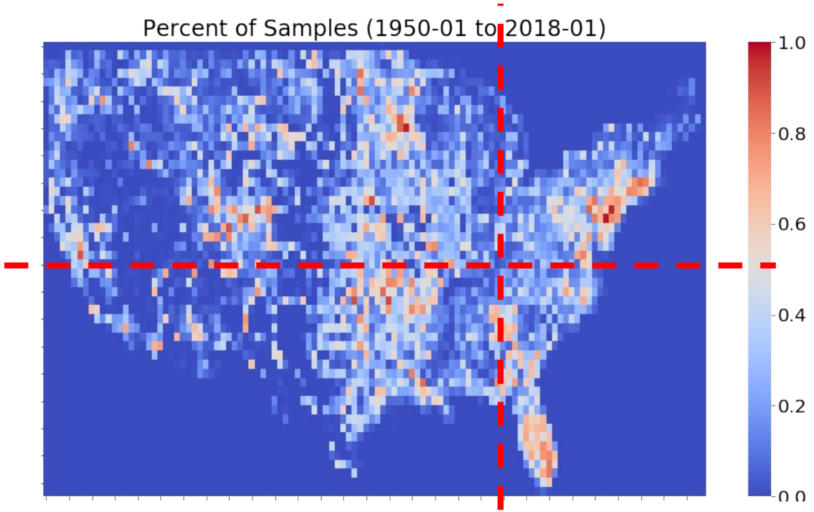
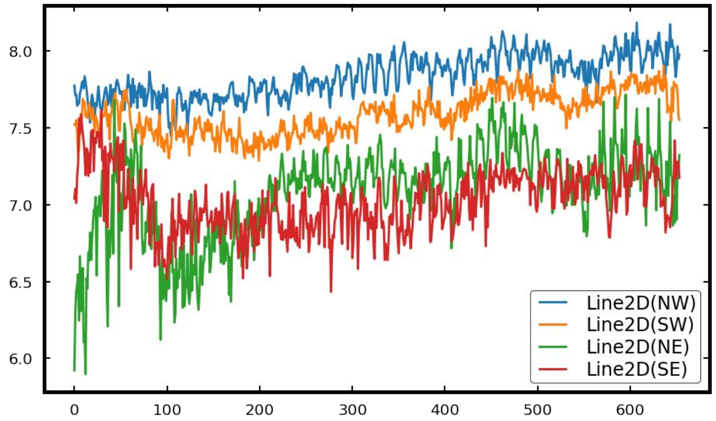
A comparison of algorithms for Spatial-Temporal Data Imputation

Mengqi Liu

Project Overview

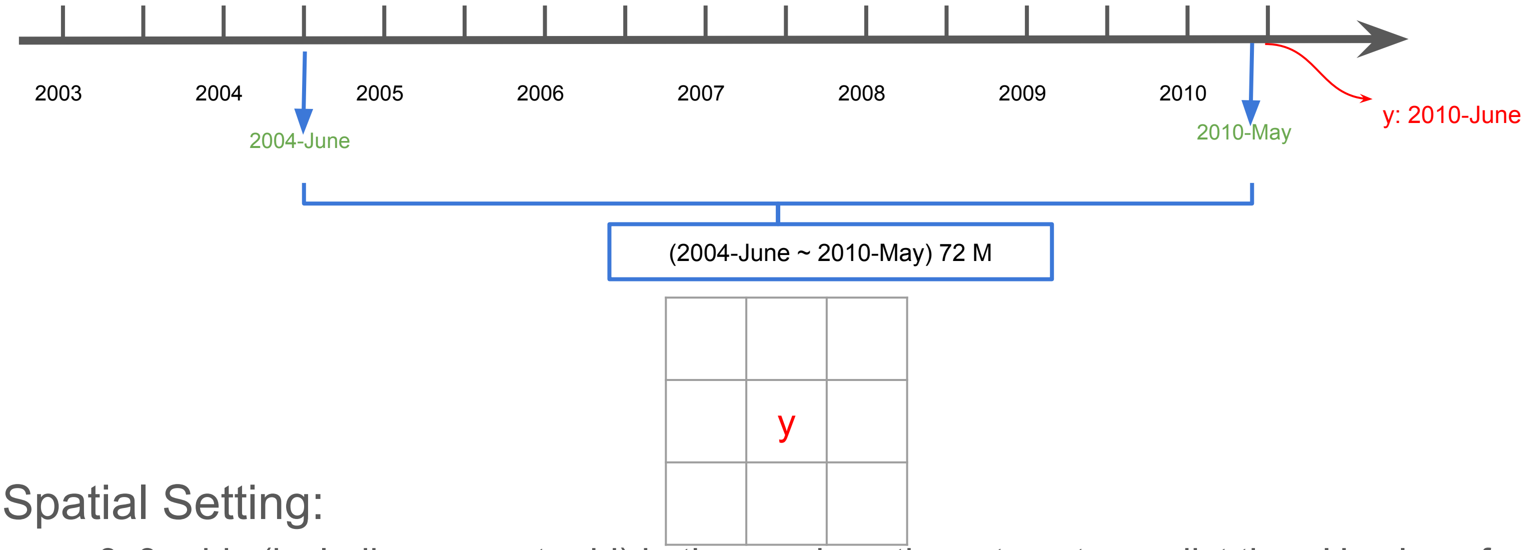
- **Objective:** impute missing values in spatial-temporal data
- **Challenge:**
  - Data does not fit MAR (missing at random)
  - In total, there are 56.7586% grid (see next slides) has value.
  - Grids in different area has different distribution





