Project Talk for physics7516: Machine Learning for Quantum Scientists

Spatiotemporal abundance forecasting of the North American Breeding Bird Survey (BBS)

Hannah Wehnes

31.07.2026

Motivation and task

- BBS: standardised route-based bird counts since 1966
- **Task**: given a route, species, year, and associated features → predict observed bird count
- Train on 1966–2008, forecast 2009–2019
- Features: BBS data + bioclimatic variables + landuse type
- Model: LightGBM (gradient boosting), compared against simple baselines

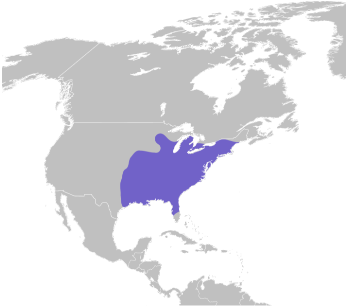


Photo [1], Map: [2]

Dataset

BBS survey data

- Count data: one row per (route, year, species) in two survey eras:
  - 10-stop protocol (1966–1996)
  - 50-stop protocol (1997–present)
- Route metadata:
  - Starting-point coordinates
  - Stratum
  - Bird Conservation Region (BCR)
- Weather run-list: reconstruct true zero counts
- **Limitation:** only route starting points available → environmental variables extracted using a buffer around each route

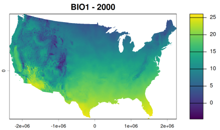
Environmental data

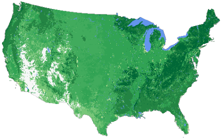
Bioclimatic data (CHELSA) [3]

- 19 BIO variables
- 1km spatial resolution

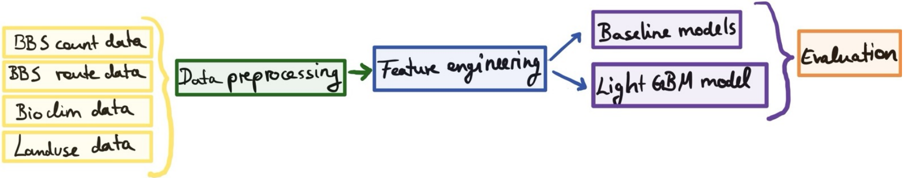
Landuse data (PANGAEA) [4]

- 14 categorical land-use classes
- 1km spatial resolution



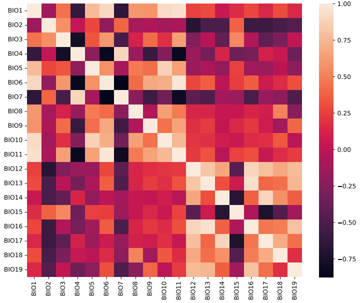


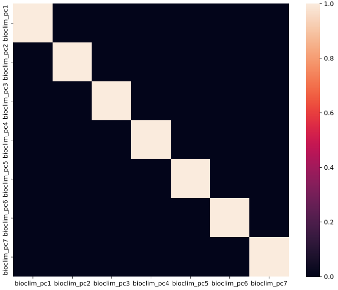
Pipeline overview



Feature engineering

- **Temporal features**
  - Lagged count: $\texttt{count\_lag1}_{r,s,y} = c_{r,s,y-1}$; route r, species s, year y
  - 5-year rolling mean: $\texttt{count\_mean\_5yr}_{r,s,y} = \frac{1}{5}\sum_{i=y-5}^{y-1} c_{r,s,i}$; smooths single-year anomalies
- **Bioclimatic features**
  - 19 BIO variables are correlated $\rightarrow$ redundant information
  - PCA: converts correlated variables into a set of uncorrelated components





Baseline models

- **Global mean**: predicts the same count for every route and year (training-set average)
- **Per-route mean**: predicts each route's historical average (captures spatial differences only)
- **Per-route trend**: linear trend fitted separately for each route (captures long-term change)
- **Moving average**: mean of the previous 5 years (captures short-term temporal persistence)

What is LightGBM?

- **Decision tree**: simple model that splits data by feature thresholds to predict a value
- **Gradient boosting**: combine many weak trees into one strong ensemble
  - each new tree is trained to minimize the loss function
  - algorithm computes gradient of loss w.r.t. to current predictions
  - the new tree is trained to predict this gradient
  - repeat until a stopping criterion is met
- **LightGBM** [5]: an open-source gradient-boosting framework

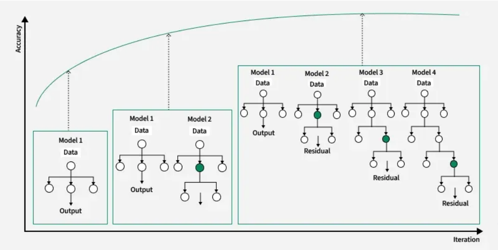
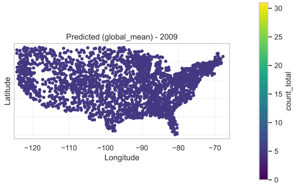


