

# Berry Project – Strawberry Data Report

Ruxin Liu

10/13/2020

## Introduction

As nowadays, more and more people are concerned about health, organic food becomes a more popular choice. In general, there is no universal way to define organic food, but since the main goal of organic agriculture is to optimize the biodiversity as well as the ecological balance, the usage of synthetic chemicals are restricted (food unfolded 2019). In this study, the berry data collected by the United States Department of Agriculture (USDA) database selector and stored online was explored and analyzed to compare the types and quantities of chemicals applied for planting strawberries during different years and in different states.

## Methodology

### Data Cleaning & Variable Selection

The berry dataset contains information about three kinds of berries, which are blueberries, raspberries and strawberries. In this study, only the data points related to strawberries are considered. Firstly, some of the variables only have one unique value for all the observations, which cannot provide any useful information for the purpose of comparison and can be removed. It turns out that 12 out of the 21 variables only have one unique value, such as “Program”, “Week Ending”, “Geo Level” and etc. After the first step, the dataset now had 8 variables and 3220 observations in total. Next, since the variable “Data Item” contains lots of different information in one single cell, it needs to be reorganized and renamed. After this step, the variable “Data Item” were separated into 4 new variables, which are “berry” (types of berry), “market” (marketing condition), “harvest” (harvest condition) and “units” (units of values collected). Similarly, the variable “Domain” was separated into “Domain\_1” and “Domain\_2”, and the variable “Domain Category” was separated into “DC\_1\_left”, “DC\_1\_right”, “DC\_2\_left” and “DC\_2\_right” first. After carefully checking the unique combinations, it was found that variables “Domain\_1” and “DC\_1\_left” produced exactly the same values. Similarly, it was found that variables “Domain\_2” and “DC\_2\_left” produced exactly the same values. Therefore, only “Domain\_1” and “Domain\_2” were kept, combined and renamed to be “chemical”, which states the type of chemicals being used.

After the above steps, the data still contained much information than needed and required deeper cleaning. The final strawberry data set was selected to include 8 variables, which were Year, State, market, harvest, materials, chemical, Values and units, that all could provide useful information for analysis. All the NA values and D values that are used to withhold to avoid data disclosure were removed from the variable “Value”. Then, this tidy data was used to subset a smaller data set of interest, which only looked at observations where chemicals were applied. This was done by only selecting rows with harvest values equal to “APPLICATIONS” and unit values equal to “LB / ACRE / APPLICATION”. Now, the smaller data set “unfood” only contained information about the pounds of chemicals used per acre in some states and years, which were all prepared to be analyzed and compared.

All the R code and data used were attached in the Appendix and also in the GitHub repository (<https://github.com/Ruxinliu97/Berry>).

## Exploratory Data Analysis (EDA)

Based on the dataset, it is found that only California, Florida and Washington used chemicals when planting strawberries during the year 2016, 2018 and 2019. In order to see any differences, histograms are plotted by using the ggplot2 package. From Fig.1, it is shown that the number of strawberry farms which applied chemicals varied a lot from states to states. For Washington, there were just 3 cases of chemical usage detected in the year 2016 only. The figure also conveyed that although none of the farms used chemicals in California and Florida for the year 2017, there was continuous and relatively large use of chemicals in other years. Among the 3 states, the farms that used chemicals when planting strawberries were mostly from California.

