

第六届湖南省研究生数学建模竞赛

题 目 **基于多视图处理技术的黄桃采摘贮藏分析**

摘 要：

如何精准的把握采摘时机、如何有效的实施贮藏运输、如何科学的刻画黄桃品质是帮助农户与商户更好的采摘、存运与售卖，推动S市黄桃产业的数字化转型的关键。

在黄桃品质试验中，需要有效预测黄桃品质，同时降低人工评价、有损测试的成本。本文利用异常值剔除、相关性分析、多元回归分析、方差分析等技术，分析了黄桃的机械指标、人工评分、贮藏条件等变量的关联关系，建立了利用机械指标预测人工评分的关联模型，构建了在不同温度条件下贮藏时间与黄桃品质的预测模型，为黄桃的采摘贮藏分析提供了可参考性的实施方案。

首先，对附件1-附件7中的黄桃实验数据进行预处理，采用Pearson系数和协方差分析了黄桃机械指标和人工评分之间的相关关系，并用Kappa系数分析了两类评判方法的一致性。结果表明，机械指标和人工评分内部部分指标具有较强的相关性，而机械指标和人工评分之间的相关性相对较差，两类评判方法具有一定的一致性，但一致性相对较差。

其次，通过附件1-附件6中的黄桃实验数据分别构建了黄桃机械指标和人工评分间的关联模型以及基于人工评分的黄桃品质定量分析模型。结果表明，构建的基于人工评分的黄桃品质定量分析模型对黄桃品质的评判结果介于“专家经验”和“普遍认知”之间。

而后，利用黄桃机械指标和人工评分间的关联模型以及基于人工评分的黄桃品质定量分析模型对附件7中“标黄”样本进行了黄桃人工评分等级（“优质”或“非优质”）判断，优质率为5:9，与往年人工评分优质率一致。

最后，利用附件1-附件7中的黄桃实验数据构建了黄桃贮藏预测模型并采用方差分析进行了试验次数设计。预测结果表明，在冷库(1℃)下，黄桃能够贮藏15天，而在冷柜(8℃)下，黄桃能够贮藏11天。设计结果表明，在冷库(1℃)下，至少应进行4次有损测试，而在冷柜(8℃)下，至少应进行6次有损测试。

基于多视图处理技术的黄桃采摘贮藏分析

图片摘要



文字摘要

如何精准的把握采摘时机、如何有效的实施贮藏运输、如何科学的刻画黄桃品质是帮助农户与商户更好的采摘、存运与售卖，推动 S 市黄桃产业的数字化转型的关键。

在黄桃品质试验中，需要有效预测黄桃品质，同时降低人工评价、有损测试的成本。本文利用异常值剔除、相关性分析、多元回归分析、方差分析等技术，分析了黄桃的机械指标、人工评分、贮藏条件等变量的关联关系，建立了利用机械指标预测人工评分的关联模型，构建了在不同温度条件下贮藏时间与黄桃品质的预测模型，为黄桃的采摘贮藏分析提供了可参考性的实施方案。

首先，对附件 1~附件 7 中的黄桃实验数据进行预处理，采用 Pearson 系数和协方差分析了黄桃机械指标和人工评分之间的相关关系，并用 Kappa 系数分析了两类评判方法的一致性。结果表明，机械指标和人工评分内部部分指标具有较强的相关性，而机械指标和人工评分之间的相关性相对较差，两类评判方法具有一定的一致性，但一致性相对较差。

其次，通过附件 1~附件 6 中的黄桃实验数据分别构建了黄桃机械指标和人工评分间的关联模型以及基于人工评分的黄桃品质定量分析模型。结果表明，构建的基于人工评分的黄桃品质定量分析模型对黄桃品质的评判结果介于“专家经验”和“普遍认知”之间。

而后，利用黄桃机械指标和人工评分间的关联模型以及基于人工评分的黄桃品质定量分析模型对附件 7 中“标黄”样本进行了黄桃人工评分等级（“优质”或“非优质”）判断，优质率为 5:9，与往年人工评分优质率一致。

最后，利用附件 1~附件 7 中的黄桃实验数据构建了黄桃贮藏预测模型并采用方差分析进行了试验次数设计。预测结果表明，在冷库(1℃)下，黄桃能够贮藏 15 天，而在冷柜(8℃)下，黄桃能够贮藏 11 天。设计结果表明，在冷库(1℃)下，至少应进行 4 次有损测试，而在冷柜(8℃)下，至少应进行 6 次有损测试。

关键词：黄桃；相关性分析；一致性分析；回归分析；方差分析；试验设计

1 问题重述

1.1 问题的背景

黄桃是分布于我国长江流域及北部地区的早熟大果型桃，一般于6月中下旬成熟，果面淡黄至金黄，果肉金黄，果汁丰富。果实成熟后可挂树10天左右，采摘对比同期其他品种的耐储运能力也较强，是商用价值较高的一款水果产品。

然而桃子是一种比较容易腐烂的水果，需要精确把握其采摘时机和存储运输方式，才能在风味最佳时上架各个水果店铺。桃子的质量与很多因素有关：可溶性固形物（TSS）、硬度、色差……，这些因素统称为机械指标，共同决定了桃子的质量好坏。对于桃子这些指标的测定，多为有损测试，如利用 FTA 硬度计、SMS 质构仪对果实质地（硬度）进行有损测试，利用手持式折光仪对果实可溶性固形物含量进行有损测试等。

位于 S 市的某黄桃种植基地通过多种实验手段对基地内的某品种黄桃（X 品牌）进行了质量评测，其实验思路主要是在黄桃成熟期的一周左右分三批进行采摘，并设置冷库（1℃）和冷柜（8℃）两种存储方式，每隔 3~5 天取出一部分样本测量各项机械指标与相对应的人工品尝评分。通过这些数据，种植基地希望能够科学刻画黄桃品质与贮藏运输的关系，以帮助农户与商户更好的采摘、存运与售卖，推动 S 市黄桃产业的数字化转型。

该种植基地提供了 2018 年至 2020 年的黄桃评测数据（有部分实验指标未测量或数据缺失），希望依据实验数据，对以下问题进行建模分析。

1.2 问题的提出

1. 黄桃品质分析：依据专家经验，优质黄桃的机械指标评分应以果实硬度 $> 1.5\text{kg/cm}^2$ 。TSS 含量下降率 $< 20\%$ （即 TSS 初始指标和贮藏期间的指标对比）来度量。而人工评分时，当色泽、质地和味道这三个指标均分 $\geq 80\%$ 分时认为样品“优质”。目前采用双评分机制，普遍认知是机械指标比较容易满足，而人工评分要求更高。当前对果实品质多以机械指标评判，而感官评价也代表了果实的综合品质和性状。如何利用机械指标和人工评分，更加科学的评判果实品质，来应对市民对高品质果品的需求，是本研究的主要目的。

问题 1：请根据附录 1 中的实验描述和附件 1~附件 7 中的实验数据，分析机械指标、人工评分之间的关系与两类方法的评判一致性，构建黄桃品质分析的定量模型，为商家推荐优质黄桃，请比较你的模型得到的结论与上述提到的“专家经验”和“普遍认知”的差异。

问题 2：2020 年度的评测中取消了人工评分，现随机抽出冷柜贮藏方式下没批采摘的 3 个测试样本（附件 7 中“标黄”命名为“样本 A ‘-’ 样本 I “），请结合问题 1 中的相关模型和结论，根据本年度的数据建立模型判断这 9 个样本的人工评分等级（“优质”或“非优质”），并将结果填入下表成像在论文中。

2. 黄桃贮藏预测模型：当黄桃从树上采摘下来后，需要及时贮藏以保持口感风味，而不同的存储温度带来了不同的运输和贮藏成本，并可能影响黄桃品质。目前普遍认为在冷库（1℃）贮藏情况下，低温会影响黄桃代谢，风味有所下降，但保质期较长；在冷柜（8℃）贮藏情况下，黄桃代谢情况较好，但保鲜期较短。

问题 3：请根据附录 1 中的实验描述与附件 1~附件 7 中的实验数据，综合考虑贮藏环境（冷库或冷柜）和贮藏时长对黄桃品质的影响，构建黄桃品质随贮藏环境和贮藏时间变化的预测模型，为商家推荐合适的贮藏方式和时间。你的数据

分析结果与上述“普遍认为“是否一致？如果下一年度果农需要用该模型预测采摘黄桃的贮藏期，按照你们的模型，果农需要进行多少次有损测试？

2 问题分析

对于问题一，我们分析了黄桃机械指标和人工评分之间的相关关系与两类方法的一致性分析，建立了黄桃品质定量分析模型。首先，对附件 1~附件 7 中的数据进行预处理，消除异常值以及个体误差对分析的影响。在此基础上，分析黄桃机械指标、人工评分之间的相关关系，并对两类方法进行一致性分析。其次，依据相关系数，选取相关性低的机械指标作为自变量，人工评分作为因变量，构建单输出多元线性回归模型及多输出多元线性回归模型。在两类回归模型中，根据拟合效果，增加非线性变量，挑选最优模型。最后，根据所构建的模型，对比“专家经验”和“普遍认知”的差异。

对于问题二，基于问题一中构建的模型，带入附件 7 中标黄的样本数据，拟合人工评分，依据人工评分等级，判断样本等级是否优质。

对于问题三，利用问题一中构建的模型以及黄桃的机械指标数据，采用非线性回归分析方法，在冷库 1℃和冷柜 8℃下分别构建黄桃品质随贮藏环境和贮藏时间变化的预测模型。采用方差分析方法，对试验次数进行设计，以此得到最终的有损测试次数。

3 模型假设

1. 假设同一批次黄桃中的每个样本与该批次总体样本趋势一致，忽略个体间的差异；
2. 假设 2019 年黄桃机械指标数据中的左右带皮硬度、左右去皮硬度、左右 TSS 以及左右色差均表示整个黄桃指标的测量值而非单个黄桃左右部分指标的测量值。

4 模型建立与求解

为更好的分析、研究和解决上述问题，我们采用以下步骤求解上述问题：

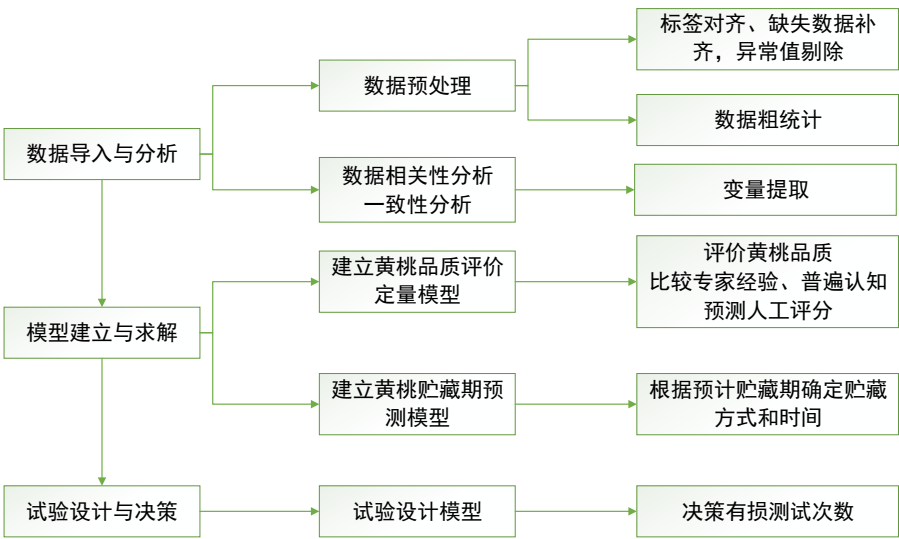


图 1 总体流程图

4.1 黄桃品质分析

桃子的质量与很多因素有关，目前多采用有损测试获得机械指标对桃子品质进行评判。为更加科学的评判果实品质，应对市民对高品质果品的需求，采用双评分机制评判果实品质，因此，问题一和问题二要求分析机械指标、人工评分之间的关系与两类方法的评判一致性，构建黄桃品质分析的定量模型，并比较模型得到的结论与上述提到的“专家经验”和“普遍认知”的差异，根据相关模型和结论，依据机械指标判断人工评分等级。

