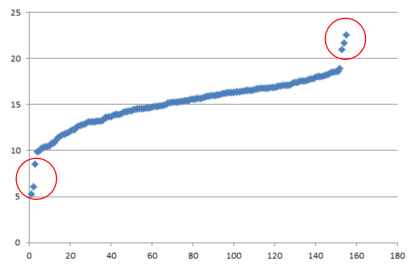
**研究缺陷处理自我备案文档**

本人在完成此论文（Entropy Weight Ensemble Framework for Yield Prediction of Winter Wheat Under Different Water Stress Treatments Using Unmanned Aerial Vehicle-Based Multispectral and Thermal Data, 2021）时，由于学术水平不足，且不够严谨和认真，故出现了很多纰漏，但并未对研究的核心结论造成明显影响，在此进行备案说明。

**一. 核心声明**

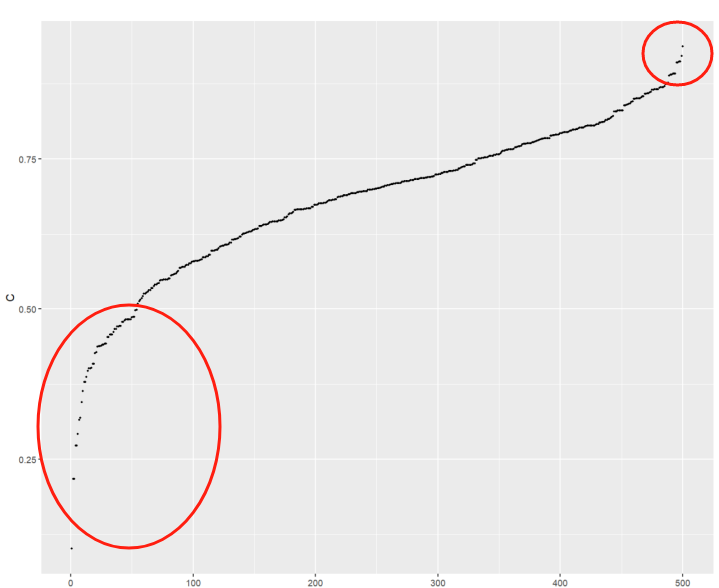
本研究（Entropy Weight Ensemble Framework for Yield Prediction of Winter Wheat Under Different Water Stress Treatments Using Unmanned Aerial Vehicle-Based Multispectral and Thermal Data, 2021）在异常值处理环节采用了基于当时研究条件的判断标准。

（1）论文中图9和图6中箱线图异常值参数处理存在不合理部分，作者硕士于中国农业科学院作物科学研究所进行客座，接触到的师兄师姐均是作物遗传育种背景。在排除异常值时参考了表型数据分析中常用的异常表型剔除方案，即将数据按从小到大的顺序排列，将两端的明显断点前的数据进行剔除（示意图1），箱线图展示去除噪声后参数样本。

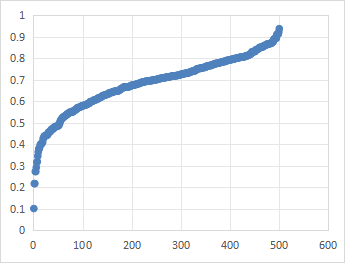


**示意图1** 去除异常值的方法示意图

在最初提交的手稿中，论文并未包含图8。图8的加入是基于审稿人的建议，要求我们补充合并多生育期特征（特征级融合）进行产量预测的内容。在第一版手稿中，作者使用R语言绘制了精度参数（R²值）从小到大排列的散点图（示意图2），并通过寻找显著断点来识别异常值。在返修稿中(距离完成初稿约8个月)，出于效率考虑，作者改用Excel进行精度参数的排序和异常值检测（示意图3）。论文出版后3年，有师弟咨询为何两种方法精度参数变化范围差距较大。在仔细回顾时发现，Excel的排序展示方式与R语言存在显著差异：Excel中的点几乎连续排列，断点极少，仅在小值处存在断点（故基本并未去除断点），这与R语言的结果形成了鲜明对比。当时，我认为图8中使用的方法变化范围更大，并因此感到“惊喜”。

