**Water Quality Prediction Dataset**

**Introduction**

Here we want to forecast the spatio-temporal water quality in terms of the “power of hydrogen (pH)” value for the next day based on the input data, which is the historical data of other water measurement indices. The input data consists of daily samples for 36 sites, providing measurements related to pH values in Georgia, USA. The input features consist of 11 common indices including volume of dissolved oxygen, temperature, and specific conductance (see details in dataset). The output to predict is the measurement of 'pH, water, unfiltered, field, standard units (Median)'.

There are two major water systems to consider: one is centered on the city of Atlanta while the other is centered on the eastern coast of Georgia. This information indicates spatial depenency among different locations which are important to the forecast.

**Processed Data**

Download link: [[Dataset](http://mason.gmu.edu/~lzhao9/pages/dataset_pages/datasets/spatial/water_dataset.mat)]

**Data format:** \*.mat (use Matlab to open)

**Data description:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Type** | **Size** | **Description** |
| features | array of strings | 1\*11 | a list of water indices to measure |
| location\_ids | array of integer | 37\*1 | IDs of the water stations |
| X\_te | array of matrices | 1\*282 | test set input data: water indices for 282 contiguous dates until 2018-01-01   * each element is a 37\*11 matrix: 37 spatial locations by 11 features |
| X\_tr | array of matrices | 1\*423 | training set input data: water indices for 423 contiguous dates from 2016-01-28   * each element is a 37\*11 matrix: 37 spatial locations by 11 features |
| Y\_te | array of matrices | 37\*282 | test set output data: water quality for 37 locations in 282 contiguous dates until 2018-01-01 |
| Y\_tr | array of matrices | 37\*423 | training set output data: water quality for 37 locations in 423 contiguous dates from 2016-01-28 |
| location\_group | array of cells | 1\*3 | the groups of water stations, each group forms a connected spatial network (i.e., water system) |

**Data Source**

This dataset is arranged and partly derived from the [United States Geological Survey](file:///C%7C/Users/zhaol/AppData/Local/Microsoft/Windows/INetCache/IE/6BQSMFDY/https%7C/www.usgs.gov/): [[External Link](http://mason.gmu.edu/~lzhao9/pages/dataset_pages/https%7C/waterdata.usgs.gov/nwis/dv/?referred_module=qw?referred_module=qw)]

**Citation**

To use these datasets, please cite the papers:

Liang Zhao, Olga Gkountouna, and Dieter Pfoser. 2019. Spatial Auto-regressive Dependency Interpretable Learning Based on Spatial Topological Constraints. *ACM Trans. Spatial Algorithms Syst.* 5, 3, Article 19 (August 2019), 28 pages. DOI:https://doi.org/10.1145/3339823