

# 自动监测水质数据 统计预报算法

算法&开发：卜泽昊

数据&调研：张静桥

数据&接口：程舒鹏

测试&集成：苗春葆

2020 年 10 月 24 日

# 1. 背景

自2019年下旬，大数据统计模型在政府的生态环境业务中占取了越来越重的地位。尤其在监测方面，各级监测单位已被明确要求需具有对其监测数据模拟分析的能力。在2020年3月，生态环境部编制了《关于推进生态环境监测体系与监测能力现代化的若干意见（征求意见稿）》并向全社会公开征求意见，现已正式发布。

上述政策不仅带来了统计模型的愈多需求，也同时利好统计模型的研发环境。现阶段，最主要的阻碍统计模型在水环境应用和发展的阻碍便是数据量&类的不足和数据质量差。而从上述政策中可以明显的觉察出国家在这方面改良进步的决心。比如：该意见中提出的“实现大监测、确保真准全、支持大保护”的发展思路；各级生态环境部门统一数据管理的方法；对企业“谁排污、谁检测”的原则这三条，分别可以解决：数据质量、数据来源和字段、数据覆盖面这三个问题。

同时，在该意见中也指出了未来监测数据的应用方向。

其一，指出“优化监测指标项目和评价方法，形成统一标准与因地制宜结合、定量评价与定性评价相结合、现状评价与预测预报相结合的生态环境质量评价体系，建立生态环境质量指数、生态环境综合质量指数等复合型评价指标并试点应用，科学反应生态环境质量和污染治理成效。”此条意见委婉的对环保一刀切的作风提出了否定，并由此提出“对环境保护标准线、企业排放限制线要在以科学为支撑的情况下合理的划定”的建议。其反映到水模型上则是：A，模型需要能够对水质变化追根溯源，说清原因，结合上下游、生态大环境特征来综合评判；B，在对环境污染和企业排放定性时，要考虑更多的因素，比如生态圈、社会、经济等；C，加强了对预测预报要求。

其二，提出“推进重点流域水环境预测预警业务和技术体系建设，逐步开展土壤风险评估和生态风险预警研究”。

其三，“推动物联网、区块链、人工智能、5G通信等新技术在监测监控业务中的应用，促进智慧监测发展”。

以上二、三两条分别从业务&技术角度强调了预测预警和大数据应用的需求。

## 2. 模型

经过对已有的水质监测数据进行多轮讨论与测试，设计两大主要模块任务，主要包括水质监测数据的线性与非线性插值模拟和水质指标预测。

### I. 水质模拟

目前自动站水质监测数据，由于仪器稳定性与数据传输等非人为因素产生大量缺失及无效数据，极大的限制了开展各类算法的测试开发，假设给定数据集是一个矩阵 $A$ ， $A$  的维度是 $m \times n$ ，数据的总体缺失率等于 $1-m$ 除以数据集无缺数据条数，各个特征的缺失率为 $1-$ 对应列非空值除以数据集无缺数据条数。

一般当特征的缺失率大于等于60%，该特征可剔除，即该特征不适合用于建模分析，当数据缺失率小于60%，可尝试用算法进行补充。目前主流用于填充的算法有：1、利用同一值填充（比如均值，中位值）。简单，易行，但误差大；2、前向填充、前后填充、线性差分填充，简单易行，只考虑了局部线性关系，忽视了非线性关系；3、基于灰色预测模型填充缺失值；4、K近邻算法，计算缺失点的最 $k$ 近邻的点，利用其相似点填充缺失值；5、当数据量较大时，可采用基于卷积神经网络的缺失数据填充方法。首先，分别针对时间序列数据的时间相关性和传感器节点间的空间相关性，使用卷积神经网络填充模型对缺失数据进行单维度相关性的填充；然后，根据时间维度和空间维度的填充结果，考虑时空相关性对缺失数据影响，进行填充。

对比自动站常规监测的数据项目，其水温、PH、电导率、浊度等数据有采样频率更低、数据稳定、缺失较少、特征易于预测等特点，故设计基于此类常规监测数据与主要水质监测指标数据建立线性和非线性模型，最大程度模拟缺失和未来预报数据。其中：

线性类模型包括LinearRegression线性回归、Lasso回归、Ridge岭回归、ElasticNet回归；非线性类模型包括随机森林模型RandomForestRegressor、极端随机树模型Extremely Randomized Trees、提升树模型AdBoost、Xgboost、GBRT等。

