海上风场SCADA数据缺失智能修复

展示团队： 三蹦子

问题背景

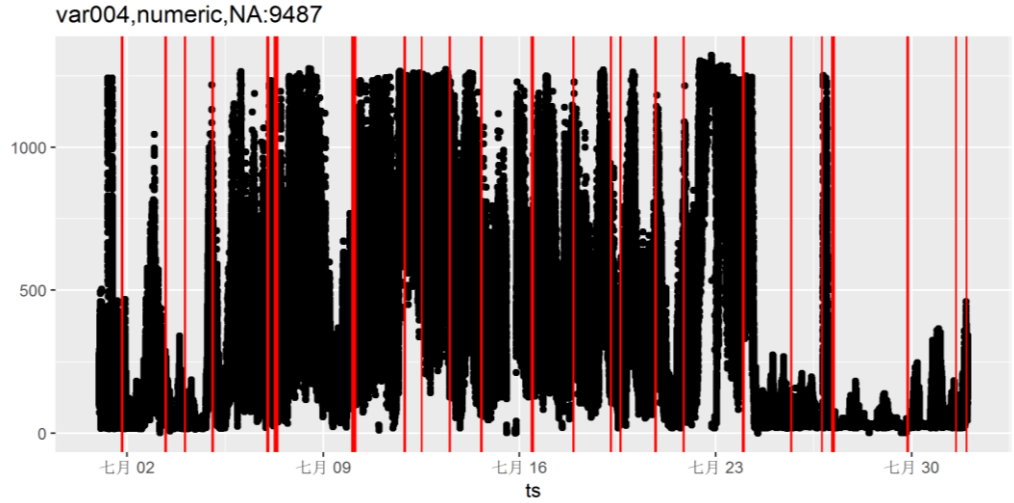
数据缺失

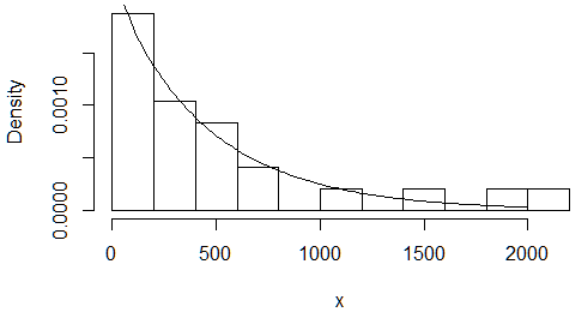
远程数据监控系统

海上风电场

SCADA

缺失数据分布规律与验证集构建

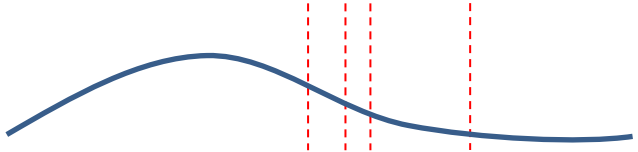




- 33个机组，68个变量（时间序列）
- 包含24个时间区间的数据簇连续缺失
- 缺失时间区间长度近似服从伽马分布

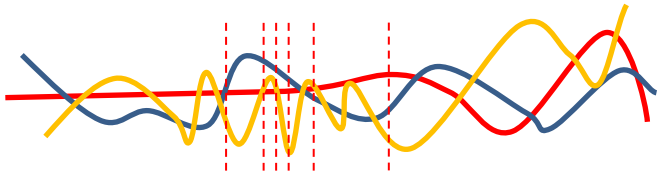
验证集策略1：依分布构建

