Final Project Report

Stephen Ling (jling9), Yicen Yang (yyang643), Chenying Mi (cmi), Xinyi Wang (xwang2454)

Introduction

Although few people pay attention, preserving food is very important. For example, the Covid Lockdown in Shanghai leads to serious supply chain problems so that many Shanghai residents are "running out of food". Under such a situation, it is important to avoid food waste caused by decay. Imagine a Shanghai resident who buys enough vegetables before the lockdown. However, it will be a tragedy if half of the vegetable gets spoiled because the resident does not know how to preserve vegetables. We want to do something that can help Shanghai people overcome the food shortage problem. To be specific, we want to provide some suggestions on how to preserve food for a longer time. So, in this research, we want to figure out what environmental factors will influence the time for vegetable spoilage and how these environmental factors influence the time for vegetable spoilage.

In fact, we believe 10 factors may affect the time for vegetable spoilage. In specific, these factors can be categorized into external factors (environmental factors) and internal factors. The external factors are *Air Contact Level (Oxygen Level)*, *Temperature*, *Humidity Level*, *Light Condition*, *Bacterial Type*, and *Fungi Type*. The internal factors are *Vegetable Type*, *Surface Area*, *Chemical Residuals on Vegetables*, *Pests hiding inside Vegetables*. The Fishbone diagram below illustrates whether these factors will be controlled, blocked, or held constant during the experiment.

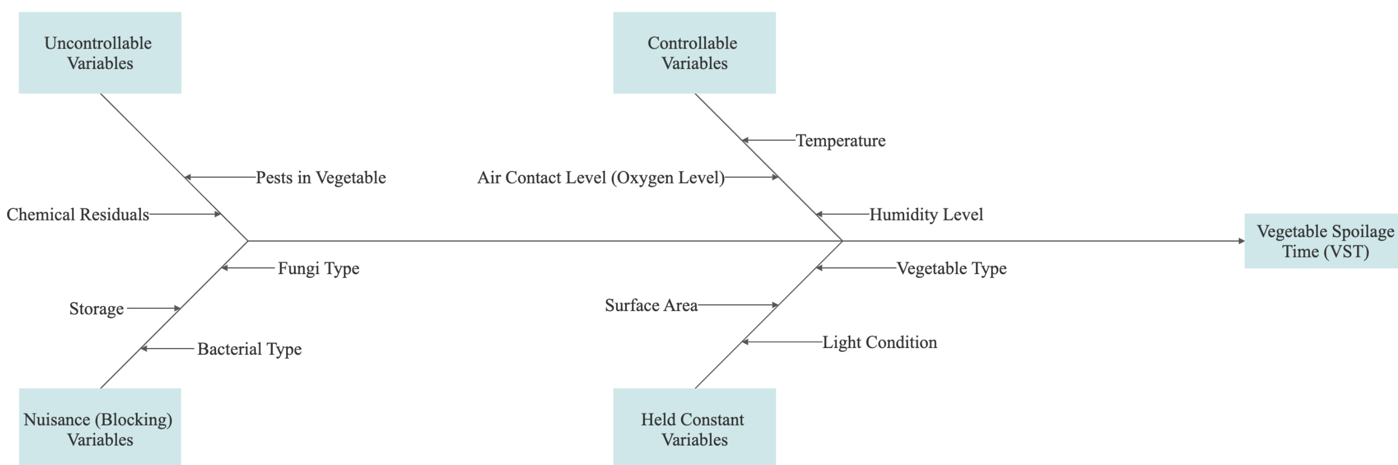

Figure 1

Through this experiment, we are trying to find: What combination of *temperature*, *humidity*, and *air contact level* will make the slice with the *same surface area* from the *same type of vegetable* under the *same light condition* have the smallest spoiled area in 6 days, while *blocking by the environment condition* (bacterial, fungi, storage)?

Methods

For this experiment, we plan to measure the spoiled area of the lettuce leaves preserved under different temperatures, humidity, and air contact level. However, there are some points we must consider when designing the experiment and choosing the statistical model: 1. hard to precisely measure "air contact" and "moisture"; 2. possible interactions between factors; 3. every group member should perform the experiment, but the experiment's environment differs by person.