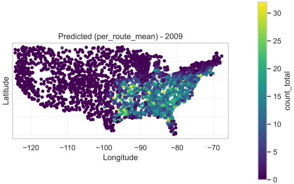
figure [6]

Prediction maps: actual vs. predicted (2009)

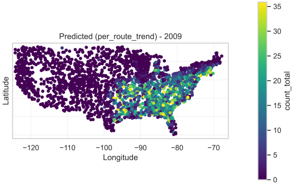
Global Mean

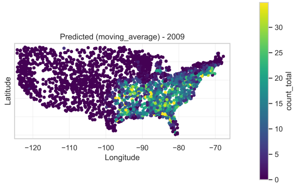


Route Mean

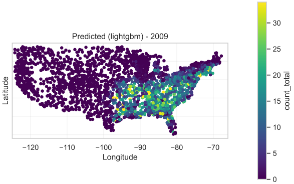


Route Trend

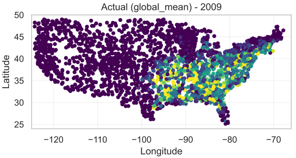




Moving Average



LightGBM



Actual Data

Evaluation metrics

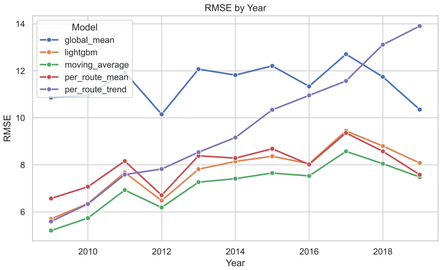
- **RMSE** (Root Mean Squared Error)
  - $\sqrt{\frac{1}{N}\sum(y_i - \hat{y}_i)^2}$
  - penalises large errors more strongly $\rightarrow$ sensitive to big misses
- **MAE** (Mean Absolute Error)
  - $\frac{1}{N}\sum|y_i - \hat{y}_i|$
  - robust to outliers $\rightarrow$ average error in bird counts
- R^2
  - $1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$
  - proportion of variance explained; $1 =$ perfect fit

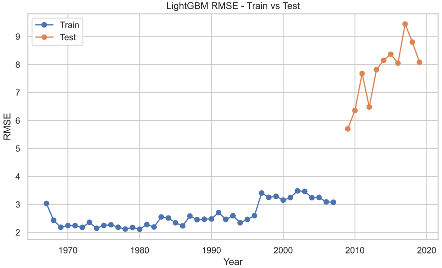
Where: $y_i =$ actual value, $\hat{y}_i =$ predicted value, $\bar{y} =$ mean value, $N =$ number of samples

Results: LightGBM vs. Baselines

Model	RMSE	MAE	R^2
Moving average	7.17	3.38	0.598
LightGBM	7.81	3.59	0.524
Per-route mean	8.00	3.77	0.500
Per-route trend	9.91	4.19	0.232
Global mean	11.50	7.89	−0.033

- Moving average achieves the best overall performance
- LightGBM overfitting
- Short-term temporal persistence dominates current environmental predictors





Key findings

- LightGBM outperforms global and route-level baselines
- Moving average performs best, highlighting strong temporal persistence
- Environmental features add value but do not outperform temporal information

Future work

- Use full BBS route geometries with ArcGIS data [7]
- Improve features and tune hyperparameters to reduce overfitting

References I

[1] Rhododendrites. *Tufted titmouse (84917).jpg*. Photograph on Wikimedia Commons. Accessed: 2026-07-30. Mar. 2025. URL: https://commons.wikimedia.org/wiki/File:Tufted_titmouse_(84917).jpg.

[2] Jebbles. *Baeolophus bicolor map.svg*. Distribution map on Wikimedia Commons. Accessed: 2026-07-30. Aug. 2021. URL: https://commons.wikimedia.org/wiki/File:Baeolophus_bicolor_map.svg.

[3] Philippe Brun et al. *CHELSA-BIOCLIM+: A Novel Set of Global Climate-Related Predictors at Kilometre Resolution*. Accessed: 2026-07-30. 2022. DOI: 10.16904/envidat.332. URL: https://www.envidat.ch/dataset/chelsa-bioclim-plus.

[4] Karina Winkler et al. *HILDA+ version 2.0: Global Land Use Change between 1960 and 2020*. dataset. 2025. DOI: 10.1594/PANGAEA.974335. URL: https://doi.org/10.1594/PANGAEA.974335.

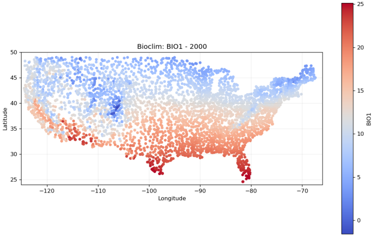
[5] The LightGBM developers Microsoft Corporation. *lightgbm-org /LightGBM*. Accessed: 2026-07-28. URL: https://github.com/lightgbm-org/LightGBM.