4.1.1 数据预处理

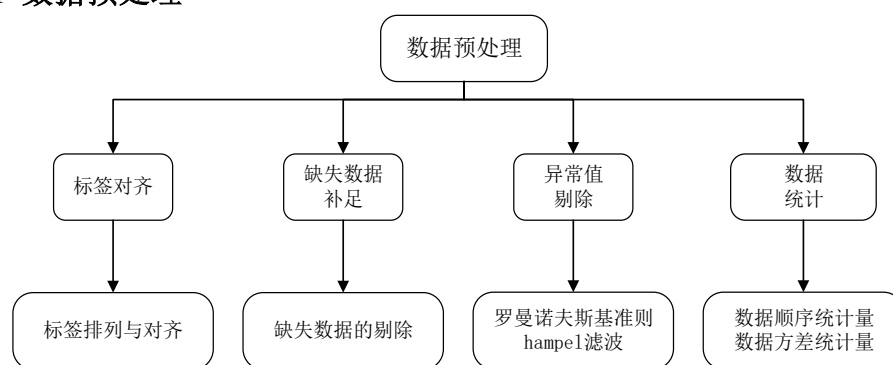


图 2 数据预处理方法图

4.1.1.1 数据标签错位和缺失数据的处理

经过对数据的查找、比较，我们发现部分原始数据存在标签错位和数据缺失情况。

为方便后续处理，针对部分原始数据标签存在错位情况，通过比对该部分数据之前以及之后的数据，对标签数据位置进行调整，使得同一个标签下的数据值在同一列。

针对数据存在实验指标未测量或数据缺失情况，由于优质黄桃的机械指标评分以果实硬度和 TSS 含量下降率来度量，而人工评分以色泽、质地和味道三个指标的均分来度量，因此，缺失果实硬度或 TSS 含量将会影响机械指标评分，缺失任一色泽、质地和味道均会影响人工评分，然而单个黄桃样本仅趋势与该批次整体相同，不同个体之间并不相关，故而果实硬度或 TSS 含量的缺失无法通过其他黄桃个体对其进行填补，另外，每次人工评分的测试员不同，测试员的人数不同，测试员的测试水平未知，测试员的编号也不总是知道，故而缺失任一色泽、质地和味道均无法通过其他人工评分数据进行增补。

综上，当任一缺失果实硬度和 TSS 含量指标时，我们舍弃与之相对的机械指标值，当任一缺失色泽、质地和味道指标时，我们舍弃与之相对的人工评分指标值。

4.1.1.2 异常值剔除

由于黄桃的成熟度没有一个量化指标，同一批次的黄桃的评测指标的分布未知，因此，每次取出一部分黄桃存在个别黄桃在置信区间外，使得测量的各项指标偏离大部分黄桃的各项指标，因此，我们必须剔除个别偏离指标值，以便能客观反映黄桃实际的趋势。所以，我们对附件 1~附件 7 的数据进行了处理，剔除异常值，达到消除个体偏离的目的。

1. 罗曼诺夫斯基准则^[1]

假设处理数据前,认为测量的黄桃数据服从正态分布，但是数理统计学可以

证明，在测量次数较少的情况下，t 分布更符合实际分布。因此，在一定测量次数 n 下，设等精度独立测得的一组黄桃数据为 $x_1, x_2, \dots, x_n$ ，若对某一数据 x_k 有怀疑，可以按照如下步骤判别：

- 1) 先将怀疑数据 x_k 去掉，计算出不包含数据 x_k 的测量数据的算术平均值 $\bar{x}$

$$\bar{x} = \frac{1}{n-1} \sum_{i=1, i \neq k}^n x_i \quad (1)$$

- 2) 计算不包含怀疑数据 x_k 的残差 v_i 和标准偏差 s

$$s = \sqrt{\frac{\sum_{i=1, i \neq k}^n v_i^2}{n-2}} \quad (2)$$

- 3) 根据选定的显著水平 α 和测量次数 n ，在 t 分布表中查出检验系数 $K(\alpha, n)$, $\delta = K(\alpha, n)s$

- 4) 若 $|x_k - \bar{x}| > K(\alpha, n)s$ ，则可判定 x_k 为异常数据，应予以剔除，若 $|x_k - \bar{x}| < K(\alpha, n)s$ ，则此 x_k 就不是异常数据，应予以保留。

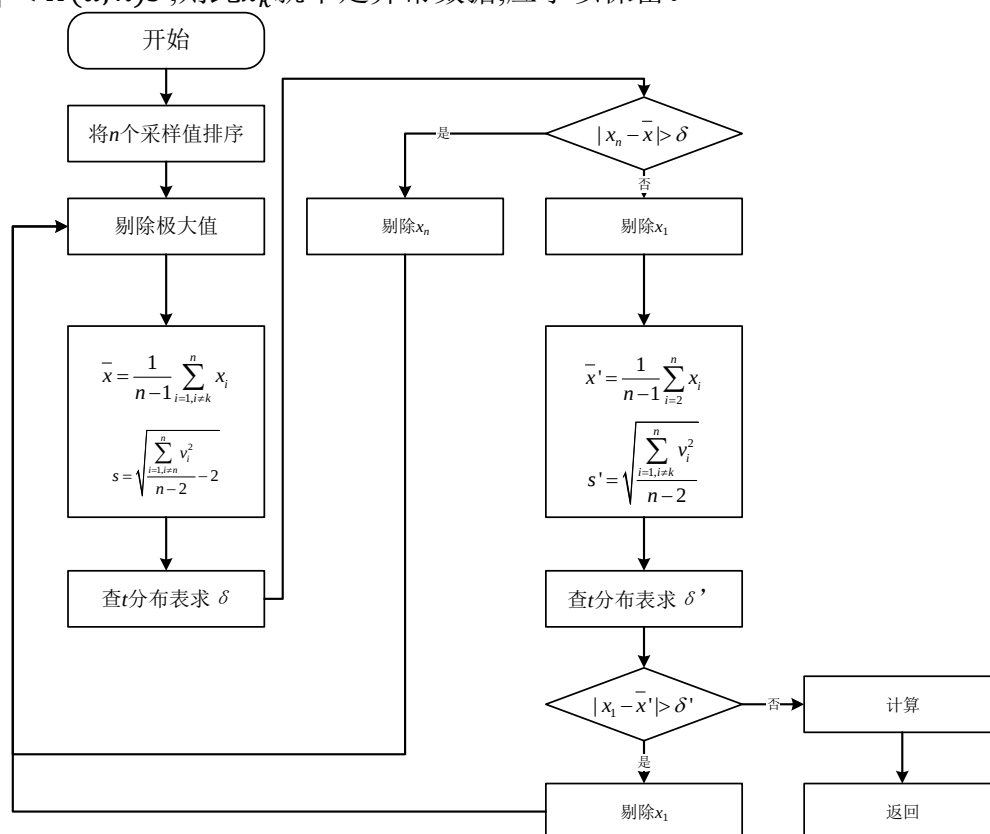


图3 罗曼诺夫斯基准则实现框

通过上述方法，对附表 1~附表 7 中的数据进行了野点剔除，发现附件 4 中 6 月 20 日冷库 1℃冷库第六天有损 TSS/左 111.5 以及附件 5 中 7 月 3 日冷库 1℃第十八天 a/右 5 数据为野点，对野点进行剔除，具体如下图所示：



图 4 附件 1~附件 7 中野点标识图

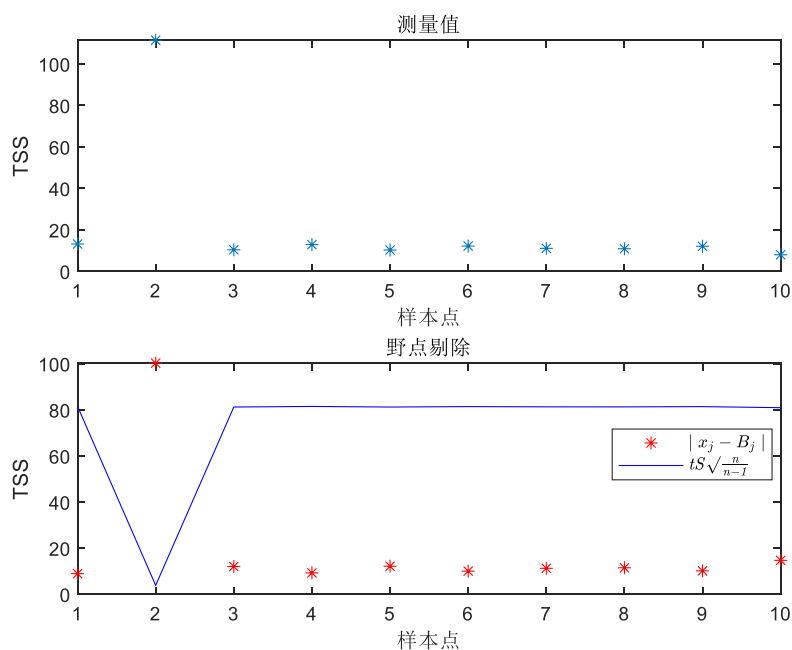


图 5 罗曼诺夫斯基野点剔除

2. hamper 辨识器^[2]

设同一批次的黄桃样本数据集为 $X = \{x_1, x_2, \dots, x_n\}$, 设 x_k 为异常值, $median(X)$ 为 X 的中位数, $median(X)$ 通过对 X 进行一阶排序获得, 如下式所示:

$$x_1 < x_2 < \dots < x_n \quad (3)$$

构建数据集 Y

$$Y = \{y_1, y_2, \dots, y_n\} = \{|x_1 - median(X)|, \dots, |x_n - median(X)|\} \quad (4)$$

则比例因子为:

$$t_i = \frac{|x_i - median(X)|}{1.4826 \times median(Y)}, i = 1, \dots, n \quad (5)$$

其中， $median(Y)$ 为 Y 的中位数。当 $t_i > 3$ 时，对应的黄桃数据 x_i 被确定为异常值。

通过上述方法，对附表 1~附表 7 中的数据进行了异常值剔除，实验发现，6 月 20 日冷柜 8℃ 冷藏柜第十四天去皮硬度 3.02 3.73 3.42 2.24 3.75 3.99 为异常值，我们对异常值进行剔除，具体如下图所示

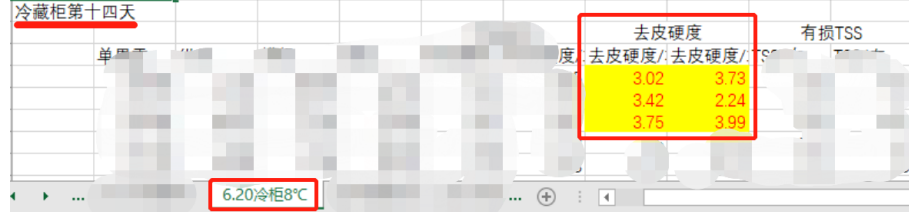


图 6 附件 1~附件 7 中异常数据标识图

4.1.1.3 数据粗统计

为便于下一步处理，对黄桃的各项指标有一个初步认知，我们通过统计方法对附件 1~附件 7 中的数据进行初步统计，对黄桃的各项指标有一个初步的认知，具体如下：

首先，我们按照下式求取同一天测量的各项指标的均值

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (6)$$

其次，为进一步方便后续处理，我们依据附录 1~附录 7 中的所得的数据，依据公式（6）所求得的指标均值，按照下式分别对数据值进行标准化处理

$$\begin{cases} x' = \frac{x - \bar{x}}{\sqrt{\text{var}(x)}} \\ \text{var}(x) = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \end{cases} \quad (7)$$

其中， $\text{var}(x)$ 为样本方差。具体均值数据如下表所示：

4.1.2 黄桃机械指标和人工评分之间的相关性分析

4.1.2.1 基于 Pearson 系数的相关性分析

设 X_k 、 Y_l 分别为矩阵 X 的第 k 列和矩阵 Y 的第 l 列，则 X_k 和 Y_l 的样本均值为

$$\begin{cases} \bar{X}_k = \sum_{i=1}^n (X_{k,i})/n \\ \bar{Y}_l = \sum_{i=1}^n (Y_{l,i})/n \end{cases} \quad (8)$$

则 Pearson 线性相关系数 ρ 为

$$\rho = \frac{\sum_{i=1}^n (X_{k,i} - \bar{X}_k)(Y_{l,i} - \bar{Y}_l)}{\sqrt{\{\sum_{i=1}^n (X_{k,i} - \bar{X}_k)^2 \sum_{i=1}^n (Y_{l,i} - \bar{Y}_l)^2\}^{\frac{1}{2}}}} \quad (9)$$

相关系数的值的范围是从 -1 到 +1。当相关系数 ρ 的值为 -1 时，表示 X_k 和 Y_l 完全负相关；当相关系数 ρ 的值为 1 时，表示 X_k 和 Y_l 完全正相关；当相关系数 ρ 的值为 0 时，表示 X_k 和 Y_l 之间没有相关性。