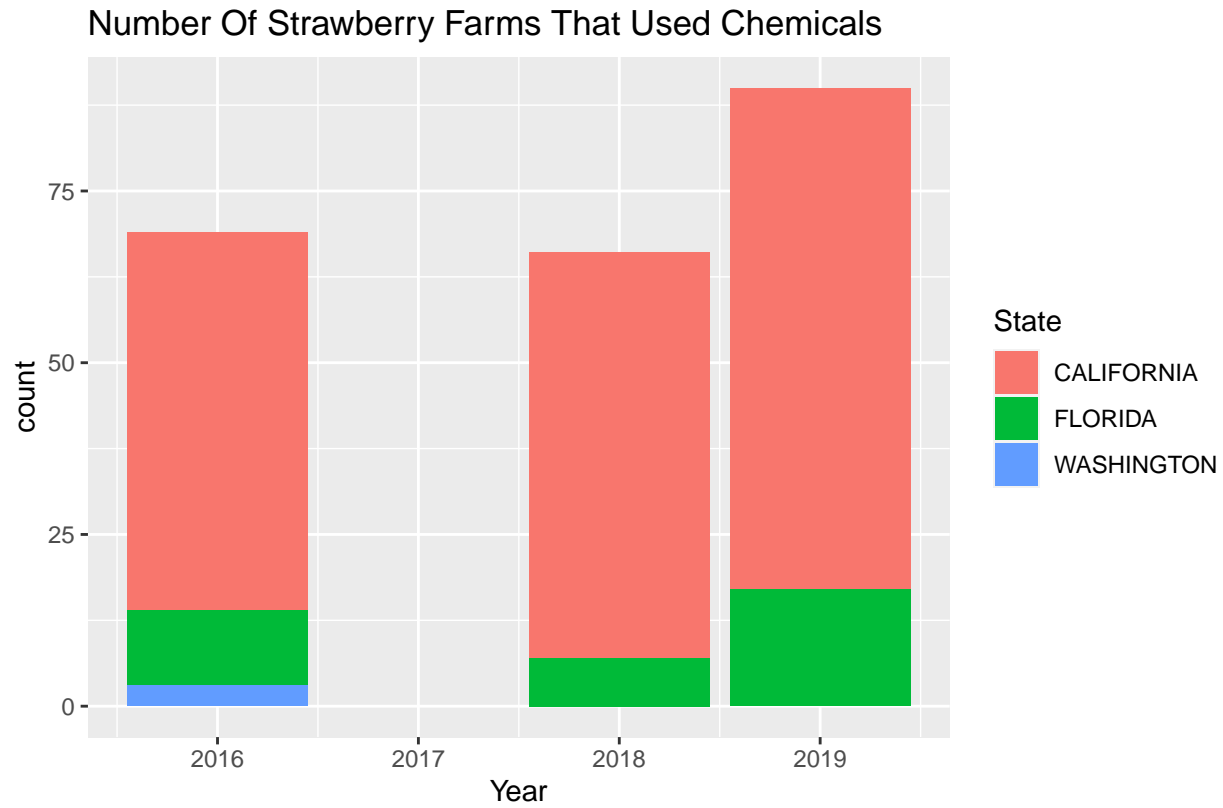


Fig.1

From Fig.2, it is shown that among the 225 observations, the kinds of chemicals being applied were also different. It is clear that for all 3 years, herbicides and insecticides were used by most farms.