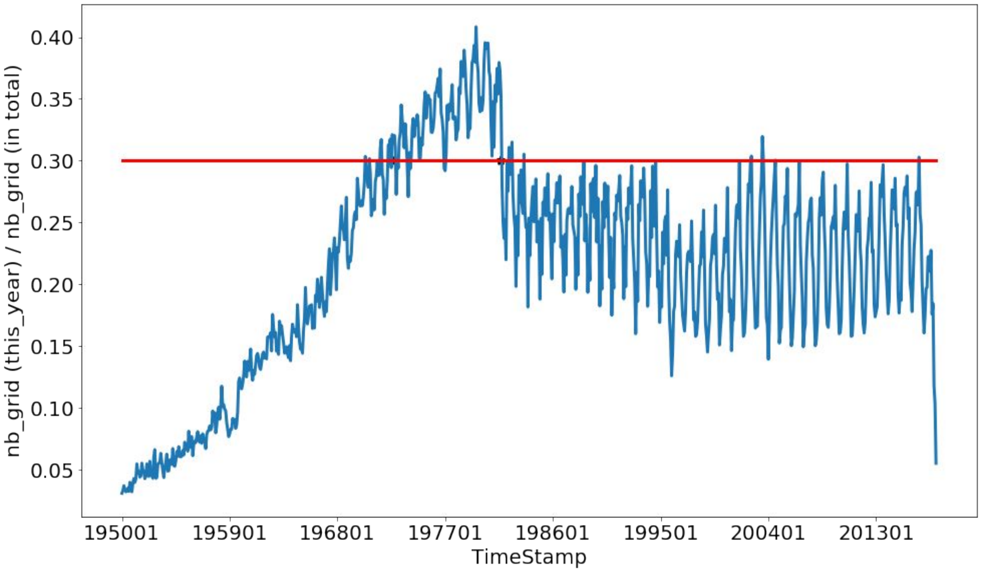
Task Description - Input & Output

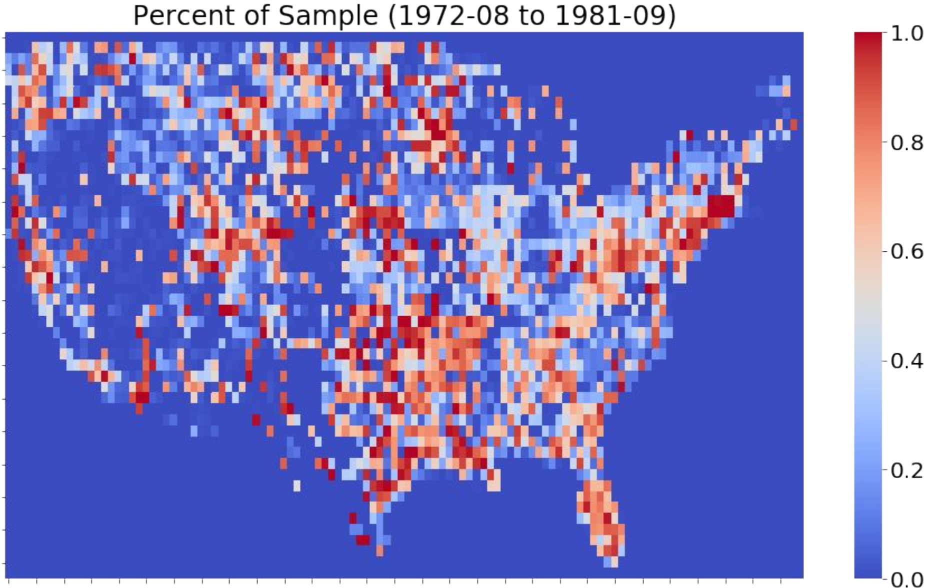
- Temporal Settings in previous experiments (X for input, y for output)



- Spatial Setting:
  - 3x3 grids (including current grid) in the previous time steps to predict the pH value of current grid at current time step.
  - grid size: ½ latitude x ½ longitude
- Model: single XGBoost for continental US

Data Used in Experiment





- Most of the time step doesn't have much data
- Only use the continuous time steps that has above 30% of data (grid) that has value

Prediction Method - XGB [1]

Data

Example	Age	Gender
X1	?	male
X2	15	?
X3	25	female

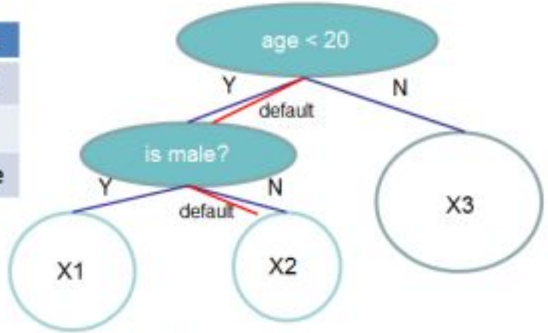


Figure 4: Tree structure with default directions. An example will be classified into the default direction when the feature needed for the split is missing.

This figure is taken from the original paper of XGB.

Experiment (Impute by grids at the same timestep)

RMSE from XGBoost			
None	1.7363	**EM**	0.3523
Fast KNN	0.2428	**Mean**	0.2338
MICE	0.2338	**Median**	0.2378
Mode	0.3058	**Random**	0.3686

Experiment (Impute by all timestep of current grid)

RMSE from XGBoost			
None	1.7363	**EM**	0.5353
Fast KNN	0.2544	**Mean**	0.2729
MICE	0.2729	**Median**	0.2729
Mode	0.2898	**Random**	0.8618

Reference

[1] Chen, T., He, T., & Benesty, M. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1-4.