引文格式：周仿荣,文刚,马仪,等. 遥感数据驱动的电力污秽等级XGBOOST预测模型[J]. 地理空间信息, 2023, 21(12): 21-24. doi:10. 3969/j. issn. 1672-4623.2023.12.004

遥感数据驱动的电力污秽等级XGBOOST预测模型

周仿荣[1]，文　刚[1*]，马　仪[1]，张　辉[1]，朱龙昌[1]，杨可意[2]，韩　舸[2]
（1. 电力遥感技术联合实验室（南方电网公司云南电网电力科学研究院），云南 昆明 650217；
2. 武汉大学遥感信息工程学院，湖北 武汉 430079）

摘　要：输电线路绝缘子积污是电网安全运行的重要威胁之一。传统的电力污秽等级评估依靠人工作业，存在效率低、准确性不足等问题。遥感和人工智能技术的发展为改进电力污秽等级预测模型提供了新的契机。本研究搜集了夜间灯光遥感数据、大气环境遥感监测数据、NDVI和污染源排放清单等多源异构数据，开发了一种基于XGBOOST算法的电力污秽等级预测模型，用于实现大范围、细粒度的电力系统污区图绘制。实验结果表明，该方法在测试集上取得了87%的总体精度，对于电力系统关注的重污染区域预测精度在82%以上。模型所需的数据源可以通过公开渠道免费获得，因此该方法将为今后电力系统低成本、高效、准确的绘制电力污区图提供重要支撑。

关键词：电力污秽；夜光遥感；ESDD；XGBOOST
中图分类号：P237　　**文献标志码：**B　　**文章编号：**1672-4623（2023）12-0021-04

XGBOOST Prediction Model for Power Fouling Level Based on Remote Sensing Data-driven

ZHOU Fangrong[1], WEN Gang[1], MA Yi[1], ZHANG Hui[1], ZHU Longchang[1], YANG Keyi[2], HAN Ge[2]
(1. Joint Laboratory of Power Remote Sensing Technology, Electric Power Research Institute, Yunnan Power Grid Co., Ltd., China Southern Power Grid, Kunming 650217, China; 2. School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China)

Abstract: Transmission line insulator fouling accumulation is one of the important threats to the safe operation of power grids. Traditional power fouling level assessment relies on manual work, which has problems such as low efficiency and accuracy. The development of remote sensing and artificial intelligence technology provides new opportunities to improve the prediction model of power fouling level. During the research, we collected multi-source heterogeneous data such as remote sensing data of nighttime lights, remote sensing monitoring data of atmospheric environment, NDVI and emission inventory of pollution sources, and developed a prediction model of power fouling level based on XGBOOST algorithm for realizing large-scale and fine-grained fouling area mapping of power system. The experimental results show that the method achieves an overall accuracy of 87% on the test set and a prediction accuracy of over 82% for the heavily polluted areas of power system concern. The data sources required by the proposed model in this paper are freely available through public channels, so the method will provide important support for low-cost, efficient and accurate mapping of power pollution areas in power systems in the future.

Key words: power fouling, nighttime light remote sensing, ESDD, XGBOOST

近年来，通过大量实验室和自然环境的实验发现，电力污秽来源于绝缘子所处环境大气中的污染物。宿志一[6]证实大气污染是引起污闪事故的重要原因；胡雱[7]等发现利用$PM_{2.5}$浓度观测可以提高绝缘子等值盐密（ESDD）的建模精度；熊宇[8]等将大气质量指数（AQI）引入ESDD动态累积模型，成功提高模型预测精度；高嵩[9]等定量分析了降雨对污秽积累的冲刷作用。目前在单点尺度上利用大气环境数据和气象数据对电力污秽的积累预测取得较好的进展。但是，利用这些驱动因子进行区域性评估和电力污秽等级制图尚需要进一步探索[8,10]。要形成一种准确的电力污秽等级图绘制手段，还存在较大的差距[11-12]。

针对这一问题，本研究提出一种多源数据融合的电力污秽等级预测模型。在前人研究的基础上进一步引入夜间灯光遥感数据作为人类活动强度的量化指标，同时利用记录污染源信息的文本数据生成空间化的污染源核密度，并采用网格化排放清单表征多种大气污染物的排放量。以云南地区作为研究