当历史水质监测数据大量缺失，形成一段的空数据时期，此时利用线性插值和填充的方式产生的数据，会产生大量单调趋势数据，对预测阶段模型的趋势分析产生较大噪声，可基于数据缺失较少的数据项进行非线性映射产生随机性强且特征维度一致的水质模拟数据。

## II. 水质预报

在目前的水环境监测中，整体数据现状是：数据量级整体较小，由于大部分省市的自动检测断面都是在近几年部署的，导致历史数据较少，严格来说目前远远不够大数据的量级；自动检测数据能够达到小时级别，但经常会因为仪器或传输的原因导致数据缺失或者检测值超出检测下线，异常情况难以判断；手工检测数据一般每月一次或几次，整体数据量较少，但地方环保部门普遍对手工数据比较倚重和信任；支持水质模拟的特征因子中：水文（流量、水位等）较难获取，历史数据量级中等，实时数据多为天数据；气象历史数据较为全面，历史预测数据也较为准确，一般为天数据；实时数据可达分钟级别，数据量最大。

在水质预测模块中，对于小时频率以下的检测数据预报需求，可以采用深度学习类算法。充分利用深度学习的非线性逼近的功能，发现检测数据之间的规律，不断优化调参，精确预报未来小时数据的预测值，选择深度学习的理由是：小时以下的数据量级可观，适合深度学习，无需人工特征工程，和深度学习强大的非线性拟合能力，建议模型主要有：LSTM，GRU,RBF神经网络模型，多RBF神经网络模型。

LSTM，GRU是作为大数据时序分析的首选。这两个算法都具有timestamp参数，即模型输入数据就是一段时间的特征，进而通过非线性函数找到输入数据之间的关系。同时又记忆门和遗忘门的设置来保证学到的规律和知识能传下去，很适合时序数据分析。

RBF神经网络除了具有一般神经网络的优点，如多维非线性映射能力，泛化能力，并行信息处理能力等，还具有很强的聚类分析能力，学习算法简单方便等优点。其中，径向基函数(RBF)神经网络是一种性能良好的前向网络利用

在多维空间中插值的传统技术，可以对几乎所有的系统进行辨识和建模。它不仅在理论上有着任意逼近性能和最佳逼近性能，而且在应用中具有很多优势。比如，与利用Sigmoid 函数作为激活函数的神经网络相比，RBD计算速度大大高于一般的BP 算法。同时，相比BP 网络，RBF不仅在理论上是前向网络中最优的网络，而且在学习方法也避免了局部最优的问题。

目前在应用中已经证明：一个RBF网络，在隐层节点足够多的情况下，经过充分学习，可以用任意精度逼近任意非线性函数，而且具有最优泛函数逼近能力，另外，它具有较快的收敛速度和强大的抗噪和修复能力。

对于日频率检测数据预报，在数据量较大的情况，也可采用上述深度学习的方式，否则建议选择Xgboost模型。极端梯度提升XGBoost (EXtreme Gradient Boosting) 是集成学习方法的王牌。在近些年Kaggle数据挖掘比赛中，大部分获胜者都采用XGBoost。XGBoost在绝大多数的回归和分类问题上表现的十分顶尖。相比于传统的回归模型xgboost算法优点如下：a) xgboost支持线性分类器，这个时候xgboost相当于带L1和L2正则化项的线性回归；b) 相比于传统的回归模型，xgboost对代价函数进行了二阶泰勒展开，同时用到了一阶和二阶导数。支持自定义代价函数，只要函数可一阶和二阶求导；c) xgboost在代价函数里加入了正则项，用于控制模型的复杂度。正则项里包含了树的叶子节点个数、每个叶子节点上输出的score的L2模的平方和，正则项降低了模型的方差，防止过拟合；d) xgboost借鉴了随机森林的做法，支持列抽样，不仅能降低过拟合，还能减少计算；f) 对于特征的值有缺失的样本，xgboost可以自动学习出它的分裂方向；g) xgboost工具支持并行。xgboost在训练之前，预先对数据进行了排序，然后保存为block结构，后面的迭代中重复地使用这个结构，大大减小计算量。这个block结构也使得并行成为了可能，在进行节点的分裂时，需要计算每个特征的增益，最终选增益最大的那个特征去做分裂，那么各个特征的增益计算就可以开多线程进行；