Taking these points into consideration, we decide to conduct a $\mathbf{2^3}$ ***Factorial Design with Blocking*** for the following reasons. First, factorial design makes it possible to collect data at two levels (low & high) instead of collecting data at a specific air contact/moisture value. Also, the factorial design enables us to investigate both the main effect and interaction between factors. Finally, blocking can remove the effects of the nuisance variable, which is the environmental condition in this case.

In this 2^3 Factorial Design with Blocking, the ***response variable*** is the spoiled area of the lettuce leaf, which is measured in mm^2 (square millimeter). The **3 *factors*** used in the experiment are temperature (whether preserved in the fridge), moisture (whether preserved in a wet environment), and air contact level (whether air is blocked). Each factor is not measured in units but has two levels: low (-1) and high (+1).

Since we are conducting a 2^3 Factorial Design with Blocking, we will use the ***effect model*** for a factorial experiment as our statistical model:

- $y_{ijk} = \mu + \tau_i + \beta_j + \gamma_k + (\tau\beta)_{ik} + (\beta\gamma)_{jk} + (\tau\beta\gamma)_{ijk} + \delta_l + \epsilon_{ijkl}$
  {i = 1, 2, ..., a}, {j = 1, 2, ..., b}, {k = 1, 2, ..., c}, {l = 1, 2, ..., n}
- $\boldsymbol{\tau_i}$: main effect of *Temperature*; $\boldsymbol{\beta_j}$: main effect of *Humidity*; $\boldsymbol{\gamma_k}$: main effect of *Air Contact Level.*
- $\boldsymbol{(\tau\beta)_{ik}}$: interaction between *Temperature* and *Humidity*; $\boldsymbol{(\beta\gamma)_{jk}}$: interaction between *Humidity* and Air Contact *Level*;
- $\boldsymbol{(\tau\beta\gamma)_{ijk}}$: interaction between *Temperature*, *Humidity*, and *Air Contact Level*; $\boldsymbol{\delta_l}$: block effect; $\boldsymbol{\epsilon_{ijkl}}$: Random Error

After performing the experiment, we will conduct ***hypothesis tests*** to find factors that have statistically significant effects on the response variable (spoiled area of the lettuce leaf). Then we will perform the ***Fisher LSD test*** to figure out how factors, which have statistically significant effects, influence the response variable (whether the factors at either level will increase or decrease the spoiled area of lettuce leaf).

Hypothesis (7):

$H_0: \tau_1 = \tau_2 = \cdots = \tau_a = 0$	$H_a: \tau_i \neq 0$ for at least one i
$H_0: \beta_1 = \beta_2 = \cdots = \beta_b = 0$	$H_a: \beta_j \neq 0$ for at least one j
$H_0: \gamma_1 = \gamma_2 = \cdots = \gamma_c = 0$	$H_a: \gamma_k \neq 0$ for at least one k
$H_0: (\tau\beta)_{ij} = 0$ for all i, j	$H_a: (\tau\beta)_{ij} \neq 0$ for at least one pair of i, j
$H_0: (\tau\gamma)_{ik} = 0$ for all i, k	$H_a: (\tau\gamma)_{ik} \neq 0$ for at least one pair of i, k
$H_0: (\beta\gamma)_{jk} = 0$ for all j, k	$H_a: (\beta\gamma)_{jk} \neq 0$ for at least one pair of j, k
$H_0: (\tau\beta\gamma)_{ijk} = 0$ for all i, j, k	$H_a: (\tau\beta\gamma)_{ijk} \neq 0$ for at least one pair of i, j, k

Experimental Procedure

Each person is assigned the same type of lettuce based on the result of a random number generator that generates 3 different integers between 0 and 2. The result of the random number generator is “2, 1, 0” which suggests Stephen should pick the first lettuce from the left in the bag, Chenying should pick the middle lettuce in the bag, and Lucia should pick the last lettuce from the left in the bag. Each person cut eight squares with a side length of five centimeters (5cm) from the lettuce (*Figure 2*). Then each person applies a “treatment” on each lettuce square (8 treatments in total). In specific, each person labels his/her 8 lettuce squares from 0 to 7, and a random number generator will generate 8 integers from 0 to 7. For example, the sequence generated by Stephen is “0, 4, 2, 3, 6, 1, 5, 7,” then the corresponding treatment on each lettuce square is described in the table below:

