第十届“泰迪杯”挑战赛
—— B题作品评析

北京工业大学 薛毅

2022年7月23日

1. 题目

根据电力系统负荷的历史数据和当地气象资料的历史数据，对未来一段时间的电力系统负荷做出预测。预测分为短期（10天）和中期（3个月）预测两种。

问题1：根据附件中提供的某地区电网间隔15分钟的负荷数据，建立中短期负荷预测模型：

(1)给出该地区电网未来10天间隔15分钟的负荷预测结果，并分析其预测精度；

(2)给出该地区电网未来3个月日负荷的最大值和最小值预测结果，以及相应达到负荷最大值和最小值的时间，并分析其预测精度。

问题2：对不同行业的用电负荷进行中期预测分析，能够为电网运营与调度决策提供重要依据。通过对大工业、非普工业、普通工业和商业等行业的用电负荷进行预测，有助于掌握各行业的生产和经营状况、复工复产和后续发展走势。请建立数学模型研究下面问题：

(1)挖掘分析各行业用电负荷突变的时间、量级和可能的原因；

(2)给出该地区各行业未来3个月日负荷最大值和最小值的预测结果，并对其预测精度做出分析；

(3)根据各行业的实际情况，研究国家“双碳”目标对各行业未来用电负荷可能产生的影响，并对相关行业提出有针对性的建议。

2. 数据预处理

- 重复数据

在气象数据中出现重复数据。

- 缺失数据与补充

在附件2中，大工业、非普工业、普通工业和商业用电缺少2021-1-26的数据。

在附件1中，共缺失388条的数据，其特点是，单个数据，如3:00、9:00、10:45等，和连续一段时间的数据，如3:00至24:00。

缺失数据的补充可使用插值方法。

3. 数据分析

- 气象数据

  可以用插值的方法近似给出一天中每一时刻的气温，如图1所示。

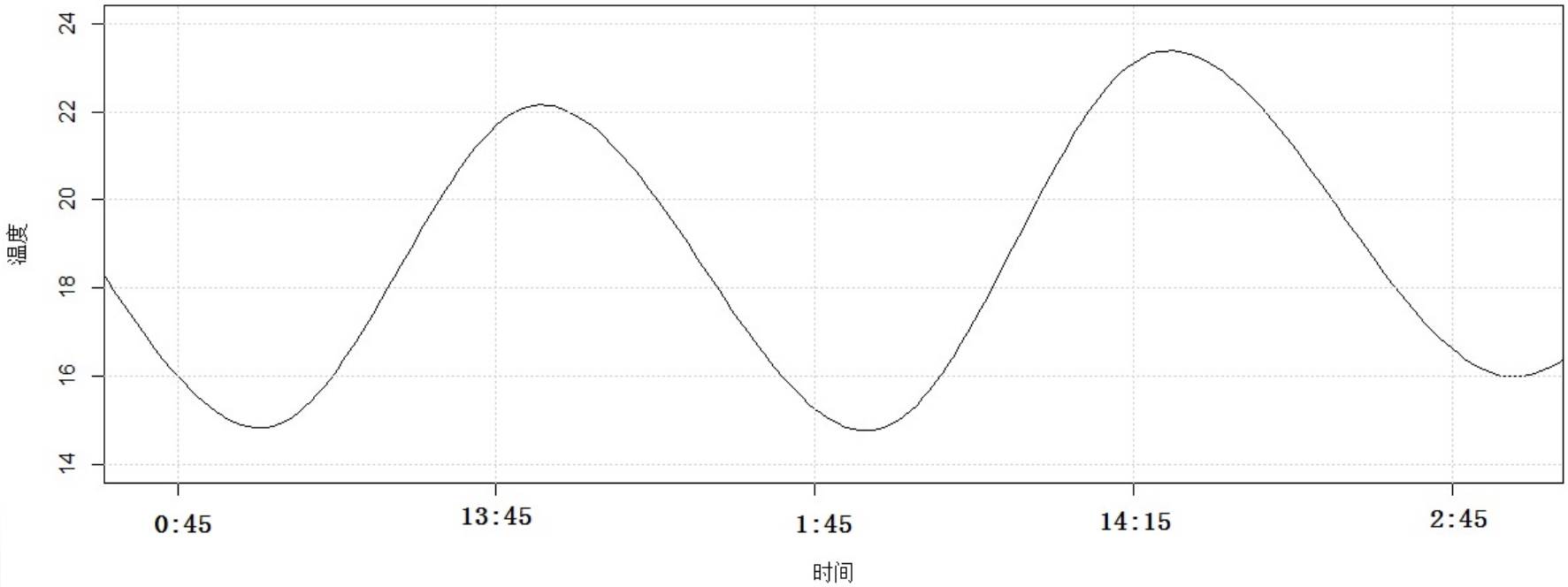

图1 利用最高气温和最低气温近似给出任意时间的气温

- 附件1数据的周期性分析