4.1.2.2 基于协方差的相关性分析

协方差是描述二维随机变量两个分量间相互关联程度的一个特征数，如果将协方差相应标准化变量就得到相关系数，由于在数据预处理部分已经对数据进行了标准化，因此，可以通过协方差评判变量间的相关关系。

假定 $\mathbf{X}$ 是经过标准化处理后的数据矩阵，称下式为第 i 与第 j 个变量的样本相关系数

$$r_{ij} = \sum_{k=1}^n x_{ki}x_{kj} \quad (i = 1, \dots, n; j = 1, \dots, n) \quad (10)$$

并称下式为样本相关系数矩阵，简称样本相关矩阵

$$\mathbf{R} = \mathbf{X}^T \mathbf{X} \quad (11)$$

当协方差值大于零时，称两个变量正相关；当协方差值小于零时，称两个变量负相关；当协方差值等于零时，称两个变量不相关。

我们采用黄桃各项指标数据的均值，求取了黄桃的硬度、TSS、色调等机械指标以及色泽、质地、味道等人工指标之间的 Pearson 系数和协方差值，发现两种方法的结果相同，因此，我们选用协方差进行相关性分析，结果如下图所示，为更为直观的表现两组变量之间的相关性，用颜色块将样本相关矩阵的样本相关系数值作图，得到对称的相关系数热力图。

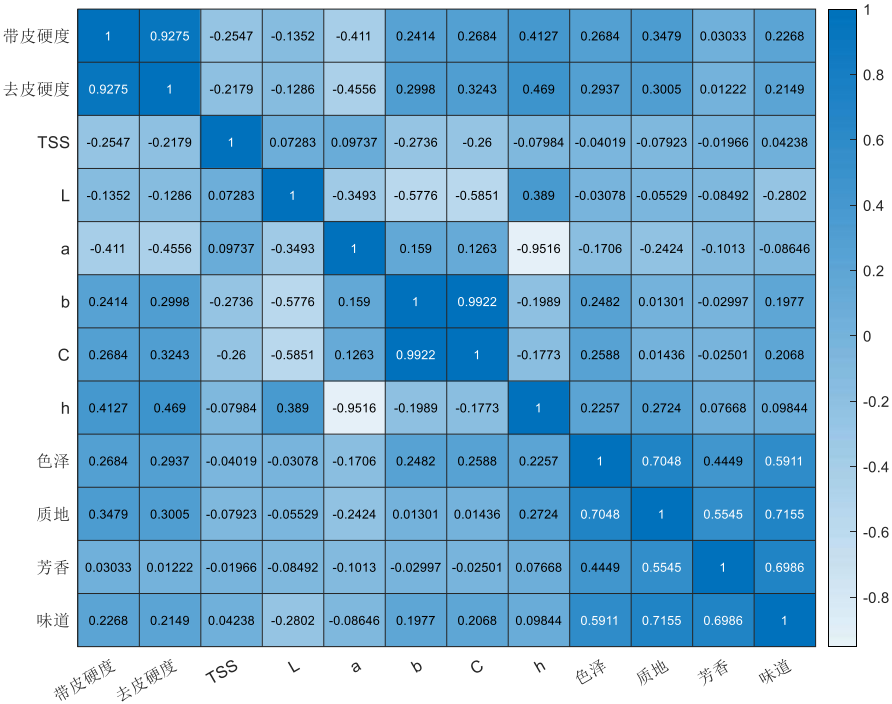


图 7 黄桃各项指标的相关系数热力图

从图中可以看出，黄桃的带皮硬度与去皮硬度的相关系数高达 0.9275，因此认为黄桃的带皮硬度与去皮硬度具有强相关性，果皮的 b（黄蓝色）与 C（色彩饱和度）强相关，人工评价的四个指标：色泽、质地、芳香、味道具有强相关性。下面分别将其相关关系进行可视化：

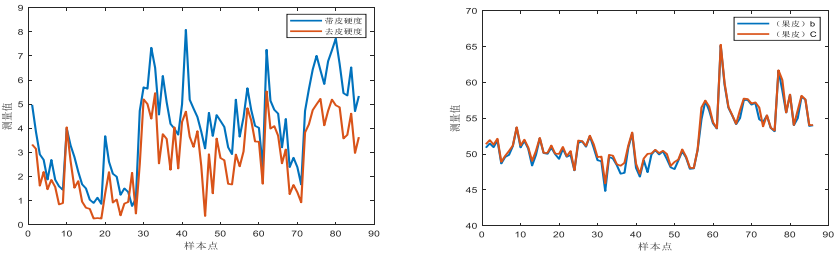


图 8 部分机械指标间的相关关系图

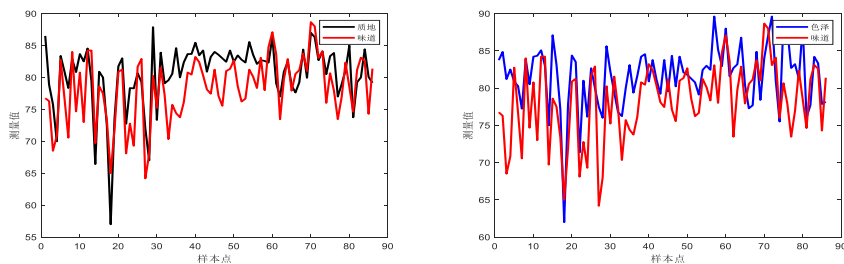


图 9 部分人工评分间的相关关系图

从图 7 中可以进一步看出，机械指标中的带皮和去皮硬度均与人工指标中的色泽、质地和味道弱相关，而机械指标中的其他指标与人工指标同样具有一定的若相关。因此，通过两类方法分别评判黄桃品质，评判结果可能呈现一定的弱相关性。

4.1.3 两类评判方法的一致性分析

上节通过求取各个机械指标和人工指标之间的相关系数，对两类方法评判黄桃品质的评判结果的相关性进行了分析。然而，由于黄桃品质是利用多个评判指标进行综合评判的结果，因此，本节中采用 Kappa 系数对机械指标评判和人工评分的评判结果进行一致性分析。

Kappa 系数的一致性分析的主要实现是：从比较两个观测者对同一事物的观测结果或同一观测者对同一事物的两次观测结果是否一致出发，以由于机遇造成的一致性和实际观测的一致性之间的差别大小作为评价基础所提出的统计指标。针对资料类型的不同，常用的 Kappa 系数包括针对无序分类资料(含二分类和多分类)的简单 Kappa 系数和针对等级资料的加权 Kappa 系数^[4]。

加权 Kappa 系数是简单 Kappa 系数的推广，是用加权的方法对两个评价结果进行量化，主要用于等级资料。对于四格表来说，简单 Kappa 系数与加权 Kappa 系数是相等的，对于行×列的列联表资料，加权 Kappa 系数 K 的计算公式为：

$$\begin{cases} K_w = \frac{P_0(w) - P_e(w)}{1 - P_e(w)} \\ P_0(w) = \frac{\sum_i \sum_j w_{ij} A_{ij}}{n} \\ P_e(w) = \frac{\sum_i \sum_j w_{ij} A_i A_j}{n^2} \end{cases} \quad (12)$$

其中， A_{ij} 为方法 1 第 i 个水平和方法 2 第 j 个水平对应的频数； w_{ij} 为方法 1 第 i 个水平和方法 2 第 j 个水平对应频数的权重， $0 \leq w_{ij} = w_{ji} < 1, i \neq j, w_{ii} = 1$ ；

A_i 为方法 1 的第 i 个水平的所有频数； A_j 为方法 2 的第 j 个水平的所有频数。

确定 w_{ij} 主要有两种方法：Cicchetti—Allison 线性权重法(CA)和 Fleiss—Cohen 平方权重法(FC)，估计方法分别为：

$$w_{ij} = \begin{cases} 1 - \frac{|C_i - C_j|}{C_{\max} - C_{\min}} & \text{CA法} \\ 1 - \frac{(C_i - C_j)^2}{(C_{\max} - C_{\min})^2} & \text{FC法} \end{cases} \quad (13)$$

其中, C_i 和 C_j 分别表示第 i 和第 j 个水平的评价分值, $C_{\min}$ 和 $C_{\max}$ 分别表示最小和最大的评价分值。Kappa 系数 K 的值介于 -1 到 1 之间。当 $K < 0$ 时, 说明 $P_a < P_e$, 即两种观测结果不一致; 当 $K = -1$ 时, 说明两种观测结果完全不一致; 当 $K = 0$ 时, 说明 $P_a = P_e$, 即观察的一致率完全是由随机因素造成的; 当 $K > 0$ 时, 说明 $P_a > P_e$, 即两种观测结果较大可能存在一致性, 由于随机因素引起的可能性较小。

应用专家经验: TSS 下降率 < 20%, 硬度 > 1.5 kg/cm^2 为优质进行一致性检验。经过计算, 所提供的所有黄桃样本的 TSS 下降率均小于 20%, 具体如下表所示:

表 1 机械指标与人工指标的评判结果

			人工指标		合计
			非优	优	
机械指标	非优	计数	7	2	9
		机械指标(%)	77.8%	22.2%	100.0%
	优	计数	29	48	77
		机械指标(%)	37.7%	62.3%	100.0%
合计		计数	36	50	86
		机械指标 中的 %	41.9%	58.1%	100.0%

表 2 对称度量

	值	渐近标准误差	近似值 T	近似值 Sig
一致性度量	0.173	0.079	2.308	0.21

根据 Kappa 一致性度量准则, 当检验的 $Kappa \geq 0.85$ 时, 认为检验结果的一致性很好, 当 $0.6 \leq Kappa < 0.85$ 时, 认为检验结果的一致性较好, 当 $0.45 \leq Kappa < 0.6$ 时, 认为检验结果的一致性一般; 当 $Kappa < 0.45$ 时, 认为检验结果的一致性较差。根据上述结果, 可以看出在给定的 TSS 下降率 < 20%, 硬度 > 1.5 kg/cm^2 条件下, 认为人工指标与机械指标的一致性较差。由于在人们对黄桃品质的实际认知中, 当 TSS 下降 2% 时就已经严重影响黄桃品质, 因此, 考虑将 TSS 下降率调整为小于 2%:

应用专家经验, 对 TSS 下降率 < 2%, 硬度 > 1.5 kg/cm^2 为优质的黄桃进行一致性检验, 结果如下所示

表 3 机械指标与人工指标的评判结果

			人工指标		合计
			非优	优	
机械指标	非优	计数	12	18	30
		机械指标(%)	40.0%	60.0%	100.0%
	优	计数	24	31	55
		机械指标(%)	43.6%	56.4%	100.0%
合计		计数	36	49	85
		机械指标(%)	42.4%	57.6%	100.0%

表 4 对称度量

	值	渐近标准误差	近似值 T	近似值 Sig
一致性度量	-0.035	0.107	-0.324	0.746

根据上述结果,看出在 TSS 下降率>2%,硬度>1.5 条件下,人工指标与机械指标具有一定的一致性,但一致性相对较差,这是由于相比于机械指标值,人的因素对黄桃品质的评价干扰较大,影响了对黄桃品质的一致性评判结果。

4.1.4 黄桃品质的定量模型

4.1.4.1 回归模型的一般形式

假设因变量 y 与一个或多个自变量 $x_1, x_2, \dots, x_p$ 之间具有相关关系, y 的取值受两部分的影响^[5]: 1) 由自变量 $x_1, x_2, \dots, x_p$ 确定的部分,记为 $f(x_1, x_2, \dots, x_p)$; 2) 随机因素的影响,这部分可看成随机误差,记为 ε 。因此可得到如下的统计模型:

$$y = f(x_1, x_2, \dots, x_p) + \varepsilon \quad (14)$$

上式中的 ε 是不可观察的随机误差,假设有

$$E(\varepsilon) = 0, D(\varepsilon) = \sigma^2 < \infty \quad (15)$$

称 $E(y) = f(x_1, x_2, \dots, x_p)$ 为回归函数,该模型称为回归模型的一般形式,当回归函数是线性函数时,模型变为如下形式:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon \quad (16)$$

其中, $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ 是未知参数,称为模型参数,上式称为线性回归模型。