收稿日期：2023-06-01。
项目来源：云南省重大科技专项资助项目（202202AD080010）。
第一作者简介：周仿荣（1982—），硕士，正高职高级工程师，研究方向为输变电设备防灾减灾及设备故障诊断，E-mail：4283590@qq.com。
通信作者：文刚（1987—），工程师，研究方向为输电线路防灾减灾，E-mail：1192381484@qq.com。

对象，利用形成的高维输入属性集合，以现有污区图为目标，采取XGBOOST进行建模以完成污秽等级空间预测。

1 研究区与数据源

1.1 研究区概况

云南省位于我国西南部，大气环境质量在全国属于上游水平[13]，因此云南省与现有电力污秽累积研究有很大的区别。云南省2018—2020年大气$PM_{2.5}$数据显示，其全省年最大$PM_{2.5}$质量浓度仅为53.2 μg/m^3。由此可见$PM_{2.5}$或AOD不可能是该地区电力污秽的主要驱动因子，这为建立预测模型带来了巨大挑战。但也正是由于研究区的特殊性，在本地区能够适用的建模方法具有更强的移植能力。

1.2 实验数据

本文搜集到的多源数据信息主要分为四大类：遥感数据（包括夜间灯光数据[14]、归一化植被指数数据、大气环境数据）污染排放企业位置文本数据、气象数据、污染物网格化排放清单。本研究目标是建立面域预测模型，需要对原始数据进行预处理。将具有空间属性的卫星遥感数据和网格化排放清单统一到1 km的分辨率和相同的坐标系，以获得一致的空间数据。气象数据采用克里金插值转换为1 km分辨率的栅格型空间数据。污染企业信息将文本类型属性转换为点状矢量数据，再利用核密度分析和距离分析转换为1 km分辨率的栅格数据，并使用排放强度作为权重。

2 流程与方法

多源环境数据融合的电力污秽评估模型流程为：①搜集多源异构数据。为了完成污秽的空间评估，要求所有建模数据是时空数据。搜集到的数据中表征排放源信息的数据是文本类型，主要包括经纬度、排放强度分级和影响范围分级数据。本研究采用核密度分析和距离分析的空间分析手段将文本数据转换为空间数据。②对所有数据进行时空配准并进行数据清洗并剔除异常值。③使用2020版云南省污区图作为标签图层，利用XGBOOST建立输入数据到标签的映射关系，并采用十折交叉验证法训练和验证模型。④为了消除重采样排放清单数据以及某一变量重要性较高引起的锯齿效应，我们将采用引导滤波进行平滑处理，以获得云南全省污染等级的最终空间预测结果。

XGBOOST是一种由GBDT(gradient boosting decision tree)算法发展而来的机器学习算法[15]，其在集成学习的基础上，结合梯度信息，完成了对目标函数的优化，获得最优解。相比于GBDT，XGBOOST除了运用了损失函数的一阶导数信息外，还通过对损失函数的泰勒展开，获取损失函数二阶导数信息，更快获得最优解[16]。XGBOOST算法由一系列决策树组合而成：

$$\hat{y}_i=\sum_{n=1}^{N}f_n(\boldsymbol{x}_i),f_n\in FS \tag{1}$$

式中，FS（forest sets）为决策树集合；$\boldsymbol{x}_i$为第i条数据的特征值所组成的向量；$f_n(x_i)$为第n个独立决策树，其中包含树的结构和权重信息；N为决策树的总量；$\hat{y}_i$为第i条数据的预测值。

XGBOOST定义了损失函数Loss，通过训练集提供的数据训练，可以获取决策树的相关信息。

$$\text{Loss}=\sum_{i=1}^{M}L(y_i,\hat{y}_i)+\sum_{n=1}^{N}\Omega(f_n) \tag{2}$$

式中，$L(y_i,\hat{y}_i)$为预测值$\hat{y}_i$和真实值y_i间的损失函数，根据任务需求不同，选取的损失函数种类不同，本研究中选取multi−softmax作为预测值$\hat{y}_i$和真实值y_i间的损失函数。M为训练集数量，$\Omega(f_n)$为决策树的正则项，防止树结构过于复杂，产生过拟合现象。