  每天的用电量会呈现周期性变化，如图2所示。

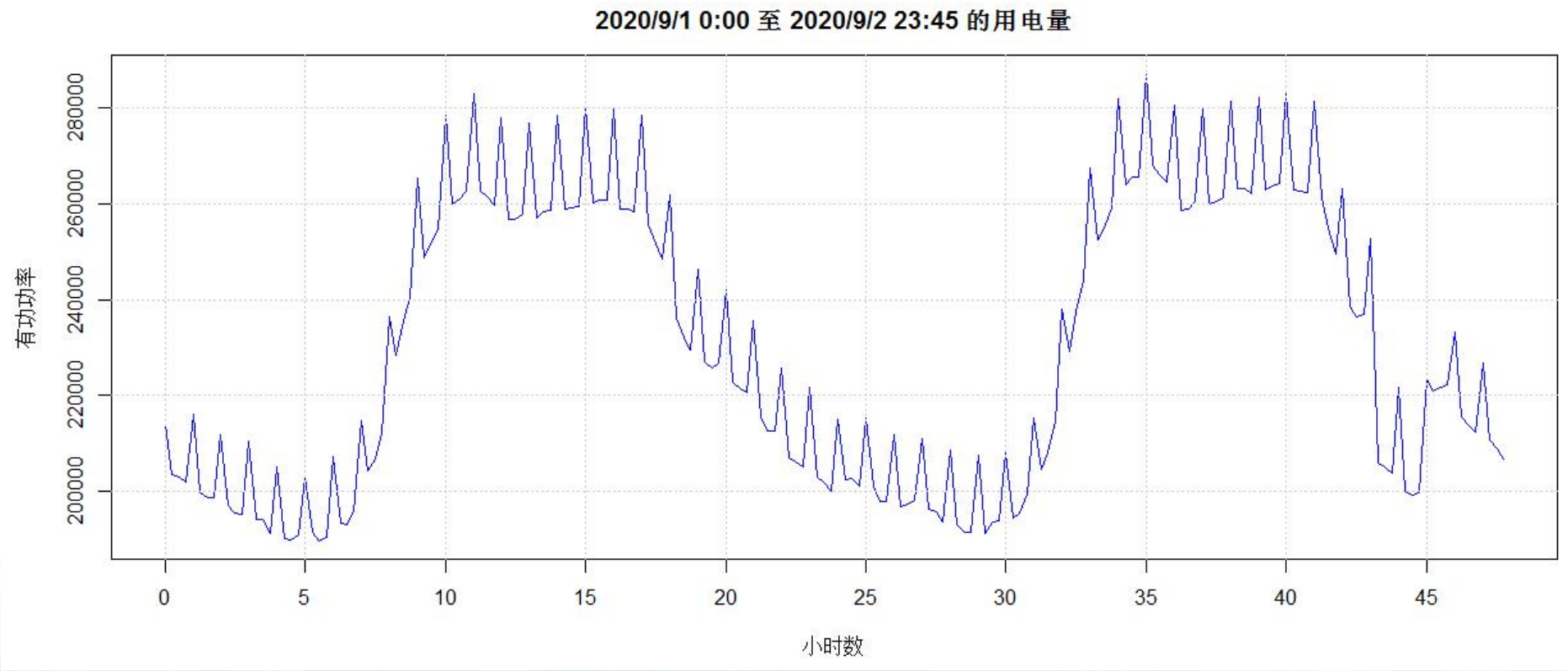

图2 2020年9月1日0时至9月2日24时48小时的用电情况

- 附件1数据的周期性分析

  呈现7天为一个周期的变化，如图3所示。

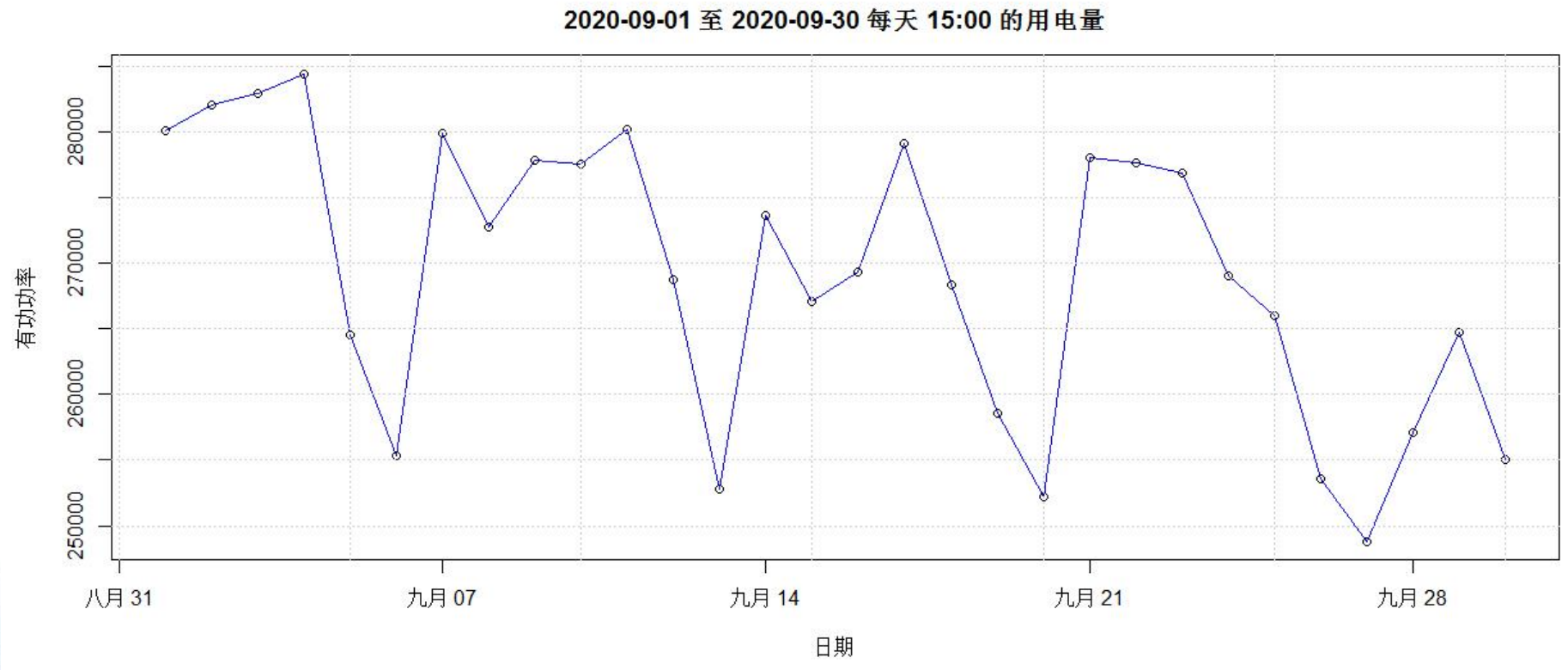

图3 用电量呈现7天一个周期的变化趋势

关于年是有周期性的，但有些是与阴历有关（如春节、端午、中秋），表1列出各年度节日的日期。图4给出各年度用电的变化情况。

表1 各年度节日对应的日期

年度	春节	元宵节	清明节	端午节	中秋节
2018	2月16日	3月2日	4月5日	6月18日	9月24日
2019	2月5日	2月19日	4月5日	6月7日	9月13日
2020	1月25日	2月8日	4月4日	6月29日	10月1日
2021	2月12日	2月26日	4月4日	6月14日	9月21日