Figure 2

Lettuce Label	***Brief***	***Treatment Details***
0	*No Treatment*	N/A
4	*Block Air*	The lettuce square will be wrapped in plastic wrap and placed in the room.
2	*Water*	Every day at 9 pm, the lettuce is soaked in 100ml water for 3 seconds and placed in a room.
3	*Water & Block Air*	Every day at 9 pm, the lettuce is unwrapped and soaked in 100ml water for 3 seconds. Then, the lettuce will be wrapped around the plastic wrap and placed in a room.
6	*Fridge*	The lettuce square is placed in the fridge without other treatment.
1	*Fridge & Block Air*	The lettuce square is placed in the fridge and wrapped around by plastic wrap.
5	*Fridge & Water*	The lettuce square is placed in Fridge. Every day at 9 am, the lettuce is soaked in 100ml water for 3 seconds, then placed back in the fridge for the rest of day.
7	*Fridge & Water & Block Air*	The lettuce square is placed in Fridge and wrapped in plastic wrap. Every day at 9 pm, the lettuce is unwrapped and soaked in 100ml water for 3 seconds. Then, the lettuce will be wrapped around by plastic wrap and placed in the fridge.

**Each lettuce square is placed under environment <u>without light</u>.*

**The experiment starts at 9:30 am on March 3 and the data is collected at 5:00 pm on March 10 (6 days).*

<u>*Data Collection*</u>

First, we obtained a grid measurement sheet (a letter size paper divided into multiple squares, and the area of each square = $1mm^2$) from the internet. Then, we cut off the spoiled area (*Definition*: the *color* or *texture* is different from fresh lettuce) of each lettuce square and put remain lettuce square on the grid measurement sheet.

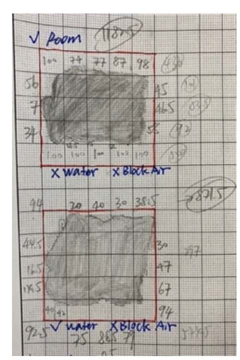

Figure 3

Press the lettuce leaf on the grid sheet carefully to make sure the lettuce leaf is not curled and outline the shape of lettuce leaves on the grid sheet. Finally, we draw a border square with a side length of 5cm and get the size of the spoiled area by counting the ***number of 1 millimeter square*** outside the outline and inside the border square (if the outline passes through the 1mm-square, that square is counted as 0.5 instead of 1).

Model Fit, Analysis, and Diagnostic

In general, the model fit and analysis are performed in *R Studio* (version 1.4.1103). In specific, for data cleaning and organizing, packages *tidyr* and *dplyr* are used; for data visualization, packages *ggplot2* (dot plot) and *stats* (interaction plot) are used; for model fitting and analysis, packages *stats* (ANOVA) and *agricolae* (Fisher's test) are used.