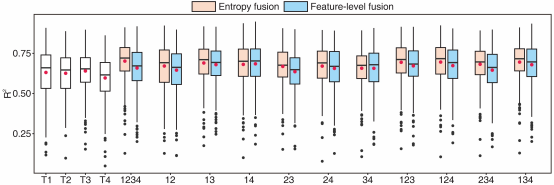


**示意图2** R语言中寻找异常断点的方法示意图



**示意图3** Excel中寻找异常断点的方法示意图 (示意图3与示意图2的绘图数据相同)

因作者转换机构且当时并未在云端备份数据，经询问原服务器现使用人，经过多次系统重装，原始数据已不可获取。核心算法与结论则通过额外数据集进行复现。我们使用其他数据集中4个生育期的植被指数和产量数据对熵权融合算法和特征级融合算法进行了复现（共组建了11种生育期组合）。结果表明，每个生育期组合均能输出比单个生育期更高的预测精度。500次精度参数（R2）的结果（示意图4和附件2）表明，熵权融合算法的产量预测精度在大部分情况下略高于特征级融合，且精度参数的变异系数更小，显示出更高的稳健性。这一结果与论文的核心结论一致。尽管在噪声数据去除过程中由于失误存在方法上的调整（如从R语言转向Excel），这一调整并未对论文的核心结论产生实质性影响。论文的核心结论——**通过熵权对不同生育期的产量预测值进行结合，大部分情况下比直接结合多生育期特征训练模型的预测精度更高且更稳定（精度参数变异系数更小）**——仍然成立。鉴于核心结论未受影响，作者认为无需联系期刊对论文进行修正。此记录旨在对相关现象进行说明备案，以便在未来可能出现的质疑或讨论中作为参考依据。这一做法不仅有助于维护研究的透明性和可追溯性，也能为后续的学术交流提供清晰的背景信息和解释。

**示意图4** 熵权融合和特征级融合在结合多生育期植被指数数据预测产量时的精度参数分布。注：建模数据和500次精度参数的数据均附在附件中

**数据注释**

复现数据（附件1）中包括了4个sheet表格，分别代表小麦的T1孕穗期，T2抽穗期，T3开花期和T4灌浆期的数据。每个sheet表格中第一列均为收获产量，后续的自变量为植被指数。

**二. 额外补充**

除上述异常值处理的主观性判断外，受限于作者当时的研究水平和学术意识，论文中尚有部分材料方法细节未详尽描述，现补充说明如下：

**1 灰色关联分析方面：**

本研究在进行灰色关联分析时发现，由于热红外参数与产量呈负相关关系，直接计算得到的灰色关联度显著偏低，这与农学常识相悖。为消除量纲和方向性影响，采用以下校正方法：

先建立各参数与产量的回归模型，取回归拟合值的绝对值作为灰色关联分析输入最终获得的关联度排序更符合实际农学意义。

**2 生育期及异常指数矫正**

在完成灰色关联分析后，意识到不同灌溉水平和品种间会存在生育期差异。为此，对指数进行了生育期校正处理，以最大可能确保各小区数据处于同一生育阶段（同时也确保了多时相数据的精度提升）。此外，还对部分异常指数值进行了矫正，主要使用相同品种小区进行代替。

**3 热红外数据的重新提取**

在模型构建过程中，我们发现热红外数据对模型性能的提升效果不如预期，这与前人研究结论存在差异。经分析可能存在以下原因：

（1）部分数据采集时机欠佳，在灌溉后即进行无人机飞行，导致冠层温度数据失真；

（2）数据提取方法有待优化。

通过询问灌溉所老师，改进了数据处理方法：

（1）在热红外数据灰色关联度排名第一的生育期重新提取热红外数据，结合裸土温度信息，得到了与产量相关性更高的热红外数据；

（2）对失真的温度数据采用相邻生育期数据进行替代。

在原始研究中，本人将上述3个数据处理方法视为研究过程中的常规技术调整，因此未在论文中详细说明。上述处理均不明显影响论文核心结论方向（且与前人研究结论吻合），现认识到完整的方法描述对研究可重复性的重要性，对此方法进行补充备案。本研究在撰写过程中，由于学术意识不足，在方法细节描述上未严格遵循最高标准。尽管如此，论文的核心结论经检验仍具有可靠性，并与现有研究相符。本研究的呈现存在瑕疵，但核心数据分析和结论仍具科学价值。