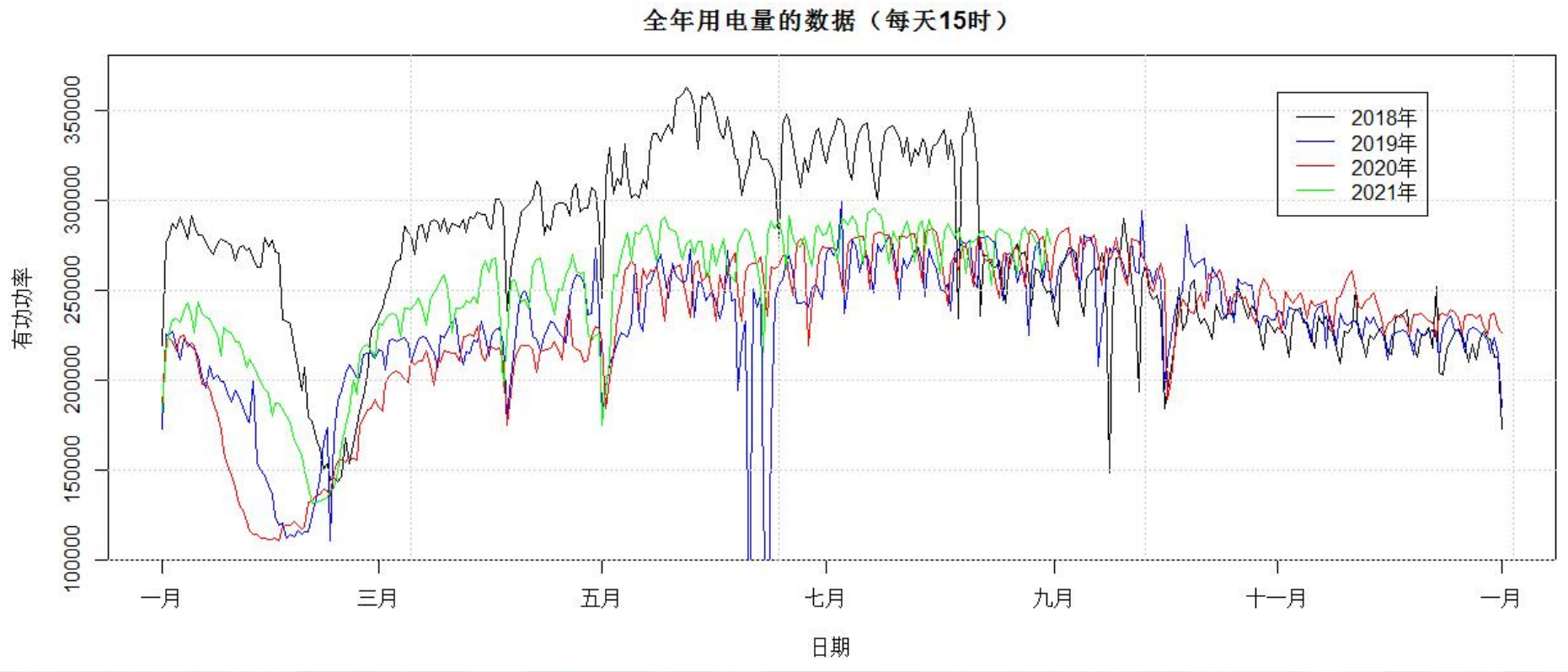

图4 用电量呈现以年为周期的变化趋势

- 相关性检验

这里主要分析用电量与气象数据的关系。例如，考虑最高气温与用电的关系，图5展示2020年9月1日至9月30日（15时）用电量与最高温度的曲线。

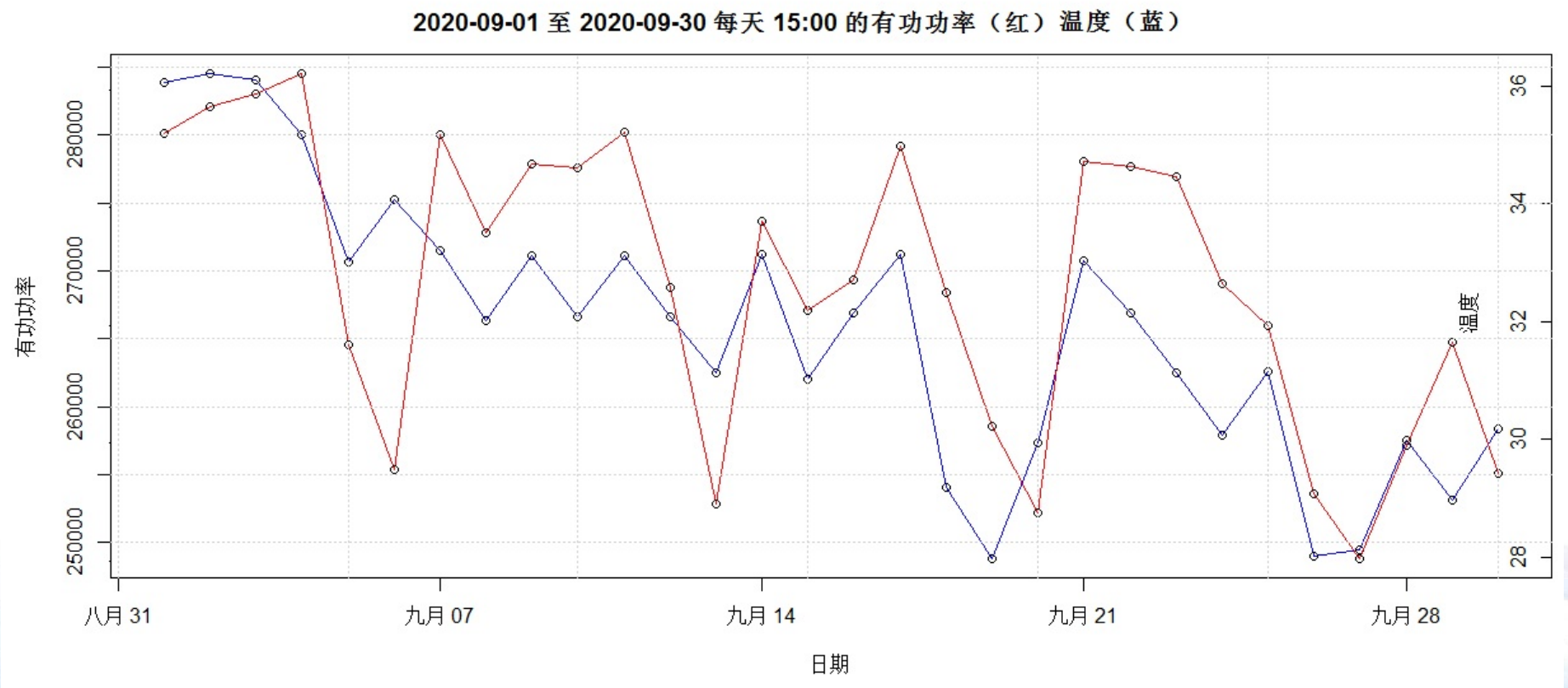

图5 每天15时的用电量与最高温度的曲线图

4. 数据的异常值检验与数据的修正

- 异常值检验