在实际问题中,参数 $\beta_0, \beta_1, \beta_2, \dots, \beta_p, \sigma^2$ 通常都是未知的,因此需要估计出它们的值,要估计出参数的值,就需要样本. 选取自变量 n 组不同的值: $(x_{i1}, x_{i2}, \dots, x_{ip}), i = 1, 2, \dots, n$ 分别进行 n 次独立的试验或观察,得到因变量的 n 个相互独立的观察值 $y_1, y_2, \dots, y_n, i = 1, 2, \dots, n$ 则由线性回归模型可知

$$\begin{cases} y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \\ E(\varepsilon_i) = 0, D(\varepsilon_i) = \sigma^2, i = 1, 2, \cdots, n \\ \varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \text{相互独立.} \end{cases} \quad (17)$$

通常将上式写成矩阵形式,记

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad (18)$$

则多元线性回归模型可以写成如下形式:

$$\begin{cases} Y = X\beta + \varepsilon \\ E(\varepsilon) = 0, D(\varepsilon) = \sigma^2 I_n \end{cases} \quad (19)$$

其中, Y 是已知的 n 维观察向量, X 是已知的 $n \times (p+1)$ 阶矩阵,称为设计矩阵, β 是未知的 $p+1$ 维未知参数, $\sigma^2 I_n$ 表示 n 维随机误差向量 ε 的协方差矩阵, 即 $D(Y) = D(\varepsilon) = [\text{Cov}(\varepsilon_i, \varepsilon_j)]_{n \times n} = \sigma^2 \mathbf{I}_n$ 线性回归模型 通常是研究问题的出发点, 特别当随机误差 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ 独立同分布服从正态分布 $N(0, \sigma^2)$ 时, 线性回归模型称为正态线性模型。

4.1.4.2 多元线性回归模型

假设因变量 y 与多个自变量 $x_1, x_2, \cdots, x_p$ 之间具有如下线性关系

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \\ E(\varepsilon) = 0, D(\varepsilon) = \sigma^2 \end{cases} \quad (20)$$

其中, y 是可观察的随机变量, $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ 是未知参数, ε 是不可观察的随机误差, $x_1, x_2, \cdots, x_p$ 是可观察或控制的确变量, 称上式为多元线性回归模型。在实际问题中, 通常会获得 n 组互不相关的观察数据 $(y_i, x_{i1}, x_{i2}, \cdots, x_{ip})$ ($i = 1, 2, \cdots, n$), 则由多元线性回归模型有

$$\begin{cases} y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \\ E(\varepsilon_i) = 0, \text{Cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 \delta_{ij} \\ i, j = 1, 2, \cdots, n \end{cases} \quad (21)$$

其中

$$\delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \quad (22)$$

记

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad (23)$$

则可写成如下矩阵形式

$$\begin{cases} Y = X\beta + \varepsilon \\ E(\varepsilon) = 0, D(\varepsilon) = \sigma^2 I_n \end{cases} \quad (24)$$

其中, Y 是已知的观测向量, X 是已知的矩阵, 称为设计矩阵, 总假设 X 是列满秩的, 即 $\text{rank}(X) = p + 1$, β 是未知参数向量, ε 是不可观察的随机误差向量.

4.1.4.3 多元线性回归模型的参数估计

记离差平方和为

$$\begin{aligned} Q(\beta) &= (Y - X\beta)^T(Y - X\beta) \\ &= \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip})^2 \end{aligned} \quad (25)$$

采用最小二乘法求未知参数向量 β 的估计值 $\hat{\beta}$, 即寻找 $\hat{\beta}$, 使得离差平方和 $Q(\beta)$ 最小:

$$Q(\hat{\beta}) = \min_{\beta} Q(\beta) \quad (26)$$

称 $\hat{\beta}$ 为未知参数向量 β 的最小二乘估计。因为 $Q(\beta)$ 是未知参数 $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ 的二次非负函数, 所以一定有最小值, 其最小值点 $\hat{\beta}$ 即为 β 的最小二乘估计。对 $Q(\beta)$ 求关于未知参数 $\beta_i (i = 0, 1, 2, \dots, p)$ 的偏导数, 令其等于 0, 得到

$$\begin{cases} \frac{\partial Q(\beta)}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip}) = 0 \\ \frac{\partial Q(\beta)}{\partial \beta_1} = -2 \sum_{i=1}^n x_{i1} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip}) = 0 \\ \frac{\partial Q(\beta)}{\partial \beta_2} = -2 \sum_{i=1}^n x_{i2} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip}) = 0 \\ \dots \dots \dots \\ \frac{\partial Q(\beta)}{\partial \beta_p} = -2 \sum_{i=1}^n x_{ip} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip}) = 0 \end{cases} \quad (27)$$

整理得到如下正规方程组

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^n x_{i1} + \beta_2 \sum_{i=1}^n x_{i2} + \cdots + \beta_p \sum_{i=1}^n x_{ip} = \sum_{i=1}^n y_i \\ \beta_0 \sum_{i=1}^n x_{i1} + \beta_1 \sum_{i=1}^n x_{i1} x_{i1} + \beta_2 \sum_{i=1}^n x_{i2} x_{i1} + \cdots + \beta_p \sum_{i=1}^n x_{ip} x_{i1} = \sum_{i=1}^n x_{i1} y_i \\ \beta_0 \sum_{i=1}^n x_{i2} + \beta_1 \sum_{i=1}^n x_{i1} x_{i2} + \beta_2 \sum_{i=1}^n x_{i2} x_{i2} + \cdots + \beta_p \sum_{i=1}^n x_{ip} x_{i2} = \sum_{i=1}^n x_{i2} y_i \\ \dots \dots \dots \\ \beta_0 \sum_{i=1}^n x_{ip} + \beta_1 \sum_{i=1}^n x_{i1} x_{ip} + \beta_2 \sum_{i=1}^n x_{i2} x_{ip} + \cdots + \beta_p \sum_{i=1}^n x_{ip} x_{ip} = \sum_{i=1}^n x_{ip} y_i \end{cases} \quad (28)$$

上述方程组可写成如下矩阵形式

$$X^T X \beta = X^T Y \quad (29)$$

由假设 $\text{rank}(X) = p + 1$ 知, 方阵 $X^T X$ 是满秩矩阵, 所以 $(X^T X)^{-1}$ 存在,

即正规方程组有唯一解

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (30)$$

称 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$ 为多元经验回归方程, 称

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip} \quad (31)$$

为观察值 $y_i (i = 1, 2, \cdots, n)$ 的拟合值或回归值, 称 $e_i = y_i - \hat{y}_i$ 为残差, 称

$$e = [e_1, e_2, \cdots, e_n]^T \quad (32)$$

为残差向量或剩余向量. 记

$$\hat{Y} = [\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n]^T \quad (33)$$

则有

$$\begin{cases} \hat{Y} = [\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n]^T = X \hat{\beta} = X(X^T X)^{-1} X^T Y \\ e = Y - \hat{Y} = (I_n - X(X^T X)^{-1} X^T) Y \end{cases} \quad (34)$$

由上式可知, 矩阵 $X(X^T X)^{-1} X^T$ 相当于一个线性变换, 将因变量向量 Y 变换为拟合值向量 $\hat{Y}$ 。

4.1.4.4 黄桃机械指标与人工评分的回归分析

1. 采用上述理论方法, 以 y 为人工评分的色泽, 质地, 味道三个指标取均值。以机械测量指标为参数 β , X 为设计矩阵。

选择附录 1-附录 6 中的黄桃各项指标数据, 依据图 7 中黄桃各项指标的相关系数热力图可知, 带皮硬度与去皮硬度有极强正相关性、 a 和 h 有极强负相关性、 b 和 C 有极强正相关性, 因此选择带皮硬度, TSS, L , a , b , TSS 变化率为自变量, 即 $\beta = (S, T, L, a, b, dt)^T$, 方程即为

$$y = x_1 S + x_2 T + \cdots + x_6 dt + k \quad (35)$$

其中, k 为常数。利用最小二乘估计求出 X , 求出回归方程为

$$y = 0.4812S + 0.3895T - 0.2383L - 0.2375b + 0.1105h - 0.0581dt + 87.3589 \quad (36)$$

对设计矩阵 X 分析可知, 硬度, TSS, 色调角 (h) 与评分有较强正相关, 明暗度 (L) 与评分正相关, TSS 变化率与评分负相关, 其结果与人们的普遍认知相同。根据计算出的 X , 带入机械测量指标值, 求出拟合人工评分, 结果如图所示。其中拟合人工总评分 >80 分为优质。

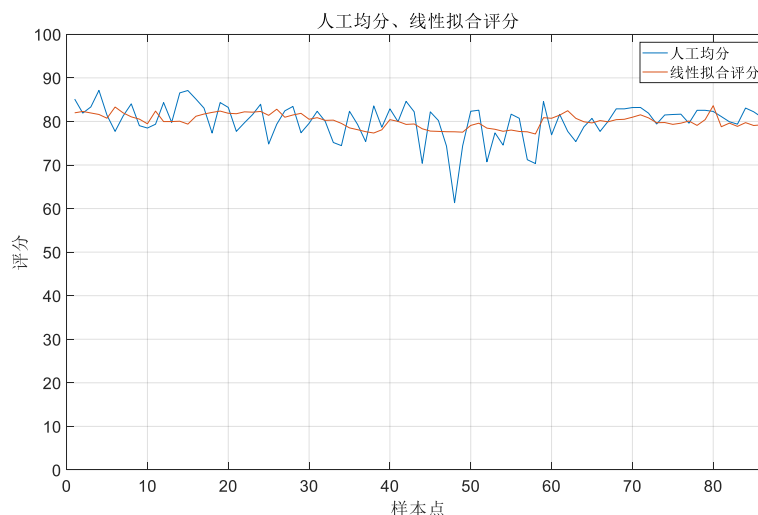


图 10 人工均分的线性拟合图

人工均分与利用线性回归方程拟合出的结果做差，得到

$$\Delta y = y - y' \quad (37)$$

则平均绝对误差(MAE)为：

$$MAE(\beta, f) = \frac{1}{m} \sum_{i=1}^m |f(\beta_i) - y_i| \quad (38)$$

计算统计信息如下表所示：

表 5 线性参数值

单输出回归模型	均值	MAE
线性参数模型	-0.0185	2.9697

综述可知，采用黄桃的机械指标值拟合人工评分，采用线性回归模型能够在一定程度上，拟合人工评分，但拟合效果不理想，因此，黄桃的机械指标与人工评分之间具有非线性关系，应考虑非线性模型拟合人工评分。

2. 由上节可以看出，对人工均分的线性拟合效果较差，为进一步提高拟合效果，对人工均分采用非线性拟合。

依据人们对黄桃硬度的普遍认知可知，黄桃硬度过高或者过低都会影响黄桃的品质。假设硬度、TSS 与评分为非线性关系，则评分 y 与硬度满足如下对应关系：

$$y \propto aS^2 + bS \quad (39)$$

其中， S 为所测硬度， a 应小于零。为进一步评判带皮硬度与评分之间的关系，依据图 7 中黄桃各项指标的相关系数热力图，对皮硬度与评分进行初步的定性分析，带皮硬度与评分之间的关系如下图所示

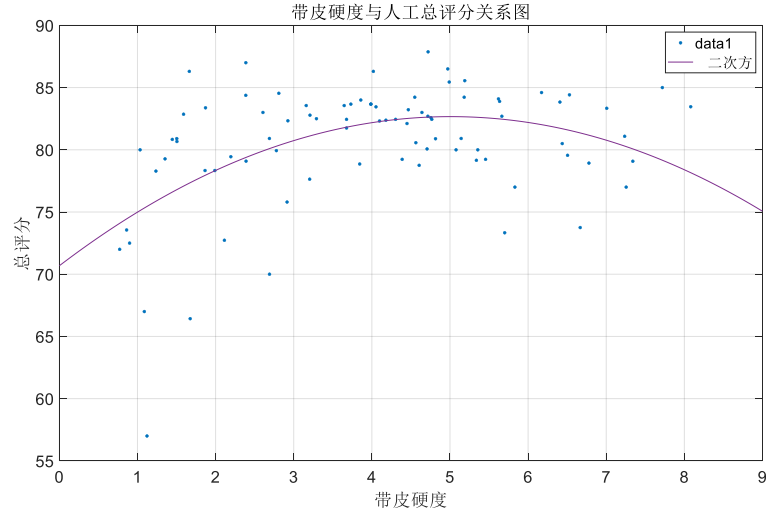


图 11 带皮硬度与人工总评分关系图

可以看出，刚开始评分随带皮硬度增加而增加，而后达到最高点，后续评分随带皮硬度增加而减少。根据以上趋势，上述线性回归方程写为

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{K} \quad (40)$$

其中， $\boldsymbol{\beta} = (S^2, S, TSS^2, TSS, L, a, b, dt, k)^T$, $\mathbf{X} = (x_1, x_2, \dots, x_9)$ 。解算设计矩阵 $\mathbf{X}$ ，求出回归方程为

$$y = -0.3411S^2 + 3.4607S - 0.0867T^2 + 3.0455T - 0.0857L - 0.0759b + 0.1888h - 0.0558dt + 46.2779 \quad (41)$$