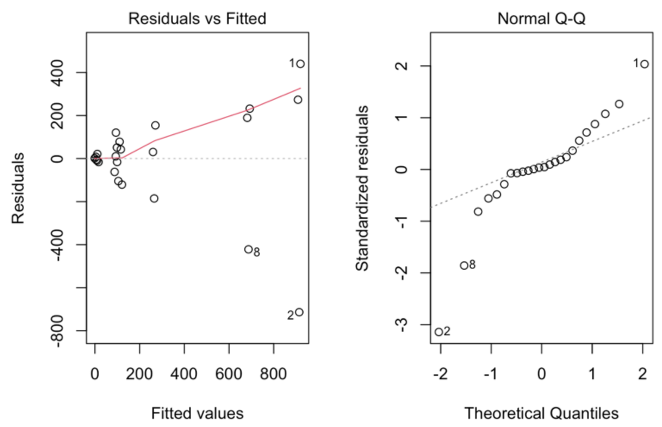

Figure 4

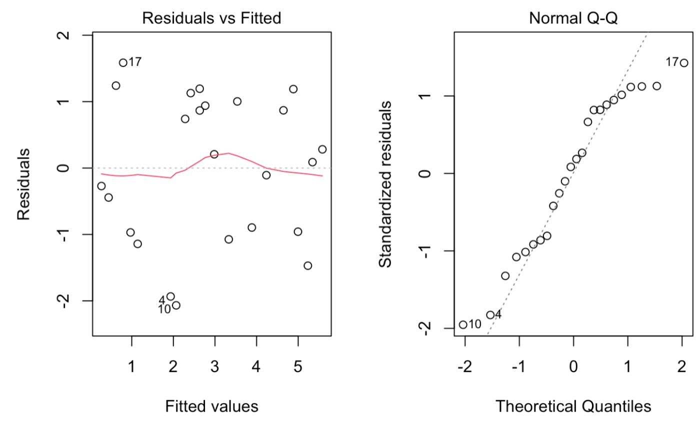

Figure 5

Since there is an obvious fanning pattern in the "Residual vs. Fitted" Plot in *Figure 4*, the assumption of equal variance is violated. So, a transformation of data is required. Besides, from the "Normal Q-Q" Plot in *Figure 4*, since the points lie mostly along the straight diagonal line with some minor deviations along each of the tails, the normality assumption is met.

To stabilize the variance, we first perform the square-root transformation, but the "residual vs. fitted" plot still indicates unequal variance. So, we perform the $\boldsymbol{y^* = y^{1/4}}$ ***Transformation***.

For the data after transformation, since the fanning pattern is no longer obvious in the "Residual vs. Fitted" Plot in *Figure 5*, the assumption of equal variance is met. Also, from the "Normal Q-Q" Plot in *Figure 5*, since the points lie mostly along the straight diagonal line with some minor deviations along each of the tails, the normality assumption is met.

In conclusion, for data after transformation, we may conclude no assumptions are violated and assume that the model is adequate.

Results & Discussion

Data Visualization

We plot two ***Scatter Plots*** (*Figure 6* & *Figure 7*) on the original data. *Figure 7* suggests: storing lettuce leaves at low temperature may decrease the spoiled area of lettuce leaf in 6 days (Effect of *Temperature*); storing lettuce at low air contact level may decrease the spoiled area of lettuce leaf in 6 days (Effect of *Air Contact Level*); storing lettuce in high humidity may decrease the spoiled area of lettuce leaf in 6 days (Effect of *Humidity*). Moreover, *Figure 7* also suggests there may be possible interaction between *Air Contact Level* and *Temperature*.

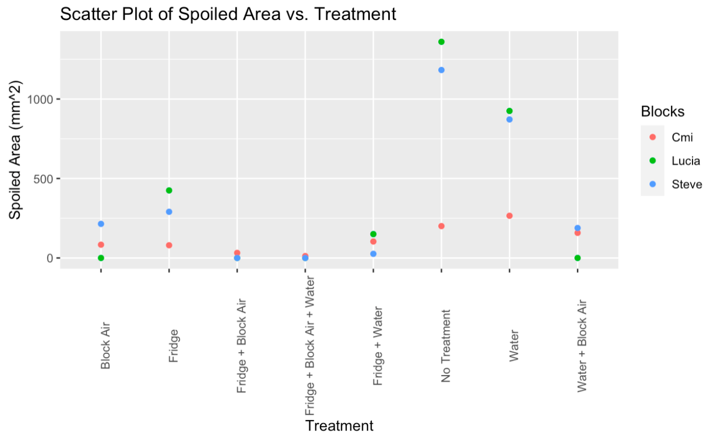

Figure 6

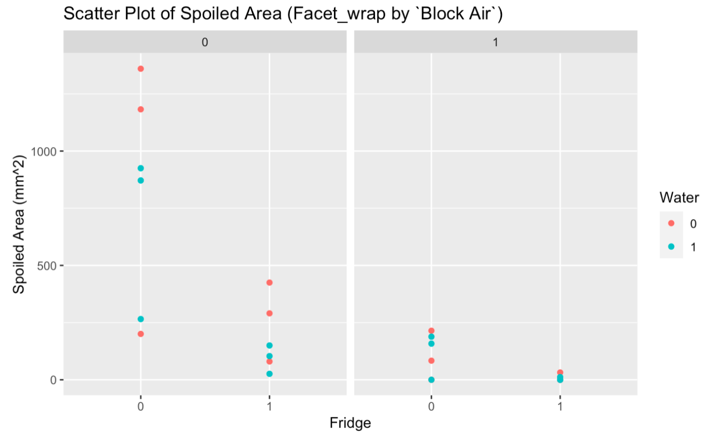

Figure 7

We further verify whether there is an interaction between variables through an ***Interaction Plot*** (*Figure 8*). In *Figure 8*, the parallel lines suggest there may not be interaction between *Temperature* and *Humidity*. However, the non-parallel lines in Figure 8 suggest there may be interaction between Temperature and Air Contact Level and interaction between *Humidity* and *Air Contact Level*.