异常值检验的方法有很多，大多数同学使用的是，Z得分法（作标准化变换）和箱线图法。图6给出每天10:00用电量的直方图和箱线图。

对于Z得分法，如果数据呈现正态分布，可用3 Sigma法则，如果数据不呈现正态分布，更保险的方法应该用切比雪夫法则。

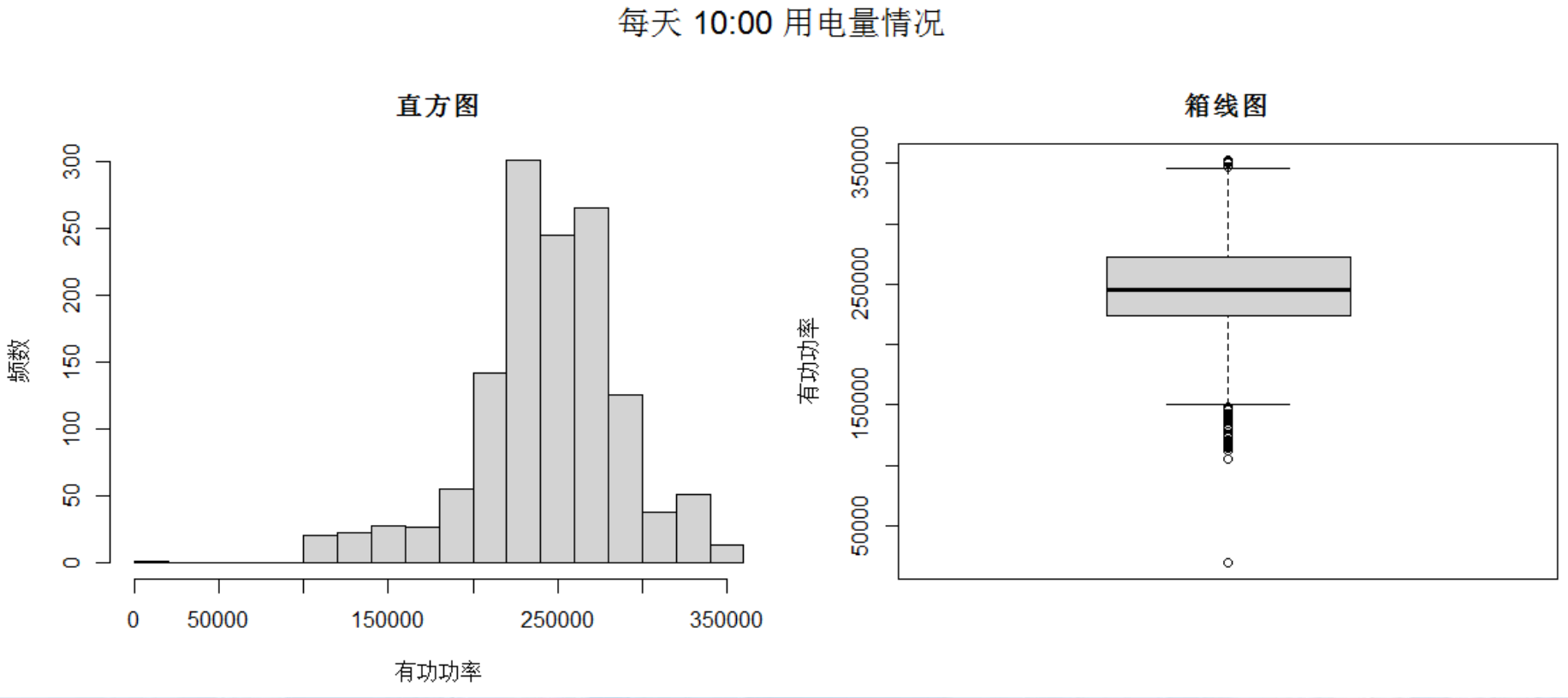

图6 每天10时用电数据的直方图和箱线图

- 数据修正

对异常值点数据的修正要看修正的目的。对于正常情况预测为目的，可以对所有检测出来的异常值点作修正。如果要保留节假日的特点，则需要分开处理，否则会淹没节假日的特征。