根据理论分析，非线性参数的二次项系数应为负数，一次项系数应为正数，与结果相符。同时，明暗度（L），与评分正相关，TSS 变化率与评分负相关，结果与常识保持一致。

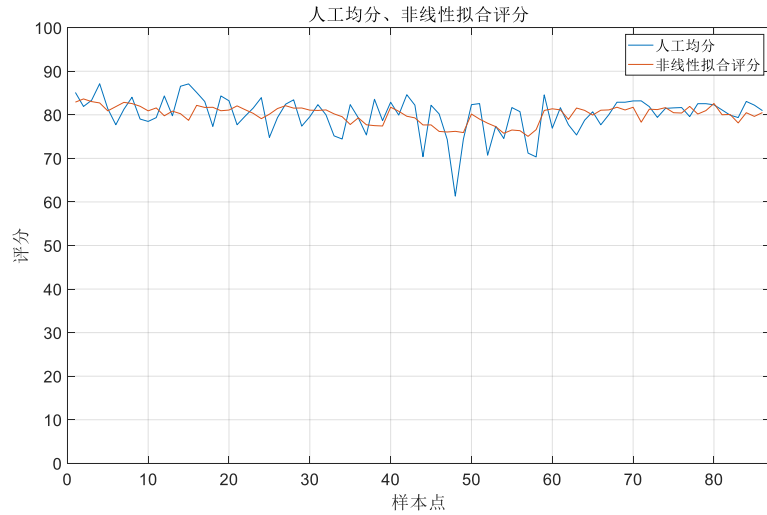


图 12 人工均分的非线性拟合图

人工均分与利用线性回归方程拟合出的结果做差，得到

$$\Delta \mathbf{y} = \mathbf{y} - \mathbf{y}' \quad (42)$$

则平均绝对误差为

$$MAE(\boldsymbol{\beta}, \mathbf{f}) = \frac{1}{m} \sum_{i=1}^m |\mathbf{f}(\boldsymbol{\beta}_i) - \mathbf{y}_i| \quad (43)$$

计算统计信息如下表所示

表 6 Δy 的统计信息值

单输出回归模型	均值	MAE
线性参数模型 1	0	2.8214

相比线性模型拟合，有 2 个非线性参数的模型拟合效果更好。假设评分与明暗度，红绿色、蓝绿色同样满足 $y \propto ax^2 + bx$ ，解算设计矩阵 X ，求出回归方程为

$$y = -0.2547S^2 + 2.7085S - 0.0565T^2 + 2.1090T + 0.0341L^2 - 5.7554L - 0.0592a^2 - 0.5108a - 0.0243b^2 + 2.7843b - 0.0467dt + 220.5085 \quad (44)$$

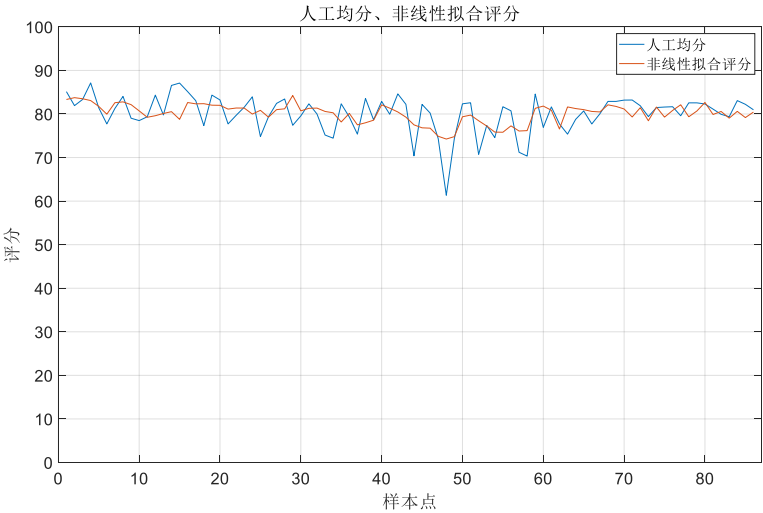


图 13 人工评分的非线性拟合图

人工均分与利用线性回归方程拟合出的结果做差，得到

$$\Delta y = y - y' \quad (45)$$

计算统计信息如下表所示

表 7 Δy 的统计信息值

单输出回归模型	均值	MAE
非线性参数模型 2	0	2.6566

相比于 2 个非线性参数的回归模型，5 个非线性参数的模型，效果更好。为进一步说明对人工指标采用非线性拟合具有更好的拟合效果，将其与线性回归模型进行对比，结果如下：

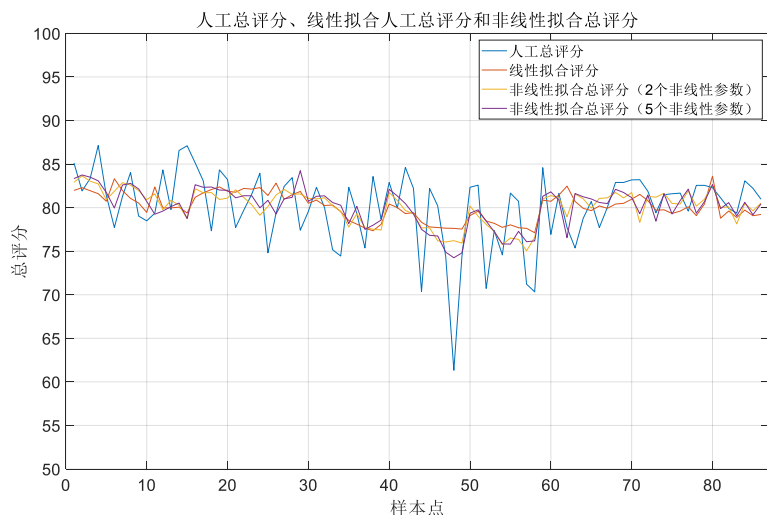


图 14 对人工评分的线性和非线性拟合对比图

表 8 不同回归模型的 Δy 统计值

单输出回归模型	均值
线性参数模型	-0.0185
非线性参数模型 2	0
非线性参数模型 2	0

表 9 不同回归模型的平均绝对误差值

单输出回归模型	MAE
线性模型	2.9697
2 个非线性参数模型	2.8214
6 个非线性参数模型	2.6566

根据上表数据显示，非线性模型拟合效果较线性模型拟合效果好；6 个非线性参数的模型拟合效果比 2 个非线性参数的模型拟合效果较好；6 个非线性参数的模型拟合效果最好。上述拟合模型中， y 是人工评分的均值，根据人工评分下判别优质黄桃标准，下面对色泽，质地，味道分别拟合，得到对应回归方程，可以更加详细的分析物理特性对单个指标（色泽、质地、味道）的影响。在此类模型下认定均分大于等于 80 时，黄桃为优质。

3. 假设黄桃的机械指标对人工评分有如下相关形式：

$$y = X\beta + K \quad (46)$$

$$\text{其中, } y = \begin{pmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ \dots & \dots & \dots \\ y_{m1} & y_{m2} & y_{m3} \end{pmatrix}, X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} & x_{1(n+1)} \\ x_{21} & x_{22} & \dots & x_{2n} & x_{2(n+1)} \\ x_{31} & x_{32} & \dots & x_{3n} & x_{3(n+1)} \end{pmatrix},$$

$$\beta = \begin{bmatrix} S_1 & T_1 & L_1 & a_1 & b_1 & dt_1 & c_1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ S_m & T_m & L_m & a_m & b_m & dt_m & c_m \end{bmatrix}^T, \mathbf{C} \text{ 为常数列向量。}$$

n 为回归模型自变量数目， m 为样本容量。挑选出带皮硬度，TSS， L, a, b, h ，TSS 变化率为线性回归模型自变量，即 $\beta = [S, T, L, a, b, dt, c]^T$ 。

求解系数矩阵，得到拟合方程为

$$\begin{cases} y_s = 0.3845S + 0.4165T + 0.1738L - 0.1640a + 0.3772b - 0.9911h - 0.0468dt + 40.6834 \\ y_z = 0.7265S + 0.0761T - 0.2821L - 0.2659a - 0.1373b - 0.0521h - 0.0511dt + 107.0530 \\ y_w = 0.3326S + 0.6758T - 0.6067L - 0.2825a + 0.0915b - 0.0754h - 0.0746dt + 114.3404 \end{cases} \quad (47)$$

根据计算出的 X ，带入机械测量值，求出线性拟合评分，结果如图所示

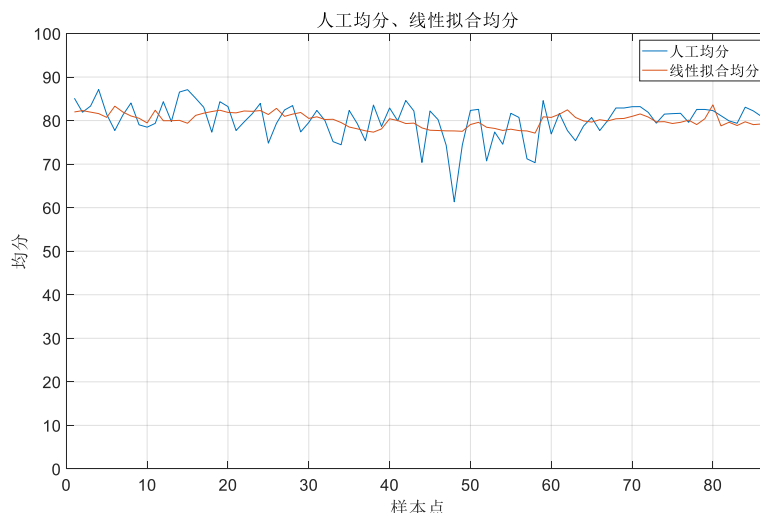


图 15 线性拟合的人工均分图

人工均分与利用线性回归方程拟合出的结果做差，得到

$$\Delta y = y - y' \quad (48)$$

计算 Δy 的统计信息如下表所示

表 10 Δy 的统计信息

名称	色泽差均值	质地差均值	味道差均值
非线性参数模型 2	0	0	0

按下式计算评价指标 MAE

$$MAE(\beta, f) = \frac{1}{m} \sum_{i=1}^m |f(\beta_i) - y_i| \quad (49)$$

MAE 的计算结果如下表所示

表 11 MAE 的值

名称	色泽 MAE	质地 MAE	味道 MAE
线性模型	2.8687	3.2578	3.7839

综上可知，虽然多输出线性回归模型对黄桃人工评分的拟合的效果好于单输出线性回归模型，但多输出线性回归模型对味黄桃人工评分总的味道拟合的结果较差。因此，该模型仍不能很好的满足我们的需求。

4. 为进一步对黄桃人工评分的拟合效果，采用多输出非线性回归模型对其进行拟合，首先假设评分与硬度、TSS 满足

$$y \propto ax^2 + bx \quad (50)$$

即模型自变量为 $\beta = (S^2, S, T^2, T, L, a, b, h, dt, c)^T$ ，求解系数矩阵，得到拟合方程为：

$$\begin{cases} y_s = -0.2514S^2 + 2.6355S - 0.1624T^2 + 5.4422T + 0.2593L - 0.0507a \\ \quad + 0.4691b - 0.0978dt - 11.6684 \\ y_z = -0.4486S^2 + 4.5852S - 0.0059T^2 + 0.2064T - 0.0517L - 0.0470a \\ \quad - 0.0717b + 0.0088dt + 77.2153 \\ y_w = -0.3233S^2 + 3.1616S - 0.0916T^2 + 3.4878T - 0.4647L - 0.1300a \\ \quad + 0.169b - 0.0783dt + 73.2867 \end{cases} \quad (51)$$

