      35.0  136.    44   864       100
 7 competition kyoto      35.0  136.    44   866       106
 8 competition kyoto      35.0  136.    44   869        95
 9 competition kyoto      35.0  136.    44   889       104
10 competition kyoto      35.0  136.    44   891       109
# i 1,070 more rows
```

Exploratory data analysis: long-term trend



Data enrichment for NYC (USA-NPN)

```
# A tibble: 1 x 2
   year bloom_doy
  <dbl>     <dbl>
1    NA        98
```

Features used in modeling

- Time: year, centered year, squared year
- Geography: latitude, longitude, altitude (log-transformed)
- Data reliability proxy: site observation count
- Source indicator (competition vs auxiliary vs NPN)

(Implemented in `add_features` in `solution.qmd`.)

Model A: local trend

- Site-wise recency-weighted quadratic (fallback to linear/mean when sparse)
- Captures local momentum and curvature
- Implemented in `predict_local_trend` in `solution.qmd`

Model B: pooled nonlinear model

- R pipeline: GAM with smooths over year, spatial terms, altitude, site depth
- Implemented in `fit_gam_model` in `solution.qmd`
- Python pipeline: Gradient Boosting Regressor in `Solution.ipynb`

Backtesting and ensemble blending

- Rolling-origin backtest over historical years
- Compute MAE for local and pooled models
- Blend weights via inverse-MAE:
  - better out-of-sample model gets larger weight
- Implemented in rolling section of `solution.qmd`

Prediction intervals

- Split-conformal style calibration
- 90th percentile of absolute residuals by location
- Fallback to global residual quantile when needed

Final predictions (from file)

```
# A tibble: 5 x 4
  location     prediction lower upper
  <chr>             <dbl> <dbl> <dbl>
1 kyoto                90    80   100
2 liestal              88    78    96
3 newyorkcity          92    85   100
4 vancouver            92    77   108
5 washingtondc         83    76    90
```

Interpretation

- Ensemble improves robustness by combining:
  1. local historical behavior
  2. cross-site transferable structure
- Intervals are calibrated to target coverage while controlling width
- Approach is reproducible and competition-aligned

Limitations and next steps

- Residual weather shocks remain hard to predict
- Potential improvements:
  - engineered temperature covariates
  - richer uncertainty modeling
  - model stacking with strict out-of-sample validation

Reproducibility

- Render slides:

```
quarto render slides.qmd
```

- Main artifacts:
  - `solution.qmd`
  - `Solution.ipynb`
  - `README.md`
  - `data/README.md`