对于月手工检测数据的预报，若数据量大于等于500，建议采用XGBoost或随机森林；若数据量较小（大于100小于500），建议采用线性回归。如 ElasticNet模型，其理由简单、易解释，可降低过拟合的可能性（数据量

小，选择复杂的模型容易过拟合）；当数据量极小（小于100），可使用无偏灰色预测模型。由于灰色预测具有少量数据建模和累加生成可增加历史数据的规律性的特点 灰色预测模型主要用于趋势性强、波动不大的短期水质预测问题，在数据较少的情况下,可以获得比较准确的预测结果。

### 3. 工程

DeepWater水环境数据智能预报是基于全国各水质监测自动站的历史数据，对未来3~7天的四类水质指标数据进行预测，其模型包括Arima、Fbprophet、LSTM、GRU等；同时考虑了大规模并行多站点多指标多时刻同时进行预测模拟和计算；相关说明可参考<http://47.92.132.84:3000/buzh/DeepWater/src/dev>。

#### environment 运行环境

本项目基于tensorflow2.0开发深度学习相关部分，由于项目需要深度学习相关库，部门服务器无法满足环境需求；

另配置miniconda3环境，相关环境打包文件可联系[buzh@3clear.com]，感谢关注和支持。

- 将miniconda3打包环境解压到指定位置，修改scripts/run.sh中export PYTHON="{your\_abspath}/miniconda3/bin/python"

#### config 配置目录

项目总体配置目录，包含公共项目配置和模块独立配置。

- common.py 公共配置模块，设计用于整体框架相关配置，

主要配置信息：

1. proj\_home 项目路径
2. common\_params 公共配置内容
  - |\_\_ FREQ 数据频率
  - |\_\_ SQLite 数据库路径
  - |\_\_ PostgreSQL 数据库信息
3. forecast\_params 预报配置内容
  - |\_\_ 预报公共配置
  - |\_\_ Arima相关配置
  - |\_\_ Fbprophet相关配置
  - |\_\_ LSTM相关配置
  - |\_\_ GRU相关配置

#### dao (Data Access Object)数据对接模块

设计功能包括数据库、文本数据对接，其中包含实时数据、离线数据对接。

- I. settings.py 模块内测试配置文件，后期统一到config模块内
- II. test.py 模块内测试脚本，测试数据读写功能
- III. orm 关系映射



IV. db\_orm.py 数据库连接

V. SQLite\_Static.py 手动更新数据表模型 不推荐

VI. SQLite\_dynamic.py 自动获取数据库对应表数据模型

VII. PostgreSQL\_dynamic.py 对接业务流程PostgreSQL的数据模型

## **data 数据目录**

I. waterTestDB.db 本地数据信息化及开发用的SQLite数据库

## **docs 文档目录**

I. 统计预报说明文档.pdf

II. ARIMA模型.pdf

III. fbprophet\_paper.pdf

## **lib 算法库封装模块**

机器学习算法

I. LinearModel 线性模型

II. NonLinearModel 非线性模型

III. EnsembleModel 集合模型

深度学习算法时间序列的神经网络模型

I. SingleShot 并行多指标一次性预测多时次框架

II. DataWindow 滑动数据集构建器

统计学习算法

I. ARIMA 算法

II. Fbprophet 算法

## **logs 日志输出目录**

其中:

\*.log为设计日志记录功能输出

\*.out为当次任务打印输出信息

## notebooks

包含数据前分析阶段的notebook，可在jupyter-notebook & jupyter-lab & VSCode等环境中打开；

**EDA**(Exploratory Data Analysis)，数据探索性分析是指在不清楚数据内容、质量、分布的情况下，对数据进行不同纬度的探索性分析；

EDA的目标是前期不受行业经验限制，对数据进行统计分析，之后结合专业经验提升对于数据的感知能力。

I. .ipynb文件环境与运行环境可能存在一些库版本不同

## scripts 控制任务脚本

I. run\_forecast.py 运行不同模型脚本

II. run.sh 运行项目主脚本，需提供预测时间，时间格式为{\$yyyy\$mm\$dd\$hh}

III. task2file.py 用于生成项目并行运行的任务列表文件脚本


## src 任务&功能模块

DataSimulation基于五项常规监测模拟四项水质指标的数据处理&非线性模型训练、预测任务。

功能设计：

基于五项常规监测指标{水温、PH、电导率、浊度、溶解氧}，建立线性&非线性模型模拟四项水质指标{高锰酸盐、氨氮、总磷、总氮}数据；

进而与水质指标线性插值填补数据进行对比，弥补间隔时间过长的线性填补劣势。

DataForecast  四项水质指标预测功能任务，包括ARIMA模型、机器学习模型、深度学习模型。

DataQC  对接总站水质监测数据的质控处理模块，包括：

1. 统计上下限分位数剔除高低异常值；
2. 线性插值填补；
3. 前向或后向填补；

运行方式 `cd {$your_proj_home}/DeepWater/scripts; {$PYTHON} -n '站点名称' -m 'qc' -i 'all' -s '2020060600' -e '2020101000'`