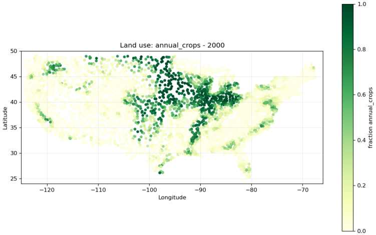
[6] GeeksforGeeks. *LightGBM (Light Gradient Boosting Machine)*. 2025. URL: https://www.geeksforgeeks.org/machine-learning/lightgbm-light-gradient-boosting-machine/ (visited on 07/28/2026).

[7] *BBS Routes Feature Service*. URL: https://services.arcgis.com/8df8p0NlLFEShl0r/arcgis/rest/services/bbs_routess/FeatureServer (visited on 07/30/2026).

Dataset: environmental extraction

- Around each route's starting point: circular buffer, radius = route length
- **Bioclimatic**: mean value of valid pixels within buffer
- **Landuse**: fraction of valid pixels per landuse class within buffer



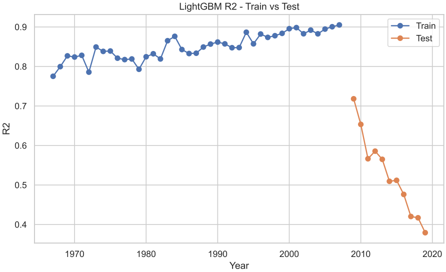


Left: mean annual temperature (BIO1) at each route, 2000. Right: annual cropland fraction at each route, 2000.

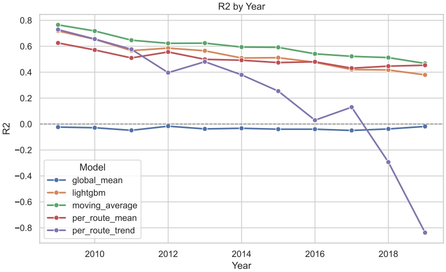
LightGBM model: setup

- **Complexity control**: max_depth=8, num_leaves=64, min_data_in_leaf=30
- **Regularisation**: L1 and L2
- **Robustness**: feature_fraction=0.7, bagging_fraction=0.7
- **Learning**: learning_rate=0.03, num_iterations=1000
- species_id, stratum, bcr treated as categorical

Results: evaluation over time



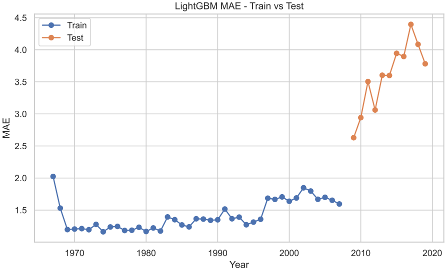
LightGBM: train vs. test R^2 by year



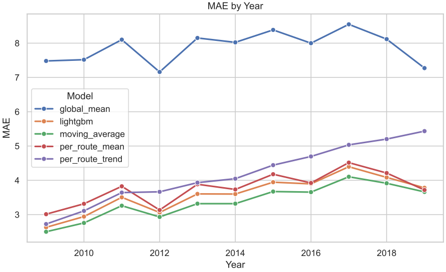
R^2 by year: LightGBM vs. all baselines

- Test R^2 declines over the forecast horizon
- Moving average stays ahead of LightGBM consistently

Results: MAE over time



LightGBM: train vs. test MAE by year



MAE by year: LightGBM vs. all baselines

- Training MAE stays near 2 bird counts; test MAE holds close to 4 throughout
- LightGBM and per-route mean track closely; both trail the moving average

Principal Component Analysis (PCA)

1. Standardise variables Variables are scaled to zero mean and unit variance :

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

- x_{ij} observation i of variable j
- $\bar{x}_j$ mean of variable j and its standard deviation σ_j

2. Compute covariance matrix

$$C = \frac{1}{n-1} X^\top X$$

3. Find principal directions Solve the eigenvalue problem:

$$C\mathbf{v}_i = \lambda_i \mathbf{v}_i$$

- $\mathbf{v}_i$: eigenvector, principal component directions
- λ_i: eigenvalue, variance explained by each component

4. Project data Transform observations into the PC space:

$$\mathbf{z}_i = X\mathbf{v}_i$$

Each PC is a linear combination of the original climate variables:

$$PC_i = X\mathbf{v}_i$$

Explained variance

$$r_i = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

Choose smallest k such that:

$$R_k = \sum_{i=1}^{k} r_i \geq 0.95$$

PCA results

- 19 climate variables reduced to **7 principal components**
- 7 PCs explain more than **95%** of the total variance
- Principal components are uncorrelated by construction
- Similar observations (colored by BCR) remain close together in the reduced space
- These 7 PCs are used as climate features for LightGBM