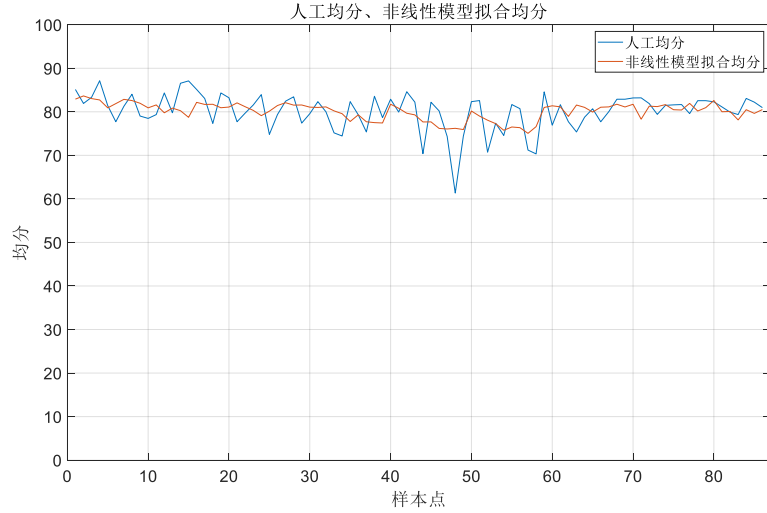


图 16 非线性模型拟合人工均分

计算 Δy 的统计信息如下表所示

表 12 Δy 的统计信息

名称	色泽差均值	质地差均值	味道差均值
非线性参数模型 1	0	0	0

MAE 的计算结果如下表所示

表 13 MAE 的值

名称	色泽	质地	味道
非线性参数模型 1	2.6177	2.9681	3.6218

为进一步提高拟合效果,使用 6 个非线性参数拟合均分,假设评分与明暗度,黄绿色,色调角同样满足 $y \propto ax^2 + bx$,即模型自变量为 $(S^2, S, T^2, T, L^2, L, a^2, a, b^2, b, dt, c)^T$,求解系数矩阵,得到拟合方程为

$$\begin{cases} y_s = -0.1231S^2 + 1.4499S - 0.1071T^2 + 3.6559T + 0.0263L^2 - 4.2118L \\ \quad - 0.1026a^2 - 0.7857a - 0.0269b^2 + 3.3047b - 0.0751dt + 118.9925 \\ y_z = -0.3414S^2 + 3.5359S + 0.0346T^2 - 1.0920T + 0.0576L^2 - 9.6511L \\ \quad - 0.0933a^2 - 0.7313a - 0.0107b^2 + 1.0554b + 0.0170dt + 458.4727 \\ y_w = -0.2994S^2 + 3.1396S - 0.0969T^2 + 3.7631T + 0.0185L^2 - 3.4034L \\ \quad + 0.0184a^2 - 0.0154a - 0.0353b^2 + 3.9930b - 0.0822dt + 84.0602 \end{cases} \quad (52)$$

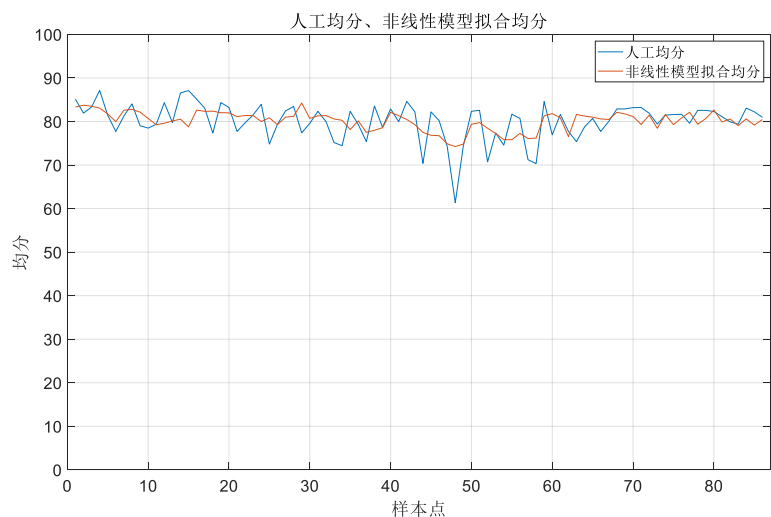


图 17 非线性模型拟合黄桃的人工均分

计算 Δy 的统计信息如下表所示

表 14 Δy 的统计信息

名称	色泽差均值	质地差均值	味道差均值
非线性参数模型 2	0	0	0

MAE 的计算结果如下表所示

表 15 MAE 的值

名称	色泽 MAE	质地 MAE	味道 MAE
非线性参数模型 2	2.5479	2.8251	3.4668

为进一步说明该方法的拟合效果，将其与上节的线性回归模型进行对比，结果如下图所示：

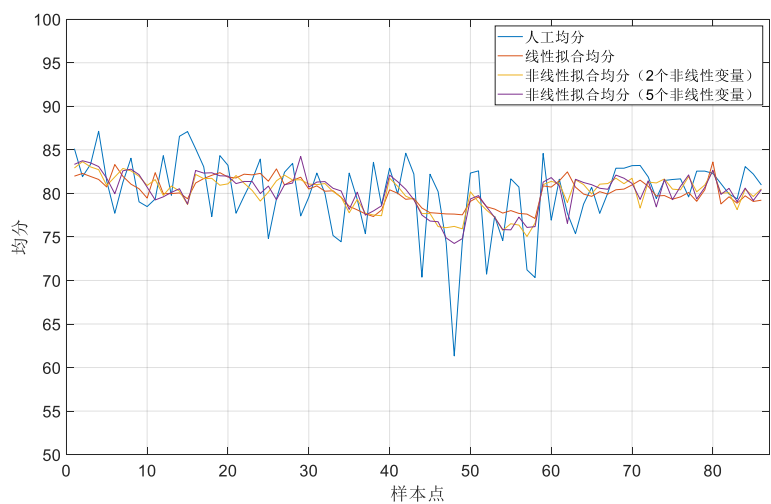


图 18 黄桃人工评分的线性与非相信拟合效果图

表 16 不同回归模型的 Δy 统计值

名称	色泽差均值	质地差均值	味道差均值
线性模型	0	0	0
非线性参数模型 1	0	0	0
非线性参数模型 2	0	0	0

表 17 不同回归模型的平均绝对误差值

多输出回归模型	色泽	质地	味道
线性模型	2.8687	3.2578	3.7839
非线性参数模型 1	2.6177	2.9681	3.6218
非线性参数模型 2	2.5479	2.8251	3.4668

由上表可知，非线性模型拟合效果较线性模型拟合效果好；6 个非线性参数的模型拟合效果比 2 个非线性参数的模型拟合效果较好。在上述模型中，假设单输出拟合模型系数矩阵为 X' ，多输出拟合模型系数矩阵为 X'' ，那么满足

$$x'_{1j} = \frac{1}{3} \sum_{i=1}^3 x''_{ij} \quad (53)$$

则单输出线性模型中的 X' 为：

$$X' = [0.4812 \quad 0.3895 \quad -0.2383 \quad -0.2375 \quad 0.1105 \quad -0.0581 \quad 87.3589] \quad (54)$$

多输出线性模型中的 X'' 为：

$$X'' = \begin{bmatrix} 0.3845 & 0.4165 & 0.1738 & -0.1640 & 0.3772 & -0.0468 & 40.6834 \\ 0.7265 & 0.0761 & -0.2821 & -0.2659 & -0.1373 & -0.0521 & 107.0530 \\ 0.3326 & 0.6758 & -0.6067 & -0.2825 & 0.0915 & -0.0754 & 114.3404 \end{bmatrix} \quad (55)$$

多输出线性模型中的 X'' 列向量相加取平均即为单输出线性模型 X' 。因此，非线性模型拟合效果较线性模型好；根据影响趋势，选择合适的非线性模型，能够有效提高拟合效果。

综上可知，单输出非线性模型拟合的拟合效果与多输出非线性模型拟合效果相当；5 个非线性参数拟合效果强于 2 个非线性参数拟合效果，而 2 个非线性参数拟合效果好于线性参数拟合效果。因此，本文采用带皮硬度，TSS，L，a，b，TSS 变化率七个指标为自变量的非线性评价模型，将所测指标代入，计算色泽，质地，口味三个评分，当三个指标的均分 ≥ 80 时认为样品为‘优质’。与专家经验相比较，本文模型中选择的评价指标更为全面，综合考虑了带皮硬度，TSS，L，a，b，TSS 变化率六个指标。

表 18 不同模型下的优质率。

评价标准	专家经验	人工评分	模型 1	模型 2	模型 3
优质率	89.53%	58.14%	52.33%	65.12%	60.47%
评价标准	专家经验	人工评分	模型 4	模型 5	模型 6
优质率	89.53%	58.14%	52.33%	65.12%	60.47%

（注：模型 1 为单输出线性回归拟合模型，模型 2 为单输出非线性回归拟合模型中非线性参数为去皮硬度和 TSS，模型 3 为单输出非线性回归拟合模型中非线性参数为去皮硬度，TSS，L，a，b；模型 4 为多输出线性回归拟合模型，模型 5 为多输出非线性回归拟合模型中非线性参数为去皮硬度和 TSS，模型 6 为多输出非线性回归拟合模型中非线性参数为去皮硬度，TSS，L，a，b。）

综上，采用黄桃的机械指标值拟合人工评分，采用非线性回归模型的拟合效果强于线性回归模型，因此，我们采用非线性回归模型拟合人工评分，建立了黄桃的机械指标与人工评分的关联关系模型，为下一步的人工评分等级预测奠定一定的理论基础。

4.1.5 2020 年度黄桃品质预测

根据问题 1 建立的模型，依据附件 7 中标黄数据，计算样本的人工评分等级，具体如下：

由于问题 1 中建立的单输出和多输出相对于的回归模型对人工评分等级（“优质”或“非优质”）的评判结果相同，因此，考虑模型的复杂度和计算效率，当只考虑人工评分等级时，我们选用单输出的三个回归模型计算附件 7 中 9 个样本的人工评分，结果如下表所示：

表 19 三个回归模型估计的人工评分值

样本	回归模型 1	回归模型 2	回归模型 3
A	83.3992	81.1851	81.1938
B	79.8059	77.9003	78.3350
C	79.4052	80.4578	81.4546
D	82.0296	83.6671	83.8739
E	78.8496	78.2526	77.8642
F	77.2199	75.7336	75.4526
G	79.4037	81.6996	81.7469
H	78.4958	78.5871	80.0615
I	79.4379	78.6388	78.7511

依照问题中所述，对于人工评分，当色泽、质地和味道这三个指标均分 ≥ 80 分时认为样品“优质”这一评判标准，依据估计的人工评分值对黄桃的品质进行评判。经计算，回归模型 3 的平均绝对误差值最小，对人工评分值的估计精度更高，因此，选用回归模型 3 的估计结果对黄桃品质进行评判，其结果如下所示：

表 20 估计的人工评分等级

样本	品质	样本	品质	样本	品质
A	优质	D	优质	G	优质
B	非优质	E	非优质	H	优质
C	优质	F	非优质	I	非优质

根据计算，人工评分 86 个样本中，优质评分有 50 个，计算人工评价下的黄桃优质率在 58.14%，在非线性回归模型 2，即回归模型中有 5 个非线性参数的模型中，判定有 5 个为优质黄桃，比例为 55.56%，与总体比例相符。

综上，我们利用相关性分析，得出带皮硬度与去皮硬度有极强正相关，a 和 h 有极强负相关，b 和 C 有极强正相关。因此构建回归模型时，选择带皮硬度，TSS，L，a，b，TSS 下降率为机械指标来拟合人工评分。结果显示，黄桃的预测均分与人工均分有较好的拟合度。但是实际中由于存在特殊情况，比如带皮硬度为与去皮硬度相差比较大，比如样本 H，带皮硬度 1.73，去皮硬度 0.33，那么根据我们拟合的模型来评分，就会判为优质，但是实际人工评分判为非优质。因此在机械指标的选取时，应该充分考虑各项影响因素。

4.2 黄桃贮藏预测模型

4.2.1 贮藏预测模型

1. 按照 4.1 节的流程，分析黄桃机械指标和人工评分随不同时间下采样次数的变化趋势，为保证分析趋势的一致性，分别绘制在同一采摘批次下，机械指标、人工评分的测量值与贮藏时间的四分位关系图，具体如下图所示：

