STATS 604 Project 3 Preanalysis Plan

Khoa Do, Surtai Han, Jialin He, Carlyle Morgan

2025-10-23

Introduction

In this project, we attempt to analyze various procedures for preserving the freshness of cilantro. Namely, we will attempt to measure if the following have any effect:

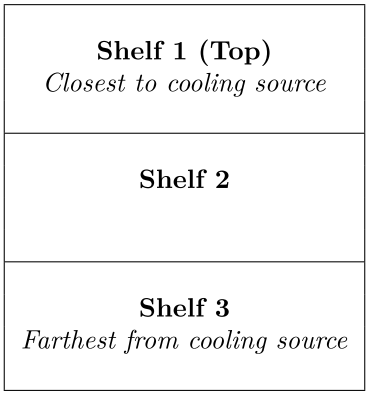
- Keeping cilantro in the fridge versus outside
- Keeping cilantro in a plastic bag versus not
- For refrigerated cilantro, keeping cilantro close to the cooling source at the top of the fridge versus farther away.

We hypothesize that cilantro kept in a fridge in a bag and far away from the cooling source will remain fresh for longer.

Methodology

Experimental setup

For convenience and budgetary concerns, we decided to use a group mate's mini-fridge to apply the treatment. A diagram of this mini-fridge is shown below:



Mini-Fridge layout

Randomization Scheme

From our 40 cilantro stalks, we assigned a group of 5 cilantro stalks to one of eight treatment groups at random. The treatment plans for each of the eight groups is as follows:

Group	Location	Bagged
1	Outside the fridge	No
2	Outside the fridge	Yes
3	Fridge shelf 1	No
4	Fridge shelf 1	Yes
5	Fridge shelf 2	No
6	Fridge shelf 2	Yes
7	Fridge shelf 3	No
8	Fridge shelf 3	Yes

Data Collection

The weight of each set of sprigs is measured in grams daily (thus, 8 measurements per day) using an AmazonBasic scale with dubious precision. One image of each set is taken daily under as similar lighting conditions as possible using the same smart phone camera.

Data Processing

To quantify cilantro freshness from daily images, we use computer vision techniques with HSV (Hue, Saturation, Value) color space analysis. Each image is converted from RGB to HSV color space, which separates color information (hue) from intensity and lighting effects. We apply a calibrated HSV threshold mask (H: 35-85, S: 40-255, V: 40-255) to segment cilantro pixels from the background, producing a binary segmentation.

After segmentation, we use contour detection to identify exactly 5 sprigs per image. Contours are identified using the RETR_EXTERNAL method, which finds outer boundaries of connected regions. To ensure we detect exactly 5 sprigs, we use an adaptive threshold approach: starting with a minimum area threshold of 10,000 pixels, we iteratively relax this threshold (reducing by 30% each iteration) until at least 5 contours are detected. Once found, the 5 largest contours are selected and assigned unique identifiers (1-5).

From each individual sprig, we extract the following metrics as indicators of freshness:

- **Mean Hue**: Average green color (0-179 scale)
- **Mean Saturation**: Color intensity (0-255 scale)
- **Mean Value**: Brightness level (0-255 scale)

Figure 1 illustrates the HSV segmentation and sprig identification process applied to a sample cilantro image. The original image (left) is converted to HSV color space and thresholded to create a binary mask (center), where white pixels represent detected cilantro and black pixels represent background. The adaptive contour detection algorithm then identifies the 5 largest sprigs (right), with each sprig outlined in a different color and labeled with a unique identifier (1-5). HSV statistics are calculated separately for each of the 5 selected sprigs.