- 单维度序列缺失值填充模型

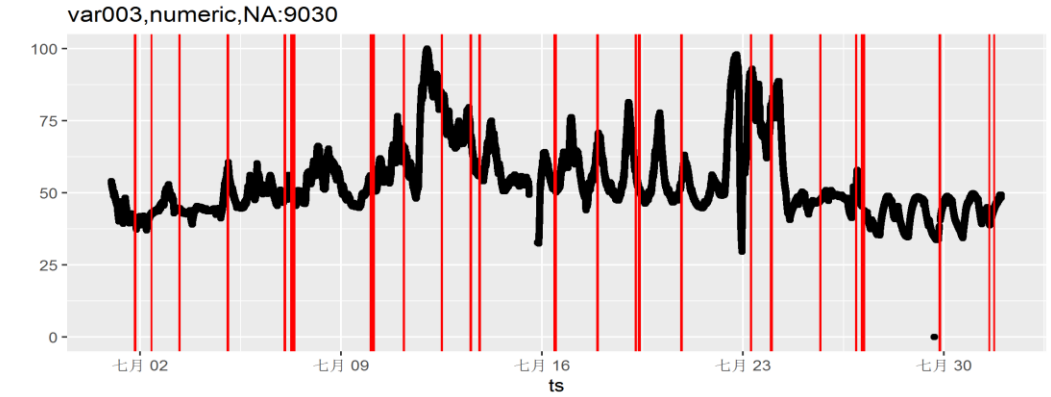


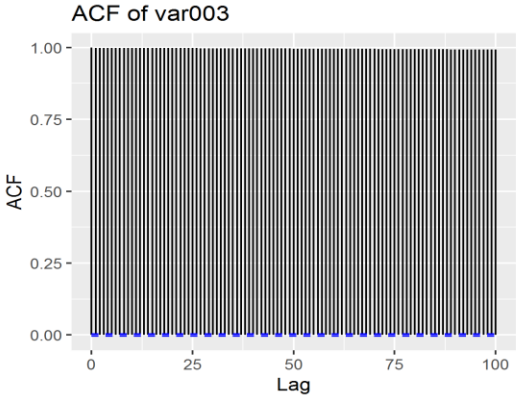
验证集策略2：依时间邻近构建

- 多维度序列缺失值填充模型

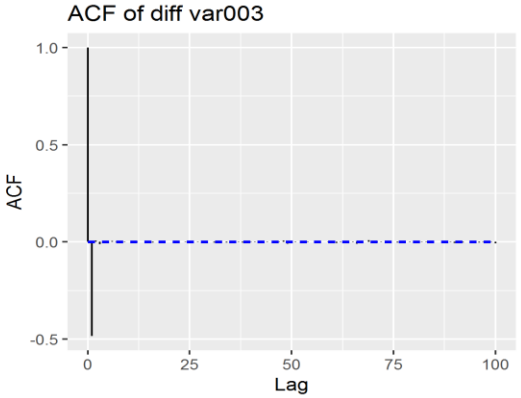


单维时间序列模型（加权平均估计）





非平稳

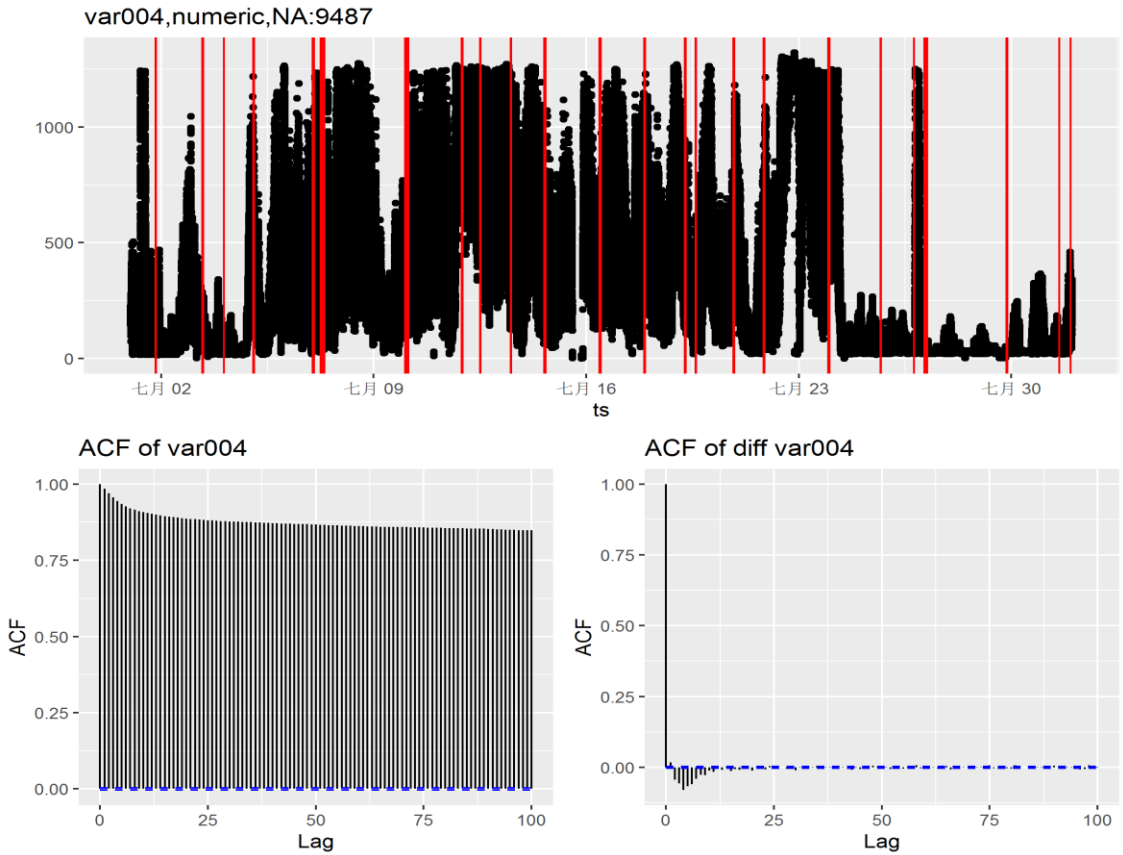


差分后自相关性不显著

$$\hat{x}_{i,j} = \frac{1}{K}\sum_{k=1}^{K} \tilde{w}(I(k,i,j),i)x_{I(k,i,j),j}$$