XGBOOST通过多轮迭代获取最优解，其中第t轮的损失函数可表示为公式（3）。

$$\text{Loss}^{(t)}=\sum_{i=1}^{M}L(y_i,\hat{y}_i)+\sum_{n=1}^{N}\Omega(f_n) \tag{3}$$

为了方便后续求导过程，将公式（4）代入公式（3）可获得t轮的损失函数的简便表示。

$$f_t(x_i)=\sum_{n=1}^{N}f_n^t(x_i) \tag{4}$$

$$\text{Loss}^t=\sum_{i=1}^{M}L(y_i,f_t(x_i))+\Omega(f_t) \tag{5}$$

为求取第t轮损失函数中叶子权重的最优解，对第t轮迭代的损失函数Loss^t进行二阶泰勒展开可以得到公式（6）：

$$\begin{aligned}\text{Loss}^{(t)}\cong\sum_{i=1}^{M}[g_i\triangle f_t(x_i)+\frac{1}{2}h_i\triangle f_t^2(x_i)]+\Omega(f_t)\cong\sum_{i=1}^{M}[g_if_t(x_i)+\\ \frac{1}{2}h_if_t^2(x_i)]+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2\cong\sum_{j=1}^{T}\left[\left(\sum_{i\in I_j}g_i\right)w_j+\right.\\ \left.\frac{1}{2}\left(\sum_{i\in I_j}h_i+\lambda\right)w_j^2\right]+\gamma T\end{aligned} \tag{6}$$

式中，$\triangle f_t(x_i)$为第t轮预测值与第$t-1$轮迭代的增量；g_i和h_i分别为$L(y_i,\hat{y}_i)$的一阶导数和二阶导数；$I_j=\{i|q(x_i=j)\}$为第j颗树所有叶子的权重；$\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2$为$\Omega(f_t)$的代数形式。

在公式（6）的基础上对w求偏导，可以得到第j

棵树的叶子权重在第 t 轮的迭代情况下的最优数值 $\widehat{w_j}$：

$$\widehat{w_j}=-\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i+\lambda} \tag{7}$$

通过设定迭代次数和决策树的结构信息，我们可以获得在给定数据集情况下训练好的XGBOOST模型 $\sum_{n=1}^{N} f_n(x_i), f_n\in FS$。

引导滤波（guided filtering）和双边滤波（BF）、最小二乘滤波（WLS）是三大边缘保持（edge-perserving）滤波器[17]。他们在保持边缘的基础上，对图像进行了平滑操作。

引导滤波定义了在给定引导图像 I 和原始图像 p 的条件下，输出图像 q 可以表示为公式（8），其中 w_k 为滤波核大小，n_i 为噪声。

$$\mathrm{q}_i=a_k\mathrm{I}_i+b_k, \forall i\in w_k \tag{8}$$

$$\mathrm{q}_i=\mathrm{p}_i-n_i \tag{9}$$

通过求解代价函数 $E(a_k,b_k)$，可获得 a_k 和 b_k，其中 ϵa_k^2 为正则约束项。

$$E(a_k,b_k)=\sum_{i\in\omega_k}\left(\left(a_k\mathrm{I}_i+b_k-\mathrm{p}_i\right)^2+\epsilon a_k^2\right) \tag{10}$$

分别对 a_k 和 b_k 进行求导，可以获得给定滤波窗口 w_k 范围内的最优估计值 $\widehat{a_k}$、$\widehat{b_k}$。

$$\widehat{a_k}=\frac{\frac{1}{|\omega|}\sum_{i\in\omega_k}\mathrm{I}_i\mathrm{p}_i-\mu_k\bar{\mathrm{p}}_k}{\sigma_k^2+\epsilon} \tag{11}$$

$$\widehat{b_k}=\bar{\mathrm{p}}_k-a_k\mu_k \tag{12}$$

3　结果分析

本文使用十折交叉验证法对构建完成的数据集进行训练，利用XGBOOST模型在测试集上进行验证，最终取得了87%的精度。图1~3展示了电力系统现行的污区图、基于多源遥感数据的XGBOOST模型直接输出结果以及采用引导滤波后的结果。