其中:

-n 站点名称

-m 模型名称, 此处唯一为qc

-i 质控指标, 此处唯一为all, 代表五常+四项

-s 质控时段开始时间, 一般小时位为00、04、08、12、16、20

-e 质控时段结束时间, 一般小时位为00、04、08、12、16、20

DataDiagnose数据诊断功能

- TODO

DataEvaluate预测评估功能

- TODO

## task 任务记录目录

名称为{task\_yyyymmddHH.txt}的文件记录了该预报时次所有任务, 方便pbatch调用进行并行计算。

文件样例

```
cd /public/home/buzh/water/DeepWater/scripts; /public/home/buzh/env/miniconda3/bin/python run_forecast.py -n '白马寺' -m 'Arima' -i 'tn' -s '2020051108' -e '2020051108' >& /public/home/buzh/water/DeepWater/logs/logs_forecast/task_2020051108.out
```

```
cd /public/home/buzh/water/DeepWater/scripts; /public/home/buzh/env/miniconda3/bin/python run_forecast.py -n '白马寺' -m 'Fbprophet' -i 'tn' -s '2020051108' -e '2020051108' >& /public/home/buzh/water/DeepWater/logs/logs_forecast/task_2020051108.out
```

```
cd /public/home/buzh/water/DeepWater/scripts; /public/home/buzh/env/miniconda3/bin/python run_forecast.py -n '白马寺' -m 'LSTM' -i 'tn' -s '2020051108' -e '2020051108' >& /public/home/buzh/water/DeepWater/logs/logs_forecast/task_2020051108.out
```

## utils 整体框架工具模块

I. ConfigParseUtils.py 命令行解析工具 

- II. ConstantUtils.py 常量枚举类工具🔧
- III. FileUtils.py 文本文件交互工具🔧
- IV. LogUtils.py 日志工具🔧
- V. SimulateUtils.py DataSimulation模块专用工具🔧
- VI. ThreadUtils.py 多线程类工具🔧

## host\_file 多节点核心数配置文件

- 利用mpirun运行时于-f指定使用。
- 修改scripts/run.sh中MPI\_HOST路径

## pbatch 批量运行任务文件程序

- pbatch和host\_file的使用均在scripts中run.sh脚本内。
- 修改scripts/run.sh中PBATCH路径

## 运行方式

scripts/run.sh

- I. 解压运行环境miniconda3，其路径修改在run.sh中的export PYTHON="{your\_abspath}/miniconda3/bin/python"
- II. git clone ssh://git@47.92.132.84:2000/buzh/DeepWater.git
- III. git checkout dev分支（目前dev为最新运行版）
- IV. 修改run.sh中export PROJ\_HOME={your\_abspath}/DeepWater
- V. 修改run.sh中export MPIRUN={your\_abspath}
- VI. 修改host\_file中节点及核心数信息，根据节点资源和测试方式选择run.sh中mpirun的方式

config/common.py

- I. 修改proj\_home
- II. 数据库信息SQLite\_dir、PostgreSQL\_info
- III. 预报配置参数：预报指标列表INDICES、预报模型列表MODELS、预报站点列表STATIONS
- IV. 预报算法配置：历史拟合天数last\_days、输入输出序列长度IN&OUT\_STEPS