较好的方法可能是数据平滑，它即去掉了数据的“毛刺”，同时还保留了节假日的特征。例如，图7给出了实际用电数据与Savitzky-Golay卷积平滑后数据的比较。

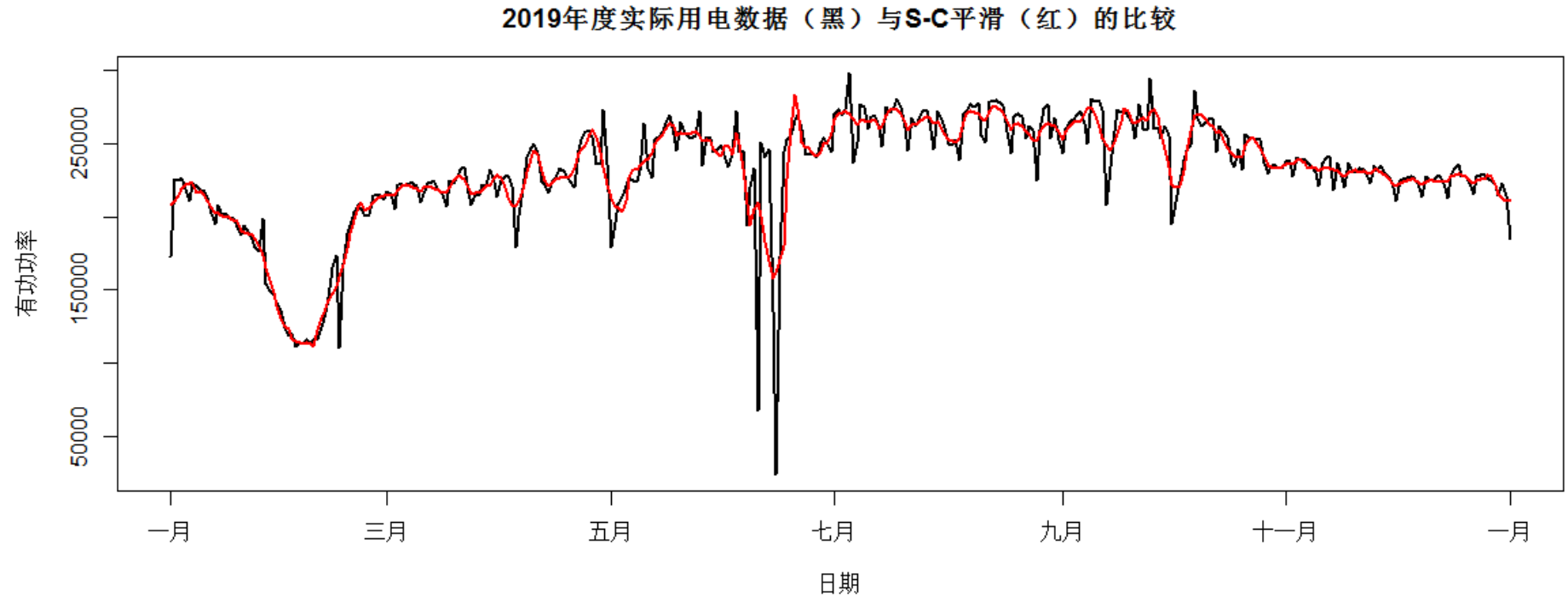

图7 实际用电数据与S-C平滑方法数据的比较

5. 数据预测模型

即不提倡使用“高大上”的模型，也不拒绝“接地气”的模型，无论什么模型，实用即好。

在考虑使用的模型时，应与前面的分析相结合，如周期、季节、节假日、趋势等。

- LSTM (Long Short - Term Memory)模型，长短期记忆模型
- Prophet模型
- Holt-Winters模型
- ARIMA模型

6. 模型检验

模型检验是评价模型的重要方法，通常的作法是k折交叉验证。但这里是时间序列，不能用交叉验证。可以一部分时间作为训练，另一部分时间作为测试。

实际上，这个时间段的预测还相对比较容易，9月1日到10日，没有节日。到11月30日，也只有中秋节和国庆节，所以远期预测会稍差一些。如果对节假日处理的较好，预测的精度会提高。