对比图1~3，发现对于电力系统最为关心的高风险区域（1~3），图2与图1非常接近，很好地还原了由工业排放引起的局部高风险区域。与熊宇等的结果相比，本研究结果对于重污染地区（1~2）的预测表现更为优秀。这是由于本研究利用夜间灯光遥感产品更准确地刻画了人为排放的分布情况，同时XGBOOST算法的性能比支持向量机等传统分类器更为优越。

图3表明引入引导滤波可以显著的抑制在局部空间上，由于排放清单的低空间分辨率导致的粗糙不平滑的边缘。可以明显观察到，通过引导滤波卷积后的预测结果在空间分布上与污秽等级实测图更接近，边缘也更加平滑，更符合污秽等级分布的真实情况。

为了更好地定量评价本文所提方法的预测精度，表1展示了最终预测结果与现行污区的混淆矩阵。从电网安全的角度看，较为严重污染等级（1~3）的地区更受重视，这些地区往往需要更为频繁的清污工作。但是，从图1可以看出，1~3等级的地区在面积上的占比不大。如果以总体精度为考察对象，最极端的模型可以通过将全部地区划分为等级4来取得0.8以上的精度，这显然与应用初衷不相符。表1显示，本文所提模型对于每一类的预测精度都达到0.8以上，并没有通过牺牲1~3类别预测精度来实现总体的高精度。这是本方法一个重要的特征和明显的优势。这表明本方法得出的结果不仅具有数学意义上的有效性，更重要的是对于指导电力系统的实际工作具有巨大的价值。

表1　模型预测性能评价混淆矩阵

数值		预测值			
		1	2	3	4
真实值	1	0.868	0.104	0.025	0.003
	2	0.020	0.847	0.109	0.025
	3	0.005	0.113	0.824	0.058
	4	0.001	0.024	0.070	0.906

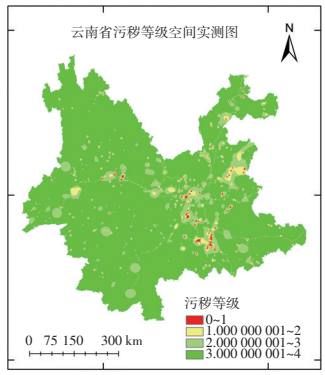


图1　云南省2021年电力系统污区图
（审图号：GS(2019)1822号）

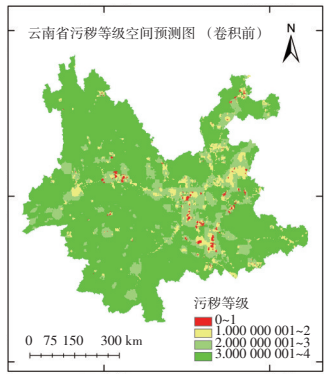


图2　利用多源环境数据和XGBOOST算法得到的云南省电力系统污区图
（审图号：GS(2019)1822号）

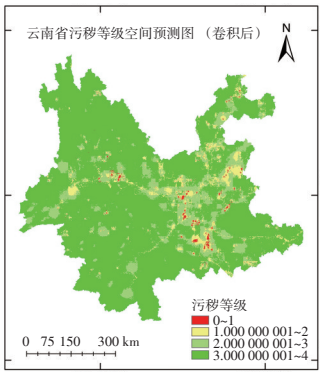


图3 经过引导滤波后的云南省电力系统污区图
（审图号：GS(2019)1822号）

4 结果与讨论

本文预测精度优于传统污秽等级预测方法，我们认为这与引入了夜间灯光数据相关。夜间灯光数据很好地表征了与污染程度存在紧密相关性的人类活动的强度，使得高分辨率的夜间灯光数据在很大程度上弥补了排放数据低分辨率的缺陷，从另一个角度表征了预测点的环境污染情况，进而使得预测精度获得了提升。