## 运行效果

```
p
buzh@node1:~  ×  common.py  ×
pid: 2, task_id: 3
pid: 3, task_id: 4
pid: 1, task_id: 0
pid: 2, task_id: 0
pid: 3, task_id: 0
pid: 4, task_id: 0
[buzh@node1 ~]$ bash water/DeepWater/scripts/run.sh 2020010112
task_2020010112 already exist, removing ...
task_2020010112 make done!
task file exist ...
pid: 1, task_id: 1
pid: 2, task_id: 2
pid: 3, task_id: 3
pid: 4, task_id: 4
pid: 2, task_id: 0
pid: 4, task_id: 0
pid: 3, task_id: 0
pid: 1, task_id: 0
[buzh@node1 ~]$ bash water/DeepWater/scripts/run.sh 2020010112
task_2020010112 already exist, removing ...
task_2020010112 make done!
task file exist ...
pid: 3, task_id: 1
pid: 4, task_id: 2
pid: 1, task_id: 3
pid: 2, task_id: 4
pid: 4, task_id: 0
pid: 2, task_id: 0
pid: 3, task_id: 0
pid: 1, task_id: 0
[buzh@node1 ~]$ bash water/DeepWater/scripts/run.sh 2020010112
task_2020010112 already exist, removing ...
task_2020010112 make done!
task file exist ...
pid: 2, task_id: 1
pid: 3, task_id: 2
pid: 4, task_id: 3
pid: 1, task_id: 4
pid: 3, task_id: 0
pid: 4, task_id: 0
pid: 1, task_id: 0
pid: 2, task_id: 0
```

## 4. 结论&探索

现有算法体系已进行了Arima与Fbprophet两类统计算法的探索，其中：Arima算法可以有效的分析给定历史一段时期该指标数据的趋势项、周期项、残差噪声等，有效依据趋势周期特征进一步预测未来一段时间的指标变化情况，其预测效果受给定的历史数据质量和情况影响较大。

Fbprophet时间序列预测框架由Facebook开源，设计之初是为较好预测商品销量，充分的将业务背景知识和统计知识融合起来，它让我们可以用简单直观的参数进行高精度的时间序列预测，并且支持自定义季节和节假日的影响，其中对于自定义影响日的特性，在后期监测站点、污染源数据达到一定精细管理的程度时，其特性可用于模拟极端天气、污染源偷排漏排等对水质指标影响较大的事件，从而有效预测水质指标变化。

然而分析算法优势和局限性时可以发现，Arima和Fbprophet算法是基于纯统计规则，拟合的特征值和预测的结果在同一批历史数据上的结果总是可以复现的，可仿照大气产品设计情景模拟与达标规划等功能产品；并且这两类算法仅支持一种特征指标的模拟和预测，该指标仅考虑其历史变化情况，无法考虑相关的其他指标对其产生的影响。

为更有效的考虑其他多种特征，基于LSTM和GRU元结构的循环神经网络在考虑时间维度变化的同时也利用了同时次其他指标特征，更有效的提高了预报算法的合理性；并且借助大量历史数据构建训练集针对每个站点、每个区域来训练神经网络。为继续丰富算法框架的内容，提高算法预报的准确性，可以从以下几个方便开展调研和算法实验：

### I. 时间序列网络

目前框架内进行了LSTM（Long Short Term Memory）与GRU（Gate Recurrent Unit）两类循环神经网络的实验与业务开发，时间序列的神经网络结构随自然语言处理技术、语音识别技术的发展已经剔除如Attention注意力机制，Seq2Seq结构，其在时间序列上表现会有更好的效果。

## II. 多源数据融合

多源数据包括拓扑结构在上下游关联的自动站点数据、气象多指标观测和预报数据、机理模型模拟数据等。

拓扑结构关联的上下游区域站点，其水质指标存在一定程度的一致性，依据空间关系合理设计监测数据在时间维度的对齐方式，在基于深度学习算法的训练集构建时可以提供更多维度特征，以提高预报效果。

气象数据在区域范围、实时监测、预测等方面均较完备，数据量较多。从特征角度，气象多要素数据为深度学习训练集的构建提供了多维特征，增加了一次、二次特征的构建和组合；从气象预报角度，气象要素的预报能力已覆盖了未来7~15天，其大量预报数据可用于水质相关指标的预报，重点可以关注区域气象要素与区域站点在空间、时间维度的融合方案设计。

机理模型模拟数据，在模拟指定时间段特定流域数据的基础上，利用历史水质与气象预报等数据，可进一步模拟预测区域水质，其数据量比独立站点数据更具连续与相关性，在二次开发的时候可与目标水质进行多种算法实验探索潜在的线性和非线性关系。

## III. 多源预报融合

基于多种统计预报算法、机器学习预报算法、深度学习预报算法、气象预报、机理模拟预报等产生的多源预报数据，依据数据量、数据质量、数据相关性梳理可进一步构建堆叠Stacking或Boosting树模型的集合预报模型，可以有效规避不同预报算法潜在的不足，以提高模型的泛化能力，提高预报准确率。