$$w(I(k,i,j),i) = \frac{1}{|I(k,i,j) - i + 1|}$$

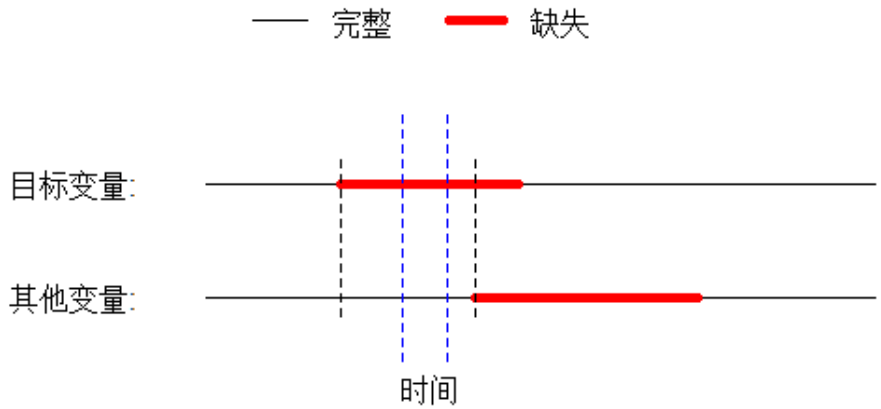
单维时间序列模型（加权最邻近估计）



当变量波动性大，缺失数据由其条件分布的众数来估计，即加权最邻近法（weighted KNN）

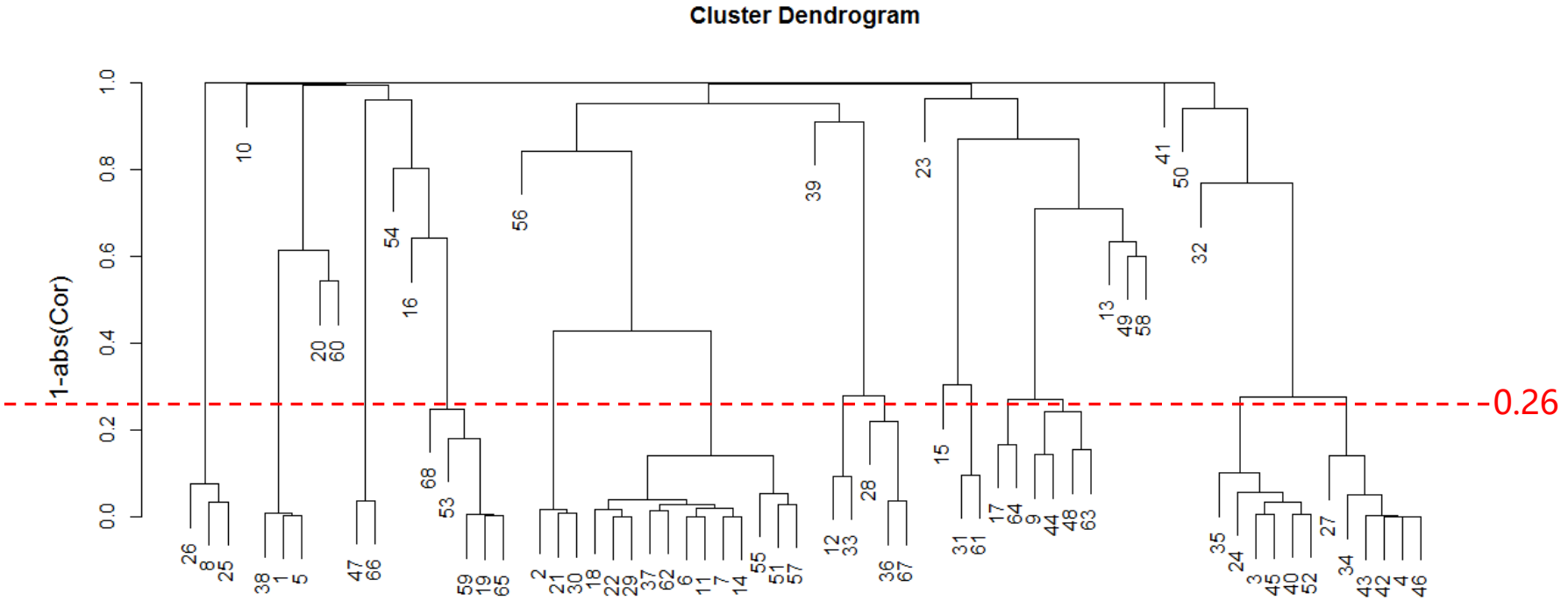
多维时间序列模型

- 单维序列模型预测在数据缺失簇两端部分（更靠近未缺失数据）效果更好
- 在多维序列模型里，我们利用同一时间点，多条序列之间的相关性来弥补这一缺点
- 我们采用了三种回归模型
  - 基于该题评价分数的线性回归模型(LM)
  - 基于最小二乘法的线性回归模型(OLS)
  - 广义相加模型(GAM)