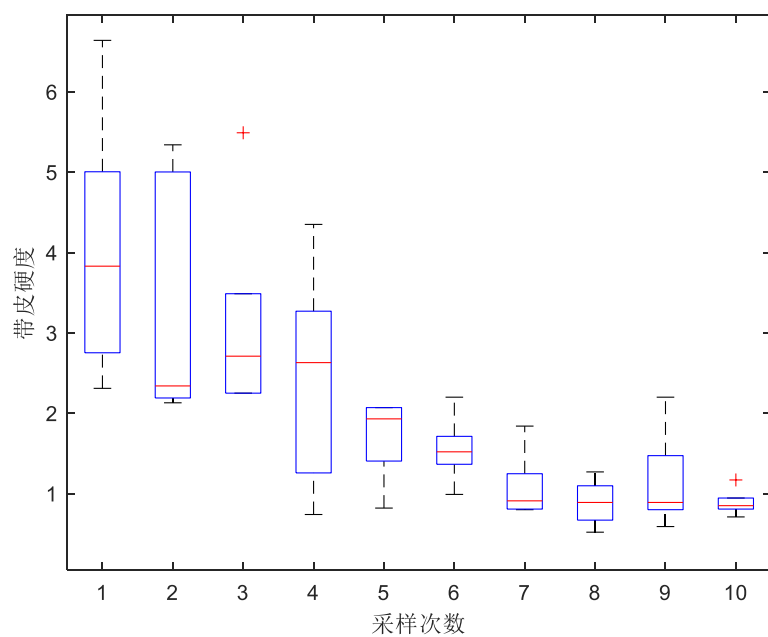


图 19 带皮硬度在不同采用次数下的变化趋势图

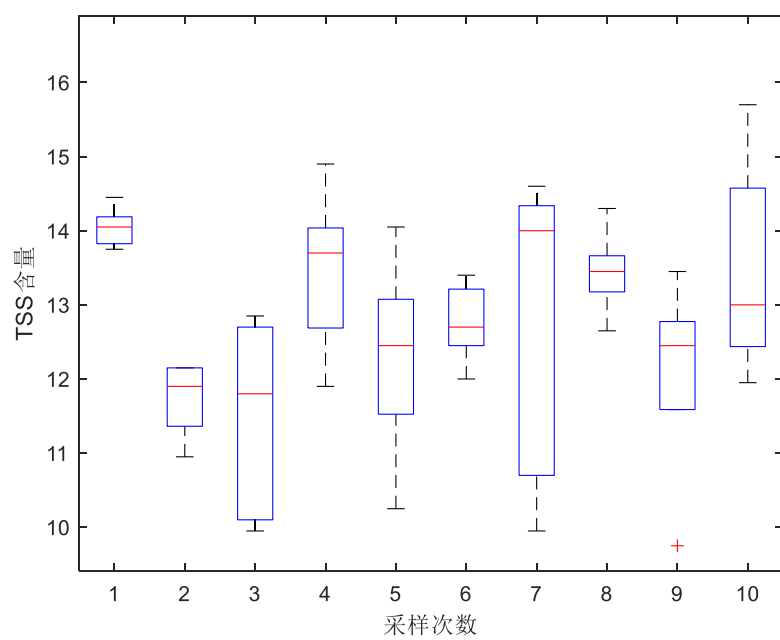


图 20 TSS 在不同采用次数下的变化趋势图

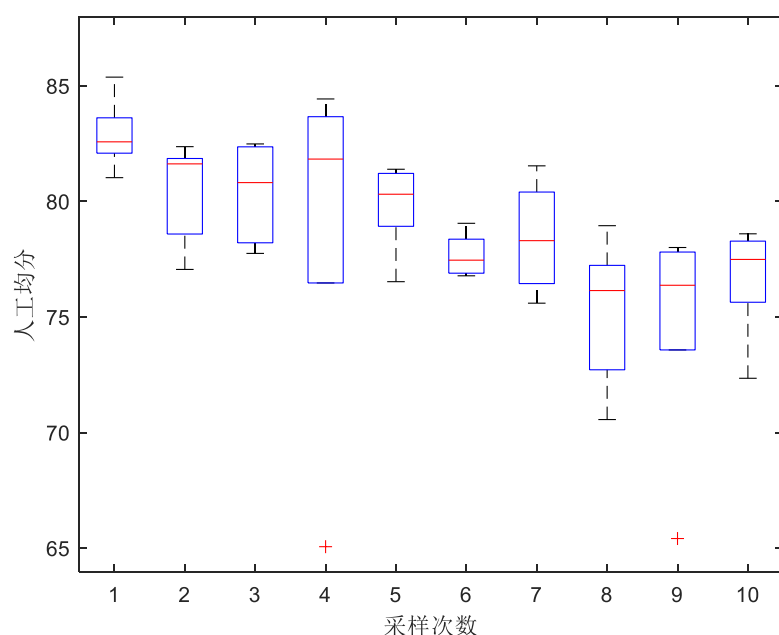


图 21 人工均分在不同采用次数下的变化趋势图

由上图可知，在不同时间的采样次数下，带皮硬度总体呈现下降趋势，TSS 呈现先下降后上升再下降的趋势，人工均分呈现先下降后上升的趋势。黄桃在采摘后的贮藏期间会出现两次呼吸高峰以及一次生长激素释放高峰。呼吸高峰出现的越早越难以贮藏。黄桃采摘后，随着黄桃逐渐分解，黄桃的硬度也开始下降，因此带皮硬度总体呈现下降趋势；黄桃采摘后依靠自身的养分维持生命活动，使得黄桃的 TSS 含量出现下降，但随着生长激素的释放，使得黄桃中的淀粉转化为糖，因此 TSS 呈现先下降后上升再下降的趋势。由于采摘后黄桃硬度以及 TSS 的变化趋势，人工均分也呈现先下降后上升的趋势。

2. 依据采摘后黄桃的变化趋势，根据附件 1 到附件 7 中的实验数据，问题一已得到了黄桃品质分析定量模型，将附件 1-2 中的 2018 年的机械指标数据与附件 4-5 中 2019 年的实验数据代入黄桃品质分析定量模型中，可得到每个样本对应的黄桃品质得分，绘制黄桃样本的品质得分与时间的散点图，并分别对冷库 1℃和冷柜 8℃下黄桃品质随贮藏时间的关系进行二次和三次回归分析，为了减少样本差异对拟合函数影响，实验中需控制变量，黄桃质量应相近，分别对 1℃和 8℃黄桃评分数据进行中心化处理，在最终的拟合结果上加上整体（即 1℃和 8℃）的评分均值，得到黄桃品质随时间的变化趋势，具体如下：

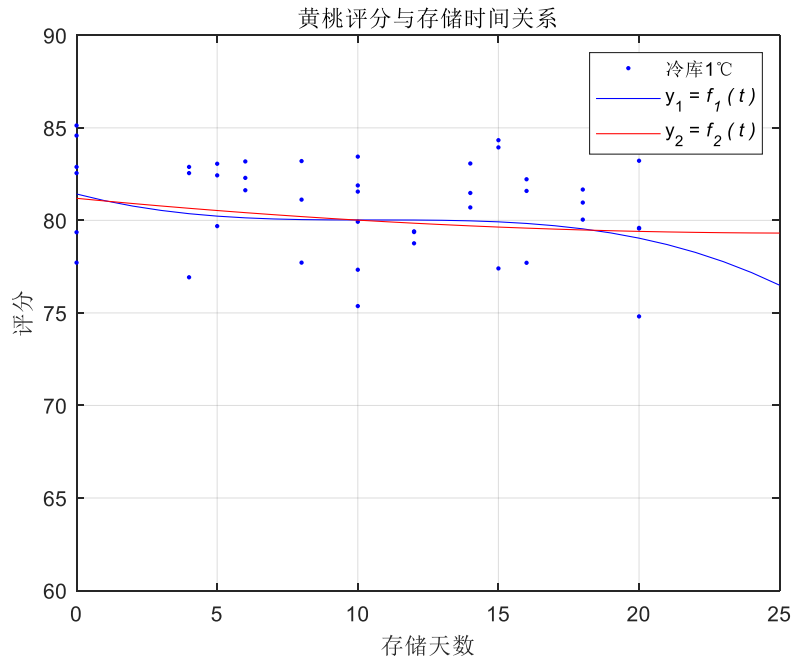


图 22 冷库(1℃)下黄桃品质得分随时间的变化趋势图

在冷库(1℃)下的条件下，分别用非线性二次拟合和非线性三次拟合对冷库(1℃)下黄桃品质得分随时间的变化趋势进行拟合，拟合模型如下式所示：

$$\begin{aligned} y_1 &= -0.001179x^3 + 0.03749x^2 - 0.3974x + 81.43 \\ y_2 &= 0.002843x^2 - 0.1462x + 81.97 \end{aligned} \quad (56)$$

二次拟合的 RMSE（均方根误差）为 15.81，三次拟合的 RMSE（均方根误差）为 15.73。在大约第 20 天，黄桃品质降至非优质，黄桃在优质范围内时，品质得分的最大值为 85 分。从图中可看出，采用非线性三次拟合黄桃品质曲线总体呈现出先下降后上升再下降的趋势，这与黄桃在采摘后的贮藏期间在两次呼吸高峰以及一次生长激素释放高峰所呈现出的黄桃品质趋势相一致。因此，我们选用非线性三次拟合模型作为冷库(1℃)下黄桃贮藏预测模型。

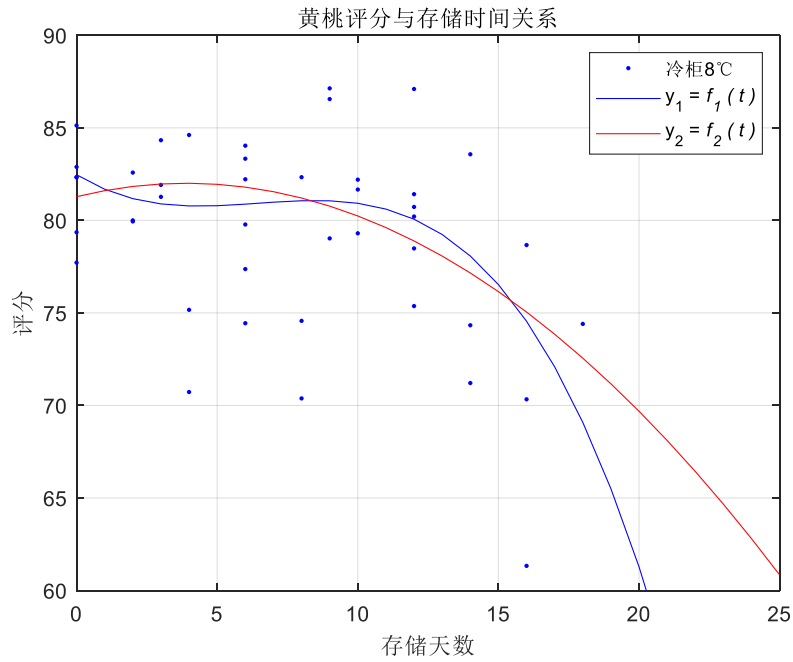


图 23 冷库(8℃)下黄桃品质得分随时间的变化趋势图

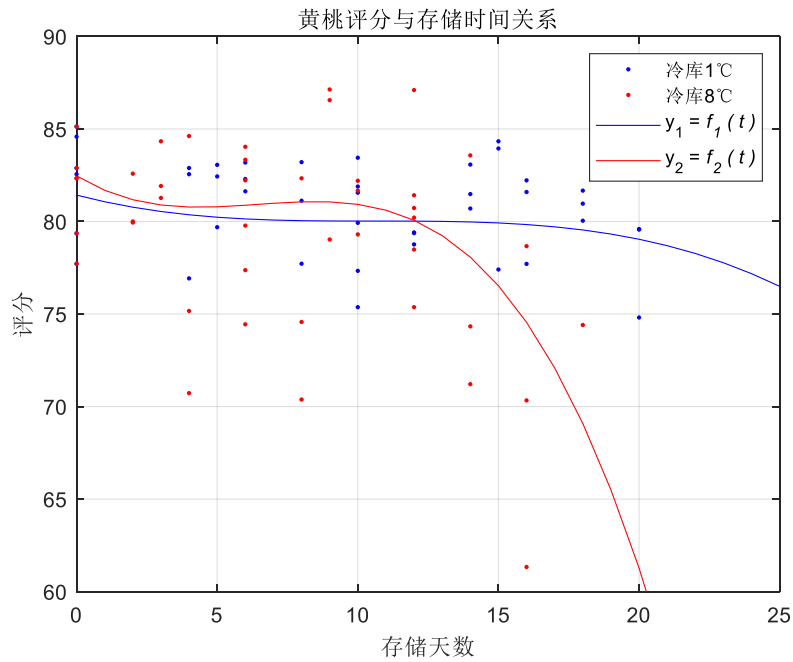


图 24 不同温度下黄桃评分与存储时间关系图

在冷库(8℃)下的条件下，分别用非线性二次拟合和非线性三次拟合对黄桃品质得分随时间的变化趋势进行拟合，拟合模型如下式所示：

$$\begin{aligned} y_1 &= -0.008443x^3 + 0.1627x^2 - 0.9373x + 82.47 \\ y_2 &= -0.047563x^2 + 0.3714x + 80.5 \end{aligned} \quad (57)$$