较为简单的方法，使用2020年的数据，预测2020年9月1日至10日，9月1日至11月30日的情况，它与2021年同期数据特征应该较为相似。

7. 问题2

- 突变

（1） Mann-Kendall突变检测

（2） 滑动t-检验

实际上，画图可能是检测突变即简单，又方便的方法。

突变的原因，很可能是电力公司利用假期检修电力造成的，也许是某些突发事件。

总之，它是无法使用统计方法预测的，图8给出商业用电数据的情况，有些用电量几乎是零。

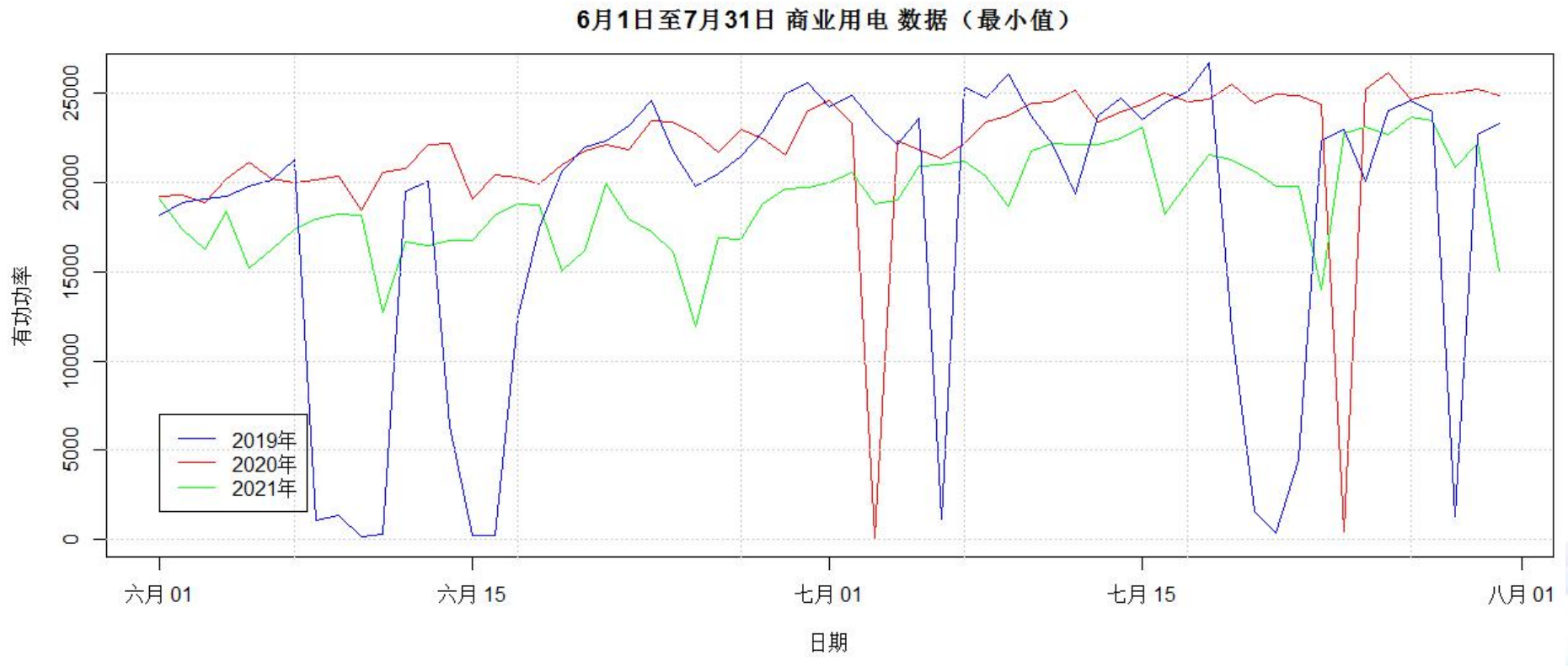

图8 六、七月份商业最低用电量的情况

- 预测

应该不用重新设计模型或算法。用电量以天为单位，少了以一天为周期的特征。

不同行业应该自己的特征。例如，

大工业用电（最大值）与气温（最高气温）呈弱相关，而与趋势（与时间有关）呈现强相关性。

而商业用电与气温呈现出强相关性，而与趋势则呈现弱相关性。

因此，在大的模型不变的基础上，增加这些小的特征，将会提高对四个行业用电量预测的精确度。

- 针对性的建议

根据各行业的实际情况，研究国家“双碳”目标对各行业未来用电负荷可能产生的影响，并对相关行业提出有针对性的建议。

或许数据不足，只能泛泛地谈一谈。

8. 数据的测试情况

将大家的预测值与实际值作了对比，计算出实际的误差的MAPE（平均绝对百分比误差(Mean Absolute Percentage Error)）。该项指标也纳入我们的评价体系。

这里可以告诉大家，凡是相似度较高的论文，这项指标基本上都很差，只能说明，论文或算法不是他们自己完成的。

9. 答辩

答辩是本轮竞赛最后一个环节。

答辩要两个目的，一是晋级，二是保级。

这里想告诉大家，答辩讲什么？要讲自己的工作、工作的创新点，以及相关的结论。

至于模型（或算法）只需要讲方法的适应性，也就是讲为什么要用文中介绍的模型解决问题，而不是别的方法。

不必讲算法本身，除非算法是你们自己创造的方法。

10. 雷同论文

按道理说，我们这类的竞赛是不应该出现雷同论文的。参加竞赛的目的不应该只是为了获奖，获奖只是人生长河中一个小小的闪光点。

学生参加竞赛的目的是学习数据挖掘、机器学习等与数据处理有关的方法，为以后的工作与学习做准备。

这类竞赛应完全凭兴趣参加，通过竞赛学习数据处理的方法，也为大数时代奠定基础。

如果认定为雷同论文，论文成绩为零分，取消竞赛成绩。

为什么还要讲雷同论文？（网上有人在贩卖论文）

向所有获奖学生表示祝贺！

谢 谢 ！