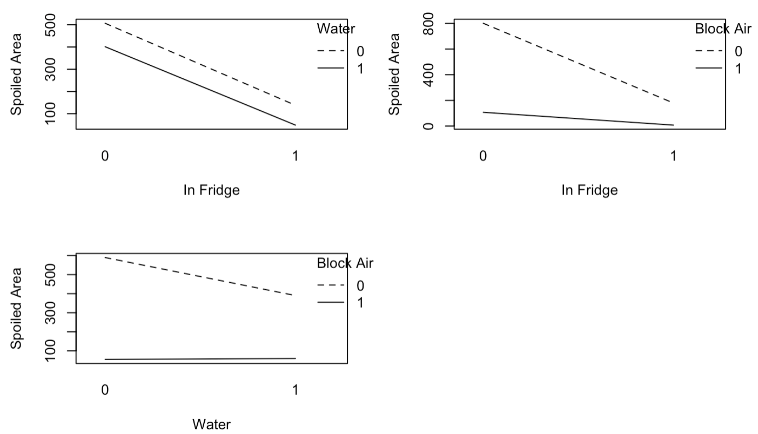

Figure 8

```
Analysis of Variance Table

Response: Spoiled.Area
                   Df Sum Sq Mean Sq F value    Pr(>F)
Fridge              1 16.565  16.565  8.9326 0.0091808 **
Water               1  0.523   0.523  0.2821 0.6030802
Block.Air           1 45.241  45.241 24.3965 0.0001782 ***
block               1  1.952   1.952  1.0526 0.3211598
Fridge:Water        1  0.354   0.354  0.1910 0.6683361
Fridge:Block.Air    1  0.002   0.002  0.0011 0.9745140
Water:Block.Air     1  0.457   0.457  0.2463 0.6268849
Fridge:Water:Block.Air  1  0.049   0.049  0.0262 0.8736446
Residuals          15 27.816   1.854
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 9

Inferences

Finally, even though the plots indicate the main effect of certain variables and interactions between variables, we would like to perform an Analysis of Variance to verify our plot interpretation and test the ***hypothesis***. (Since the original data have unequal variance, which does not satisfy the model's assumption, the *ANOVA* and the *Fisher LSD* are performed on transformed data)

With a P-value of 0.0091808, there is strong evidence against the null hypothesis. At an α = 0.05 level, we reject the null hypothesis. So, we can conclude that $\tau_i \neq 0$ for at least one i, which suggests the factor *Temperature* affect the spoiled area of lettuce leaf in 6 days.

Also, with a P-value of 0.0001782, there is strong evidence against the null hypothesis. At an α = 0.05 level, we reject the null hypothesis. So, we can conclude that $\gamma_k \neq 0$ for at least one k, which suggests the factor *Air Contact Level* affect the spoiled area of lettuce leaf in 6 days.

However, with a P-value of 0.6030802, there is NOT enough evidence to reject the null hypothesis $H_0: \beta_1 = \beta_2 = \cdots = \beta_b = 0$ at an α = 0.05 level. We can conclude that there is not statistically significant evidence to support the claim that *Humidity* affects the spoiled area of lettuce leaf in 6 days.

With a P-value of 0.6683361, there is NOT enough evidence to reject the null hypothesis $H_0: (\tau\beta)_{ij} = 0$ for all i, j at an α = 0.05 level. We can conclude that there is not statistically significant evidence to support the claim that there is an interaction between *Temperature* and *Humidity*.

With a P-value of 0.9745140, there is NOT enough evidence to reject the null hypothesis $H_0: (\tau\gamma)_{ik} = 0$ for all i, k at an α = 0.05 level. We can conclude that there is not statistically significant evidence to support the claim that there is an interaction between *Temperature* and *Air Contact Level*.

With a P-value of 0.6268849, there is NOT enough evidence to reject the null hypothesis $H_0: (\beta\gamma)_{jk} = 0$ for all j, k at an α = 0.05 level. We can conclude that there is not statistically significant evidence to support the claim that there is an interaction between *Humidity* and *Air Contact Level*.

With a P-value of 0.8736446, there is NOT enough evidence to reject the null hypothesis $H_0: (\tau\beta\gamma)_{ijk} = 0$ for all i, j, k at an α = 0.05 level. We can conclude that there is not statistically significant evidence to support the claim that there is an interaction between *Temperature, Humidity,* and *Air Contact Level*.

Finally, with a P-value of 0.3211598 > 0.05, we may conclude that *blocking* is NOT effective.

In addition to the hypothesis testing, we would like to further investigate how factors that have statistically significant (*Temperature*, *Air Contact Level*) influence spoiled area of lettuce leaf in 6 days. In specific, we choose the ***Fisher LSD*** to solve our questions.

Fridge:Block Air	Spoiled.Area ^(1/4) <dbl>	groups <chr>