二次拟合的 RMSE（均方根误差）为 30.64，三次拟合的 RMSE（均方根误差）为 29.85。在大约第 11 天，黄桃品质降至非优质。从图中同样可看出，采用非线性三次拟合黄桃品质曲线趋势与采摘后黄桃品质趋势相一致。因此，我们选

用非线性三次拟合模型作为冷柜(8℃)下黄桃贮藏预测模型。另外,从不同温度下黄桃评分与存储时间关系图中可看出,冷柜 8℃下黄桃品质变化趋势较冷库 1℃下更显著。这与普遍认为的在冷库中贮藏风味有所下降但保质期长,冷柜 8℃中贮藏风味佳但保质期较短相一致。

4.2.2 试验次数设计

方差分析是英国统计学家费希尔于 20 世纪 20 年代提出的一种统计方法,用于分析试验条件的改变是否会对所失心的数量指标有影响^[6]。即在过程中,哪些因素的影响较大,哪些因素的影响较小,以及影响较大的因素处于何种状态时,对实验结果影响较大,这就是方差分析所要解决的问题。因此我们采用方差分析的方法研究黄桃贮藏环境和时长对黄桃品质的影响以及更加科学地指导实验设计。

在方差分析中,我们所关心的试验结果称为试验指标,在试验中对试验指标可能产生影响的因子称为因子,因子常用符号 $A, B, C, \dots$, 而把因子所处的状态称为水平,用符号 $A_1, A_2, \dots, A_r$ 表示因子 A 的 r 个不同水平。设在单因子试验中因子 A 有 r 个不同水平 $A_1, A_2, \dots, A_r$, 在第 i 个水平 A_i 下的试验指标记为 $X^{(i)}$, 即有 r 个总体,在水平 A_i 下进行了 n_i 次独立试验,得到的样本记为

$$X_{i1}, X_{i2}, \dots, X_{in_i} \quad (i = 1, 2, \dots, r) \quad (58)$$

假设

1) 不同水平 A_i 下的试验指标 $X^{(i)}$ 服从正态分布 $X^{(i)} \sim N(\mu_i, \sigma^2)$ ($i = 1, 2, \dots, r$);

2) 这 r 个正态总体的方差相等,即方差具有齐性;

3) r 个样本相互独立。

方差分析的目标就是在上述假设下,检验假设

$$H_0: \mu_1 = \mu_2 = \dots = \mu_r \quad (59)$$

记

$$\varepsilon_{ij} = X_{ij} - \mu_i, i = 1, 2, \dots, r \quad (j = 1, 2, \dots, n_i)$$

ε_{ij} 为在水平 A_i 下,第 j 次试验的随机误差,是不可观察的随机变量,由

上述假设可知, $\varepsilon_{ij} (j = 1, 2, \dots, n_i, i = 1, 2, \dots, r)$ 相互独立,且 $\varepsilon_{ij} \sim N(0, \sigma^2)$ 。 μ_i 就是水平 A_i 下试验数据的理论均值,反映了水平 A_i 对试验指标的影响。由此得到方差分析的数学模型

$$\begin{cases} X_{ij} = \mu_i + \varepsilon_{ij} \\ \varepsilon_{ij} \sim N(0, \sigma^2) \\ \varepsilon_{ij} \text{相互独立}, j = 1, 2, \dots, n_i, i = 1, 2, \dots, r. \end{cases} \quad (60)$$

为了分析方便,记

$$n = \sum_{i=1}^r n_i, \quad \mu = \frac{1}{n} \sum_{i=1}^r n_i \mu_i, \quad \delta_i = \mu_i - \mu \quad (i = 1, 2, \dots, r) \quad (61)$$

称 μ 为理论总均值, δ_i 为因子 A 在第 i 个水平 A_i 下的效应, 反映了因子 A 在第 i 个水平 A_i 对试验指标的作用。易知有

$$\sum_{i=1}^r n_i \delta_i = 0 \quad (62)$$

由方差分析的数学模型可以表示为如下形式

$$\begin{cases} X_{ij} = \mu + \delta_i + \varepsilon_{ij} \\ \varepsilon_{ij} \sim N(0, \sigma^2) \\ \varepsilon_{ij} \text{相互独立}, j = 1, 2, \dots, n_i, i = 1, 2, \dots, r, \\ \sum_{i=1}^r n_i \delta_i = 0 \end{cases} \quad (63)$$

由此可见, 方差分析模型实际就是一个多元正态线性模型, 检验假设 H_0 等价于

$$H_0: \delta_1 = \delta_2 = \dots = \delta_r = 0 \quad (64)$$

构造用于检验假设 H_0 的统计量. 记

$$\begin{aligned} \bar{X}_i &= \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, S_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 \quad (i = 1, 2, \dots, r), \\ \bar{X} &= \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^{n_i} X_{ij} = \frac{1}{n} \sum_{i=1}^r n_i \bar{X}_i \\ S_T &= \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 \\ S_e &= \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 = \sum_{i=1}^r n_i S_i^2 \\ S_A &= \sum_{i=1}^r \sum_{j=1}^{n_i} (\bar{X}_i - \bar{X})^2 = \sum_{i=1}^r n_i (\bar{X}_i - \bar{X})^2 \end{aligned} \quad (65)$$

其中, $\bar{X}_i$ 称为第 i 个样本的样本均值, S_i^2 称为第 i 个样本的样本方差, $\bar{X}$ 称为样本总均值, S_T 称为总离差平方和, 反映了所有试验数据的波动, S_e 称为误差平方和, 主要反映了由于随机误差而引起的波动, 也称为组内离差平方和, S_A 称为因子 A 的效应平方和, 主要反映了因子 A 的不同水平所产生的效应而引起的波动。

1. 冷柜 8℃ 下试验次数设计

由方差分析的基本条件:

1) 各样本是相互独立的随机样本, 均服从正态分布;

2) 各样本方差相等，即满足方差齐性。

因此首先对样本进行方差齐性检验：

表 21 方差齐性检验

Levene 统计量	df1	df2	显著性
7.608	8	425	<0.01

从上表中看出，显著性 $p < 0.1$ ，说明方差齐性，因此，可以对黄桃品质与贮藏时间的关系进行单因子方差分析。

表 22 冷柜 8 度下不同贮存天数对黄桃品质的影响

天数	N	$\bar{x} \pm s$
0	63	81.44 ± 3.96
2	27	80.74 ± 3.76
3	32	82.30 ± 4.96
4	35	77.28 ± 8.15
6	66	80.70 ± 5.19
8	23	76.33 ± 6.28
9	38	84.35 ± 5.42
10	28	80.86 ± 3.05
12	63	80.58 ± 4.85
14	27	75.80 ± 7.61
16	17	71.17 ± 8.63
18	9	74.41 ± 6.24
20	6	66.94 ± 5.72

原假设 H_0 ：不同贮存天数下黄桃的品质不存在显著差异。

上表描述性分析了在不同贮存天数下黄桃品质评分的均值和方差。由于偏差平方和与自由度的比值服从 F 分布，由表中可知 F 分布的观测值为 11.997，其对应的 p 值 < 0.01 。因此，在显著性水平 $\alpha = 0.1$ 的前提下，概率 p 值小于 α ，故拒绝原假设，即认为不同的贮存天数间至少有两组存在显著性差异。

根据不同的贮存时间进行分组，设显著性水平为 $\alpha = 0.1$ ，分析组内与组间是否有至少两个组间存在显著性差异：

表 23 组间与组内的显著性差异

	平方和	df	均方	F	显著性
组间	2419.877	8	302.485	11.997	.000
组内	10715.725	425	25.213		
总数	13135.603	433			

从表中说明 13 个组间至少有两个组存在显著性差异，即不同的贮藏时间对黄桃的品质有显著的影响。

下面，通过对两两贮藏天数之间黄桃品质得分均值进行比较，可以得到当显著性水平 $\alpha = 0.1$ 时，贮藏天数是否会造成品质间的显著性差异：

表 24 不同贮藏天数下黄桃品质间的显著性差异

(I) 天数	(J) 天数	均值差 (I-J)	标准差	显著性
0	2	.69515670	1.20338138	0.564
	3	-.86358173	1.13566628	0.448
	4	4.15970696	1.10291726	0.000
4	6	-3.42709235	1.09392838	0.975
	8	.94768806	1.40426880	0.003
6	8	4.37478041	1.26675699	0.001
8	9	-8.01711162*	1.38211249	0.000
9	10	3.48582567	1.30297278	0.008
10	12	.28476946	1.32230418	.830
	14	5.06966490*	1.57029197	.002
14	16	4.61855483*	1.80251945	.011
16	18	-3.23583878	2.39994297	.180
	20	4.22712418	2.76454584	.128

根据上文冷柜 8℃下黄桃保质期为 11 天，因此，只需考虑在 11 天之内进行有损测试的次数。由上表可以得出，至少应进行 6 次有损测试。

2. 冷库 1℃下试验次数设计：

首先对数据进行方差齐性检验，具体如下表所示：

表 25 方差齐次性性检验

Levene 统计量	df1	df2	显著性
3.299	5	243	.007

从中看出，显著性 $p < 0.1$ ，说明方差齐性，满足方差分析的前提。

表 26 不同贮藏天数下黄桃品质间的显著性差异

天数	N	$\bar{x} \pm s$
0	62	81.6344086 $\pm$ 4.17
4	27	80.7901235 $\pm$ 3.71
5	35	81.5333333 $\pm$ 3.61
6	27	82.3703704 $\pm$ 2.20
8	37	80.6576577 $\pm$ 3.49
10	61	80.1311475 $\pm$ 4.57
12	16	79.1250000 $\pm$ 3.42
14	19	81.0701754 $\pm$ 2.49
15	41	81.8536585 $\pm$ 5.33
16	17	79.7647059 $\pm$ 4.97
18	17	80.9019608 $\pm$ 3.05
20	55	78.9818182 $\pm$ 4.60

经过上述方差分析，可得出每组贮藏天数的样本量 N，组内均值及标准差。

下表的方差分析得到组间的 F 统计量的测量值以及显著性水平 p 。在显著性水平 $\alpha = 0.1$ 下， $p > 0.1$ ，说明组间的黄桃品质数据不存在显著性差异。

表 27 黄桃品质数据的显著性差异

	平方和	df	均方	F	显著性
组间	134.938	5	26.988	1.783	.117
组内	3678.035	243	15.136		
总数	3812.973	248			

由上文预测的黄桃在冷库 1°C 下的保质期为 20 天，只需考虑在 20 天之内进行有损测试的次数。由上表可以得出，至少应进行 4 次有损测试。

表 28 不同贮藏天数下黄桃品质间的显著性差异

(I) 天数	(J) 天数	均值差 (I-J)	标准差	显著性
0	4	.84428514	.89706116	0.348
	5	.10107527	.82254710	0.902
	6	-.73596177	.89706116	0.413
	8	.97675094	.80821269	0.228
	10	1.50326106*	.70161129	0.033
10	12	1.00614754	1.25318880	0.423
	14	-.93902790	1.17219355	0.424
	15	-1.72251100	.90103008	0.057
	16	.36644166	1.22364087	0.765
	18	-.77081324	1.22364087	0.529
	20	1.14932936	.82961918	0.167
15	10	1.72251100	.90103008	0.057
	12	2.72865854*	1.31516899	0.039
	14	.78348310	1.23823463	0.528
	16	2.08895265	1.28704478	0.106
	18	.95169775	1.28704478	0.460
	20	2.87184035*	.92057334	0.002

5 参考文献

- [1] 吴孟达, 数学建模教程[M], 北京: 高等教育出版社, 2011。
- [2] 吴翊, 数学建模的理论与实践[M], 湖南长沙: 国防科技大学出版社, 244-249, 1999。
- [3] 宋超. 基于 Sogou 日志的个性化信息检索分析与建模[D]. 黑龙江: 哈尔滨工业大学, 2010。
- [4] 彭荣, 凌莉. 两类健康评价指标的一致性研究[J]. 数学的实践与认识, 38(23):5, 2008。
- [5] 吴翊, 概率论与数理统计, 北京: 高等教育出版社, 305-320, 2016。
- [6] 王松桂. 线性统计模型: 线性回归与方差分析[M], 北京: 高等教育出版社, 20-80, 1999。