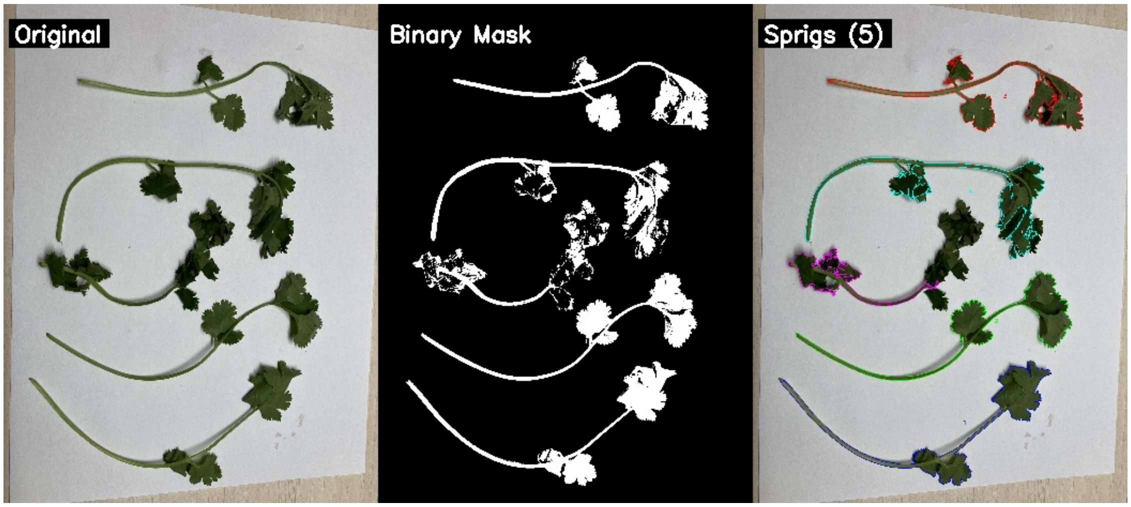


Figure 1: HSV segmentation and sprig identification pipeline showing: (left) original cilantro image, (center) binary mask from HSV thresholding, and (right) the 5 largest sprigs identified via adaptive contour detection, each outlined in a distinct color and labeled with a unique identifier (1-5) for per-sprig tracking.

Testing Procedure

EDA

Following the data processing, we will have a time-series dataset where each of the 40 replicates has 3 individual sprig measurements. To visualize the decay of freshness over time, we will first generate a heat map where the Y-axis is the 8 different experimental conditions, the X-axis is Day, and the cell's color indicates the average Mean Saturation (or Mean Hue) for that condition on that day. This graph can provide a clear comparison of which conditions lose their freshness more rapidly. Following this, to compare different conditions, we will generate summary box plots of the Mean Saturation values. These box plots will be organized by our 8 experimental conditions every three days after the start of our experiment.

Permutation Test Approach

We will then conduct a series of permutation tests:

- A permutation test on whether different layers of the fridge (which presumably have different temperature and moisture) have different effects on preserving freshness. If so, we may choose to pool all data from the fridge together in subsequent tests. Explicitly, we will convert HSV values to ranks among all sprigs in the fridge and then compare the distribution of ranks.
- A permutation test on whether putting cilantro in a plastic bag has an effect toward freshness. This will be conducted on the 30 sprigs inside the fridge. We will use conditioning if data from 3 layers differ and could not be pooled together. This will also be done using a rank testing procedure.
- A permutation test on whether in fridge and out of fridge sprigs differ in freshness. The ranks will be recalculated for all 40 sprigs instead of the 30 sprigs as before, but the overall testing procedure will remain the same.

Model Based Scheme

Regression Model

First, to compare the decay speed in the different 8 conditions, we will fit 8 individual simple linear regression models.

For each condition i, we will model its Mean Saturation (or other selected metric) as a function of Day:

$$Mean\ Saturation_i = \beta_{0i} + \beta_{1i} \times Day + \epsilon_i$$

From each model i, we use the data during a week and extract the slope coefficient β_{1i}, which represents the rate of decay for each condition. A slope near zero indicates slow decay (high freshness), while a large negative slope indicates rapid freshness loss. We can get a overview on β_{1i} of which condition keeps freshness better.

ANOVA on Slope

Next, we will fit a Two-Way ANOVA to the 8 slope values:

$$Slope\ Location + Bag + (Location \times Bag)$$

We use the F-test in the above model, which represents the ratio of the variance explained by that factor to the residual variance. We will first test the interaction term of Location and Bag. Here the null hypothesis is that the effect of the bag on the slope is the same for all locations. The result of this F-test determines the following steps. If p<0.05, we conclude there is a significant interaction between the two factors. If the interaction is not significant, we can remove the interaction term and proceed to the simpler model. For the effect of Location, where the null hypothesis is that the mean slopes are equal across all four locations. If this F-test is significant, we can conclude that locations affect the freshness. Finally, we will test the main effect of Bag, where the null hypothesis is that the mean slopes are equal for the “In Bag” and “Out of Bag” conditions. If the F-test is significant, we can conclude that there is effect of bagging in keeping herbs fresh.