0:0	5.114099	a
1:0	3.434484	b
0:1	2.350100	b
1:1	0.706604	c

Figure 10

Figure 10 suggests that storing lettuce leaves at low *temperature* will have **a smaller** spoiled area of lettuce leaf in 6 days than storing lettuce leaves without any treatments; storing lettuce leaves in low *air contact level* will have **a smaller** spoiled area of lettuce leaf in 6 days than storing lettuce leaves without any treatments.

Also, *Figure 10* suggests there is **no significant difference** in the spoiled area of lettuce leaves between storing lettuce leaves at low *temperature* and storing lettuce leaves at low *air contact level* for 6 days.

Finally, Figure 10 suggests that storing lettuce leaves at low *temperature* & low *air contact level* will have the **smallest** spoiled area of lettuce leaf in 6 days when being compared with other treatments.

Discussion

The statistical inferences suggest that based on our experiment results, both Temperature and Air Contact Level have statistically significant effects on the spoiled area of lettuce leaves stored for 6 days. In specific, both lower temperature and air contact level can decrease the spoiled area of lettuce leaves stored for 6 days. In addition, preserving lettuce leaves at low temperature & air contact level will have the greatest decrease in the spoiled area when compared with other preserving methods. This suggests if people want to preserve vegetables like lettuce for a long time, they should store vegetables in an environment that has low temperature and less air contact: storing vegetables in the fridge and wrapping the vegetable with plastic food wrappers or in sandwich bags.

There are several limitations in the experiment's design and analysis. First and foremost, when doing the randomization, we should randomly assign the lettuce squares should be randomized instead of the leaves of lettuce (this may make our model inadequate). Also, the sample size is not big enough, and more replicates for each block can lead to more accurate conclusions.

In future research, there are two things worth further investigation. 1. The scatter plot (*Figure 7*) indicates an effect of *Humidity*, but based on the hypothesis tests' results, there is no statistically significant evidence to claim *Humidity* can affect the spoiled area of lettuce leaves. So, in future research, we could perform a Completely Randomized Design (CRD) to further explore whether *Humidity* will influence the spoiled area of lettuce leaves. 2. Also, even though the 2-way interaction plot suggests there can be an interaction between certain variables, the hypothesis tests' results suggest there is no statistically significant evidence to claim there are many kinds of interaction between variables. So, in the future experiment, we can perform Completely Randomized Design (CRD) in the following way:

Interaction	**Sample 1 Treatment**	**Sample 2 Treatment**
Air Contact & Temperature	Block Air	Block Air + Fridge
Air Contact & Humidity	Block Air	Block Air + Water

Conclusion

In conclusion, according to the experimental result, we can conclude that the factor ***temperature*** and ***air contact level*** have statistically significant effects on the spoiled area of lettuce stored for 6 days. In specific, with lower temperature or less air contact, the lettuce leaves tend to have a smaller area of decay.

In the future, we will perform CRDs to further explore 2 questions informed by the results of the exp eriment: whether *Humidity* will influence the spoiled area of lettuce leaves; whether an interaction between *Air Contact Level* and *Temperature* or interaction between *Air Contact Level* and *Humidity* exists.