Data Imputation for IoT Using Spatio-Temporal Variational Autoencoder数据集使用说明

1. PRSA
2. **数据集介绍：**

This data set includes hourly air pollutants data from 12 nationally-controlled air-quality monitoring sites. The air-quality data are from the Beijing Municipal Environmental Monitoring Center. The meteorological data in each air-quality site are matched with the nearest weather station from the China Meteorological Administration. The time period is from March 1st, 2013 to February 28th, 2017. Missing data are denoted as NA.

1. **属性说明：**  
   No: row number  
   year: year of data in this row  
   month: month of data in this row  
   day: day of data in this row  
   hour: hour of data in this row  
   PM2.5: PM2.5 concentration (ug/m^3)  
   PM10: PM10 concentration (ug/m^3)  
   SO2: SO2 concentration (ug/m^3)  
   NO2: NO2 concentration (ug/m^3)  
   CO: CO concentration (ug/m^3)  
   O3: O3 concentration (ug/m^3)  
   TEMP: temperature (degree Celsius)  
   PRES: pressure (hPa)  
   DEWP: dew point temperature (degree Celsius)  
   RAIN: precipitation (mm)  
   wd: wind direction  
   WSPM: wind speed (m/s)  
   station: name of the air-quality monitoring site
2. **使用说明**

每一个文件当做一个device，不同的属性当做sensor。数据预处理时，把所有NA值替换为前后两值的加和平均，使用时对比算法只使用PM2.5, PM10, SO2, NO2, CO, O3六个属性VAE则使用该六个属性作为Monitor data，其他六个作为ST-VAE算法External Environment。

1. GHG

**1） 数据集介绍：**

This data set contains time series of greenhouse gas (GHG) concentrations at 2921 grid cells in California created using simulations of the Weather Research and Forecast model with Chemistry (WRF-Chem). Each grid cell covers an area of 12 km by 12 km, and there is one data file per grid cell. Each file contains 16 time series of GHG concentrations. The data points in the time series are spaced 6 hours apart (4 samples per day) over the period May 10 â€“ July 31, 2010. The first 15 rows are time series of GHG tracers released from 14 distinct spatial regions in California and one outside of California. The last row corresponds to the time series of â€œsynthetic GHG observationsâ€ generated with EDGAR emissions of HFC-134a scaled by a factor 0.7 and with noise added..

**2） 属性说明：**

Each file in the data set is labeled ghg.gid.siteWXYZ.dat, where WXYZ is an integer location ID described in our manuscript.  
  
At each location,  
Rows 1-15: GHG concentrations of tracers emitted from regions 1-15  
Row 16: GHG concentrations of synthetic observations  
Columns 1-327: GHG concentrations every 6 hours from May 10 â€“ July 31, 2010.  
  
All GHG concentrations are in units of parts per trillion.

**3） 使用说明**

每一个文件当做一个device，不同的行当做不同的sensor（总共16个）。随机选7个文件当做7个设备（不要连续选取或者相同间隔选取），16个属性全部作为对比算法和ST-VAE的数据。

1. GNFUV（备选）
2. **数据集介绍：**

The data-set comprises (4) sets of mobile sensor readings data (humidity, temperature) corresponding to a swarm of four (4) Unmanned Surface Vehicles (USVs). Each USV set contains records of the format: {'USV-ID'; 'humidity-value'; 'temperature-value'; 'experiment-id';'sensing-time'}  
The swarm of the USVs is moving according to a GPS pre-defined trajectory, whose relative way-points are specified in the README.pdf file. The USVs are floating over the sea surface in a coastal area of Athens (Greece).

1. **属性说明**

'device' = USV ID (String)  
'humidity' = sensed humidity value from the USV sensor (real value)  
'temperature' = sensed temperature value from the USV sensor (real value)  
'experiment' = 1 (constant real value)  
'time' = the sensing and reporting time (real value)

1. **使用说明**

每一个文件当做一个device，不同的属性当做sensor。humidity, temperature, experiment三个属性全部作为对比算法和ST-VAE的数据。