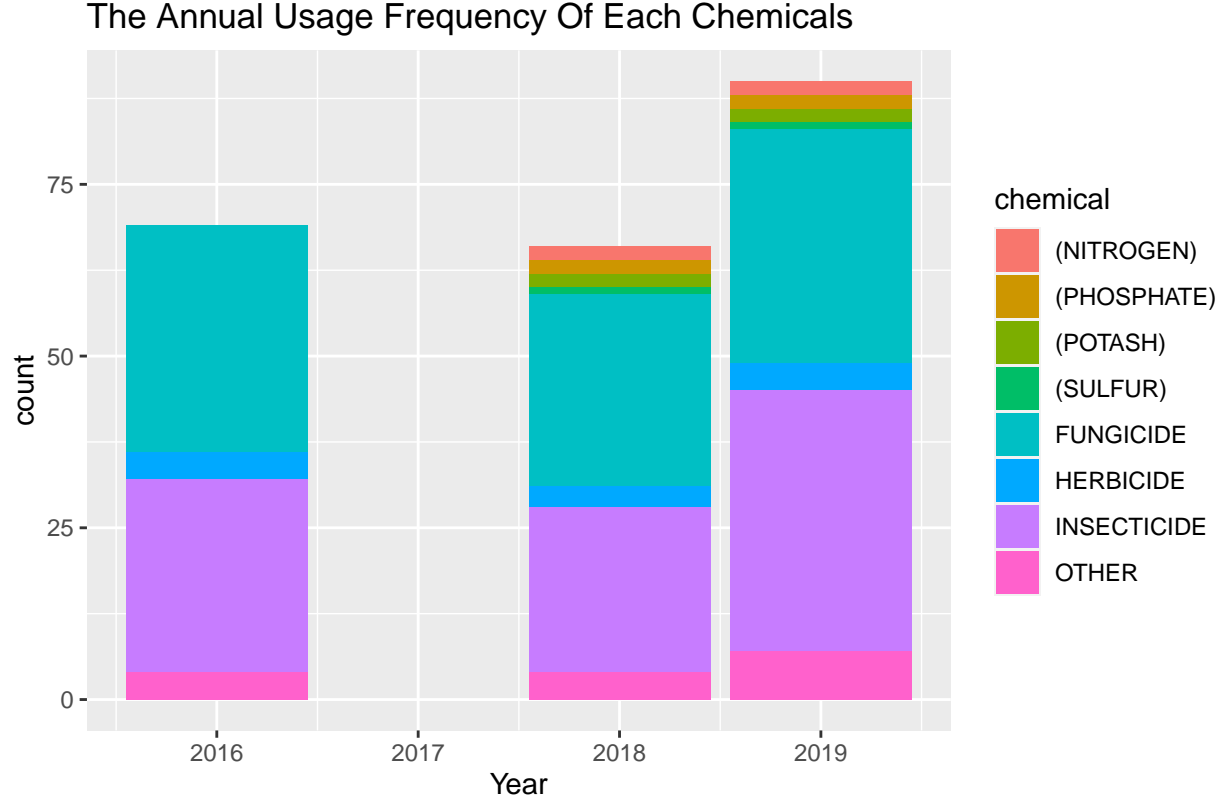


Fig.2

In order to make a better comparison, the 4 kinds of synthetic chemicals, which are fungicide, herbicide, insecticide and other, were also explored and summarized separately. Based on the results of Table.1, in 2016 one farm in Florida had the lowest pounds of fungicide usage per acre per application, which is 0.01, while one farm in California had the highest pounds of fungicide usage per acre per application, which is 3.712. In 2018, one farm in California had the lowest pounds of fungicide usage per acre per application, which is 0.022, while another farm in California had the highest pounds of fungicide usage per acre per application, which is 4.864. In 2019, one farm in California had the lowest pounds of fungicide usage per acre per application, which is 0.023, while one farm in Florida had the highest pounds of fungicide usage per acre per application, which is 3.997. The sum of the fungicide pounds may not be suitable for comparisons since each State has different number of strawberry farms. However, from the mean value, it is clear that in recent years, Florida on average had a noticeably higher average pounds of fungicide usage per acre per application.

Table 1: Fungicide Application

Year	State	min	max	mean	sum
2016	CALIFORNIA	0.022	3.712	0.568	13.634
2016	FLORIDA	0.010	1.909	0.514	3.600
2016	WASHINGTON	0.155	0.276	0.216	0.431
2018	CALIFORNIA	0.022	4.865	0.676	16.225
2018	FLORIDA	0.221	1.858	1.010	4.041
2019	CALIFORNIA	0.023	3.997	0.697	19.527
2019	FLORIDA	0.206	2.098	1.142	6.851

Table 2: Herbicide Application

Year	State	min	max	mean	sum
2016	CALIFORNIA	0.091	1.382	0.647	1.940
2016	WASHINGTON	0.228	0.228	0.228	0.228
2018	CALIFORNIA	0.096	1.364	0.604	1.813
2019	CALIFORNIA	0.024	1.202	0.407	1.628

Table 3: Insecticide Application

Year	State	min	max	mean	sum
2016	CALIFORNIA	0.019	5.860	0.542	13.017
2016	FLORIDA	0.039	0.460	0.168	0.670
2018	CALIFORNIA	0.019	6.340	0.518	12.433
2019	CALIFORNIA	0.019	5.151	0.413	12.394
2019	FLORIDA	0.045	0.141	0.072	0.575

Table 4: Other Application

Year	State	min	max	mean	sum
2016	CALIFORNIA	0.114	251.559	89.052	356.209
2018	CALIFORNIA	0.113	236.901	80.109	320.435
2019	CALIFORNIA	0.112	244.066	110.023	770.161