变量聚类分析

- 用$1-|cor(x,y)|$度量变量x与变量y之间的距离，做阶层式聚类



- 用相关性绝对值0.74作为阈值，68个变量可以被归类为28个聚类。

多维与单维模型融合算法

聚类**4**,27,34,42,43,46 → 机组1变量**4** → 合并机组，保留聚类

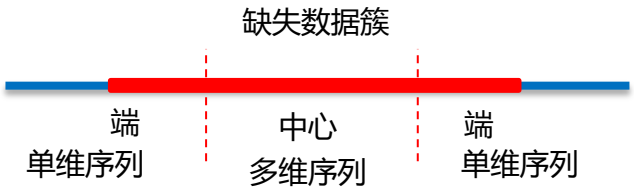
合并机组，保留聚类 → 策略2生成验证集

策略2生成验证集 → 单维序列模型在验证集i上单点分数$\hat{F}_h, h = 1, \dots, H_i$

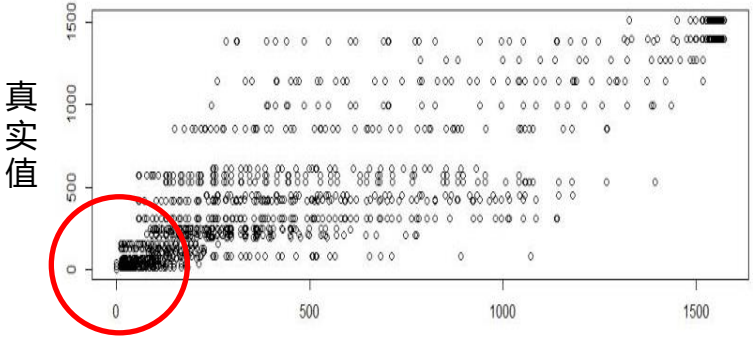
策略2生成验证集 → 多维回归模型在验证集i上单点分数$\tilde{F}_h, h = 1, \dots, H_i$

单维序列模型在验证集i上单点分数$\hat{F}_h, h = 1, \dots, H_i$ → 若$\hat{F}_h - \tilde{F}_h > 2 \times \mathrm{std}(\hat{F}_h - \tilde{F}_h) + \mathrm{c}$，则用多维回归模型估计替换

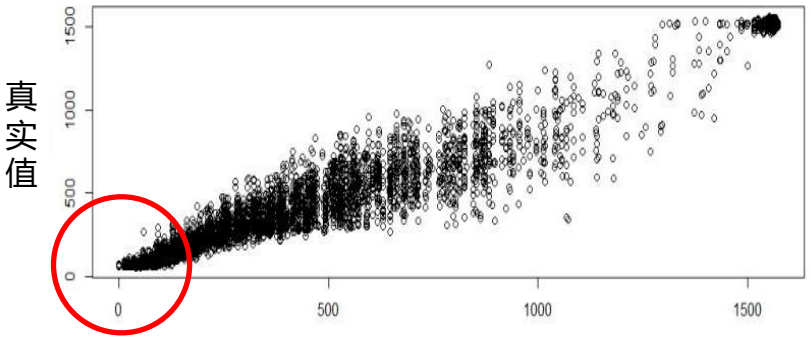
多维回归模型在验证集i上单点分数$\tilde{F}_h, h = 1, \dots, H_i$ → 若$\hat{F}_h - \tilde{F}_h > 2 \times \mathrm{std}(\hat{F}_h - \tilde{F}_h) + \mathrm{c}$，则用多维回归模型估计替换



评分标准影响与控制



单维序列模型估计值



多维序列模型估计值

方法	单序列	多序列
均开方误差RMSE	183	**90**
比赛评分F	**0.06285**	0.0285

方案潜力

创新性地提出局部线性模型的集成方法(Ensemble method)：

- 模型精度高
- 可解释性：模型非黑箱，变量间与机组间关系可解释
- 易拓展性：在多维序列模型中可拓展为分布滞后时间序列模型
- 可并行：聚类后缺失数据的估计过程互相独立，可并行计算

致谢

感谢2019数字中国创新大赛组委会组织此次比赛！

感谢金风科技股份有限公司提供的宝贵数据！

感谢三蹦子团队全员坚持不懈的努力！

谢谢观看，请您提问