决策树中，子树的分裂往往决定了最终整体树的分类效果。特征参与子树分裂的次数越多，该特征在分类中的作用越大。在上述预测精度的情况下，图4按照数据集的统计顺序给出了每种属性数据参与子树分裂的次数占总次数的频率。从图4可见，污染气体排放数据如OC、NOX、SO_2、VOC以及NO_2和SO_2的遥感数据在预测中仍然起了较大作用，几种排放数据的累计重要性占比达到了51.9%，超过一半。对于$PM_{2.5}$和PM_{10}，由于2种数据在统计性质上高度相关，它们的重要性占比相近。此外，夜间灯光属性占比达10.4%，证明了夜间灯光数据在实际分类过程中起到了很大作用，提高了预测精度。实验结果表明，提高污染气体遥感分辨率和精度，引入夜间灯光数据，有助于提高污秽监测的预测精度。

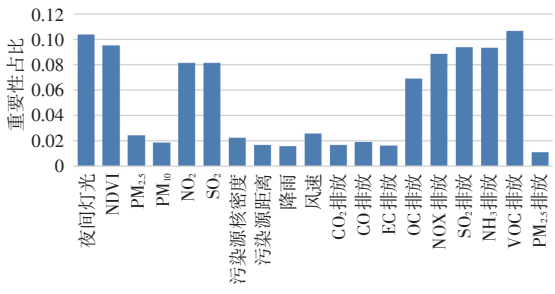


图4 输入参数在模型中的重要性排名

5 结 论

本文借助机器学习中XGBOOST算法，利用包括大气环境、气象、夜间灯光遥感数据、污染源核密度，网格化排放清单表在内的多源数据，作为驱动因子对云南省进行污秽等级预测，并与实测污秽等级分布进行比较。实验结果表明，本文所提方法在污秽等级预测上与现有污区图吻合率达到87%，单一种类的预测精度均达到80%以上，表明了多源数据融合的电力污秽等级XGBOOST预测模型在污秽等级预测上具有良好的潜力。

参考文献

[1] 宿志一. 雾霾天气对输变电设备外绝缘的影响[J]. 电网技术，2013,37(8):2 284-2 290

[2] 胡雾,万丽芬,陈怡,等. $PM_{2.5}$对理论盐密值的影响研究[J]. 高压电器,2016,52(5):60-64

[3] 熊宇,陈孝明,阮羚,等. 卫星AOD数据驱动的区域电力污秽等级评估方法[J]. 测绘通报,2016(8):39-43

[4] 高嵩,刘洋,曹彬,等. 自然降雨对绝缘子自然污秽冲洗效果的影响研究[J]. 电瓷避雷器,2020(6):6-9

[5] Ruan L,Han G,Zhu Z,et al. Spatial and Temporal Characteristics of Insulator Contaminations Revealed by Daily Observations of Equivalent Salt Deposit Density[J]. Sensors,2015,15(2):3 023-3 040

[6] Stefenon S F,Grebogi R B,Freire R Z,et al. Optimized Ensemble Extreme Learning Machine for Classification of Electrical Insulators Conditions[J]. IEEE Transactions on Industrial Electronics,2019,67(6):5 170-5 178

[7] Salem A A,Abd-rahman R,Al-gailani S A,et al. Risk Assessment of Polluted Glass Insulator Using Leakage Current Index Under Different Operating Conditions[J]. Ieee Access,2020,8:175 827-175 839

[8] Pei Z,Han G,Ma X,et al. Response of Major Air Pollutants to COVID-19 lockdowns in China[J]. Science of the Total Environment,2020,743:140 879

[9] Oda T,Maksyutov S. A Very High-resolution (1 km× 1 km) Global Fossil Fuel CO_2 Emission Inventory Derived Using a Point Source Database and Satellite Observations of Nighttime Lights[J]. Atmospheric Chemistry and Physics,2011,11(2):543-556

[10] Wei J,Savtchenko A,Vollmer B,et al. Advances in CO_2 Observations from AIRS and ACOS[J]. IEEE Geoscience and Remote Sensing Letters,2013,11(5):891-895

[11] 叶景,李丽娟,唐臻旭. 基于CNN-XGBoost的短时交通流预测[J]. 计算机工程与设计,2020,41(4):1 081-1 086

[12] 刘丽霞,李宝文,王阳萍,等. 改进Canny边缘检测的遥感影像分割[J]. 计算机工程与应用,2019,55(12):54-58