## Conclusion

## Appendix

### Code

#### Importing the data

```
# Load the data
library(tidyverse)
library(magrittr)
berry <- read_csv("berries.csv", col_names = TRUE)

# Look at number of unique values in each column
berry %>% summarize_all(n_distinct) -> unique
# Make a list of the columns with only one unique value
one <- which(unique[1,] == 1)
# Remove the 1-unique columns from the dataset
berry %>% select(-all_of(one))
# State name and the State ANSI code are redundant -- keep the name only
berry %>% select(-4)

# Select data on Strawberries only from the berry data
strawberry <- berry %>% filter((Commodity == "STRAWBERRIES") & (Period == "YEAR"))
strawberry %>% select(-c(Period, Commodity))

# The original format of one specific domain type
# (CHEMICAL, INSECTICIDE: (CYFLUMETOFEN= 138831)) is not processed successfully
```

```

# with the separate function, therefore they need to be changed manually.
manual <- c(1036, 1148, 1258, 1368, 1480, 1977, 2075, 2171, 2267, 2365,
           2469, 2542, 2613, 2684, 2757, 2985, 3019, 3052, 3085, 3120)
strawberry$`Domain Category`[manual] <- "CHEMICAL, INSECTICIDE: (CYFLUMETOFEN = 138831)"

# Organize the column data item which contains lots of information in a cell
# unique(strawberry$`Data Item`)
strawberry %<>% separate(`Data Item`, c("s1", "s2"), sep = "-")
strawberry %<>% separate(s1, c("berry", "type"), sep = ",")
strawberry %<>% separate(s2, c("data_item", "unit"), sep = ",")

unique(strawberry$berry)
# Since all the data selected are about strawberries, there is no need to
# contain name of berries anymore
strawberry %<>% select(-berry)

# Organize the column Domain
strawberry %<>% separate(Domain, c("Domain_1", "Domain_2"), sep = ", ")

# Organize the column Domain Category
strawberry %<>% separate(`Domain Category`, c("DC_1", "DC_2"), sep = ", ")
strawberry %<>% separate(DC_1, c("DC_1_left", "DC_1_right"), sep = ": ")
strawberry %<>% separate(DC_2, c("DC_2_left", "DC_2_right"), sep = ": ")

# Check and delete the redundant variables
paste(strawberry$Domain_1, strawberry$DC_1_left) %>% unique
strawberry %<>% select(-DC_1_left)

# Check and delete the redundant variables
paste(strawberry$Domain_2, strawberry$DC_2_left) %>% unique
strawberry %<>% select(-DC_2_left)

# Change all NA values into a space
strawberry[is.na(strawberry)] <- " "
strawberry %<>% mutate(Domain_1 = "CHEMICAL", Domain_1 = "")
strawberry %<>% mutate(Chemical = str_trim(paste(Domain_1, Domain_2)))
strawberry %<>% mutate(market = str_trim(type))
strawberry %<>% mutate(harvest = str_trim(data_item))

strawberry %<>% rename(chem_family = DC_1_right, materials = DC_2_right)
# All the rows with information on chem_family have no information on Chemical,
# so it is reasonable to combine these columns.
strawberry %<>% mutate(chemical = str_trim(paste(chem_family, Chemical)))

strawberry %<>% separate(unit, c("u1", "u2", "u3", "u4", "u5", "u6", "u7",
                                "u8"), sep = " ")

# Only leave the useful information after "measure in"
strawberry[is.na(strawberry)] <- " "
strawberry %<>% mutate(units = str_trim(paste(u4, u5, u6, u7, u8)))
# Select the variables that are kept in the final data
strawberry %<>% select(Year, State, market, harvest, materials, chemical,
                      Value, units)

# Check that all the variables are now tidy
# unique(strawberry$Year)
# unique(strawberry$State)

```

```

# unique(strawberry$market)
# unique(strawberry$harvest)
# unique(strawberry$chemical)
# unique(strawberry$units)

# Look at chemicals being applied to strawberry
unfood <- strawberry
unfood <- unfood %<>% filter(harvest == "APPLICATIONS")
# Remove NAs
unfood %<>% filter(Value != "(D)")
unfood %<>% filter(Value != "(NA)")
unfood %<>% filter(units == "LB / ACRE / APPLICATION")
unfood_1 <- unfood %>% select(Year, State, chemical, Value)
unfood_1 %<>% pivot_wider(names_from = chemical, values_from = Value)

# Convert the characters into numeric values for further calculation and analysis
for(i in 1 : nrow(unfood_1)) {
  unfood_1$FUNGICIDE[[i]] <- as.numeric(unfood_1$FUNGICIDE[[i]])
}
for(i in 1 : nrow(unfood_1)) {
  unfood_1$HERBICIDE[[i]] <- as.numeric(unfood_1$HERBICIDE[[i]])
  unfood_1$INSECTICIDE[[i]] <- as.numeric(unfood_1$INSECTICIDE[[i]])
  unfood_1$OTHER[[i]] <- as.numeric(unfood_1$OTHER[[i]])
}

# Calculate the total count of each type of chemical usage
unfood_1 %<>% mutate(total_fungi = "NA")
unfood_1 %<>% mutate(total_herb = "NA")
unfood_1 %<>% mutate(total_insect = "NA")
unfood_1 %<>% mutate(total_other = "NA")
for(i in 1 : nrow(unfood_1)) {
  unfood_1$total_fungi[i] <- length(unfood_1$FUNGICIDE[[i]])
  unfood_1$total_herb[i] <- length(unfood_1$HERBICIDE[[i]])
  unfood_1$total_insect[i] <- length(unfood_1$INSECTICIDE[[i]])
  unfood_1$total_other[i] <- length(unfood_1$OTHER[[i]])
}

# Fig.1
library(ggplot2)
unfood %>%
  ggplot(aes(x = Year)) +
  geom_bar(aes(fill = State))

# Fig.2
unfood %>%
  ggplot(aes(x = Year)) +
  geom_bar(aes(fill = chemical))

# Table 1: Summarize the fungicide application
library(kableExtra)
fungicide <- unfood %>%
  filter(chemical == "FUNGICIDE") %>%
  group_by(Year, State) %>%
  summarise(min = min(as.numeric(Value)),
            max = max(as.numeric(Value)),

```

```

        mean = mean(as.numeric(Value)),
        sum = sum(as.numeric(Value))
knitr::kable(fungicide, caption = "Table.1: Fungicide Application ", digits = 3)

# Table 2: Summarize the herbicide application
herbicide <- unfood %>%
  filter(chemical == "HERBICIDE") %>%
  group_by(Year, State) %>%
  summarise(min = min(as.numeric(Value)),
            max = max(as.numeric(Value)),
            mean = mean(as.numeric(Value)),
            sum = sum(as.numeric(Value)))
knitr::kable(herbicide, caption = "Table.2: Herbicide Application ", digits = 3)

# Table 3: Summarize the insecticide application
insecticide <- unfood %>%
  filter(chemical == "INSECTICIDE") %>%
  group_by(Year, State) %>%
  summarise(min = min(as.numeric(Value)),
            max = max(as.numeric(Value)),
            mean = mean(as.numeric(Value)),
            sum = sum(as.numeric(Value)))
knitr::kable(insecticide, caption = "Table.3: Insecticide Application ", digits = 3)

# Table 4: Summarize the other applications
other <- unfood %>%
  filter(chemical == "OTHER") %>%
  group_by(Year, State) %>%
  summarise(min = min(as.numeric(Value)),
            max = max(as.numeric(Value)),
            mean = mean(as.numeric(Value)),
            sum = sum(as.numeric(Value)))
knitr::kable(other, caption = "Table.4: Other Application ", digits = 3)

```

## Reference of R Packages

1. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43),1686, <https://doi.org/10.21105/joss.01686>
2. Stefan Milton Bache and Hadley Wickham (2014). magrittr: A Forward-Pipe Operator for R. R package version 1.5. <https://CRAN.R-project.org/package=magrittr>
3. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
4. Hao Zhu (2019). kableExtra: Construct Complex Table with ‘kable’ and Pipe Syntax. R package version 1.1.0. <https://CRAN.R-project.org/package=kableExtra>

## Bibliography

1. MA615 notes (class 14 - class 17)
2. food unfolded (2019). *What is organic food / Is it really chemical free?* [online]. Available from: <https://www.foodunfolded.com/how-it-works/what-is-organic-food-is-it-really-chemical-free> [accessed 18 October 2020].
3. USDA (2019). *Quick Stats* [online]. Available from: <https://quickstats.nass.usda.gov/> [accessed 16 October 2020].

4. USDA (2019). *Quick Stats* [online]. Available from: <https://quickstats.nass.usda.gov/results/D416E96E-3D5C-324C-9334-1D38DF88FFF1> [accessed 16 October 2020].