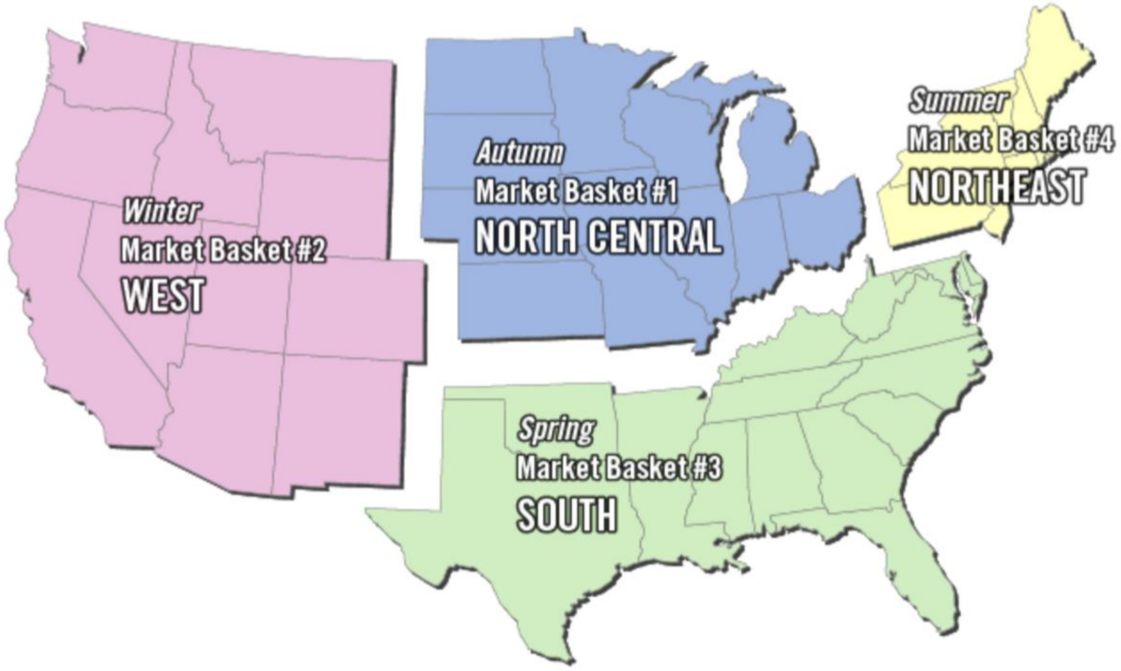
Analyzing the FDA’s Total Diet Study (TDS)

Presented at the 2016 Joint Statistical Meetings

Arjun Panda, Jacob Holman, Ayona Chatterjee

09/14/16

Market Baskets

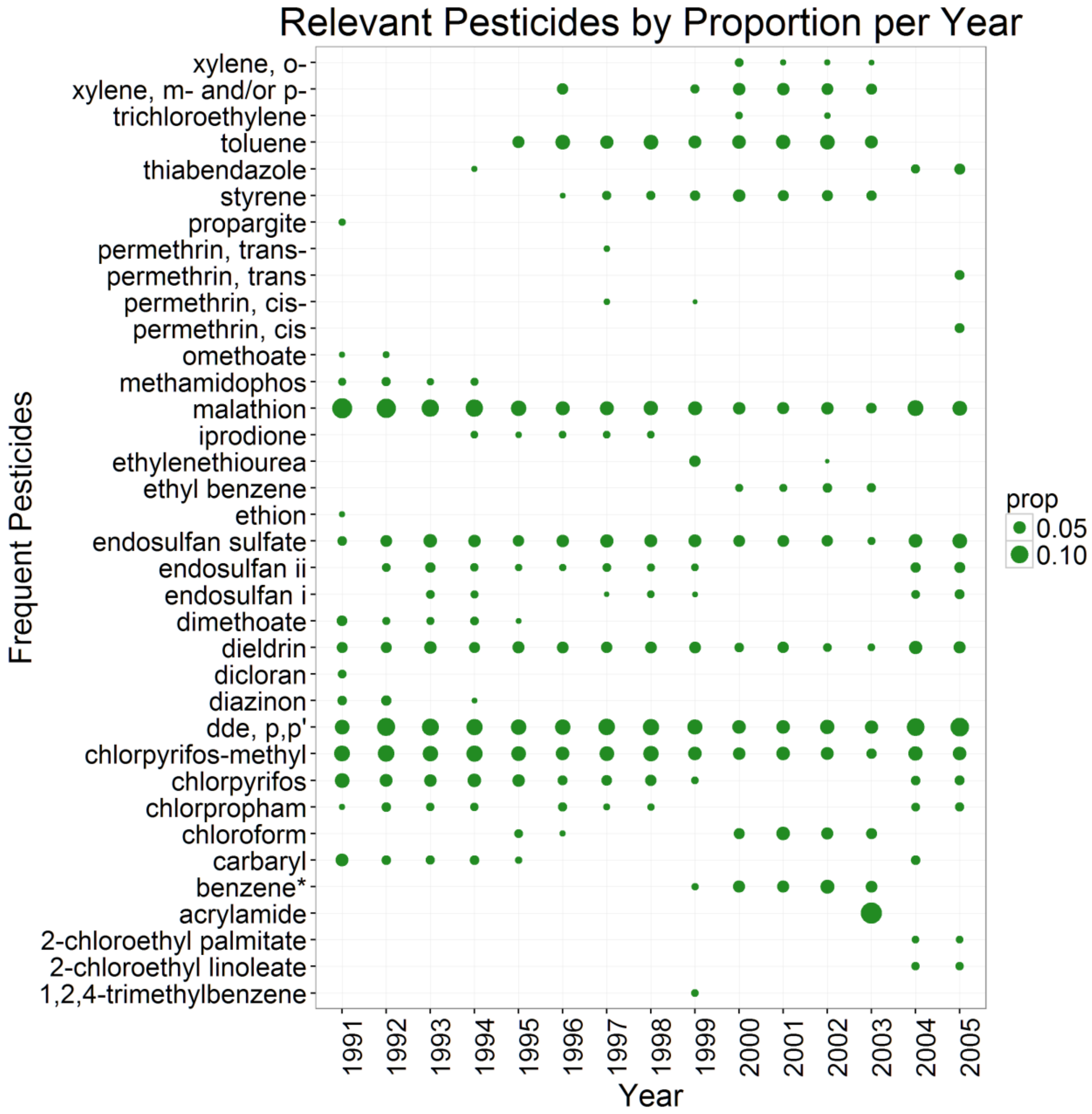


Total Diet Study collection ("market basket") regions

Research Goals:

1. Exploratory data analysis
2. The FDA grouped the foods into 25 intuitive categories (grains, fruits, processed food, etc). In contrast, we created new groupings based on pesticide profiles using machine learning techniques.

First: Exploratory Analysis

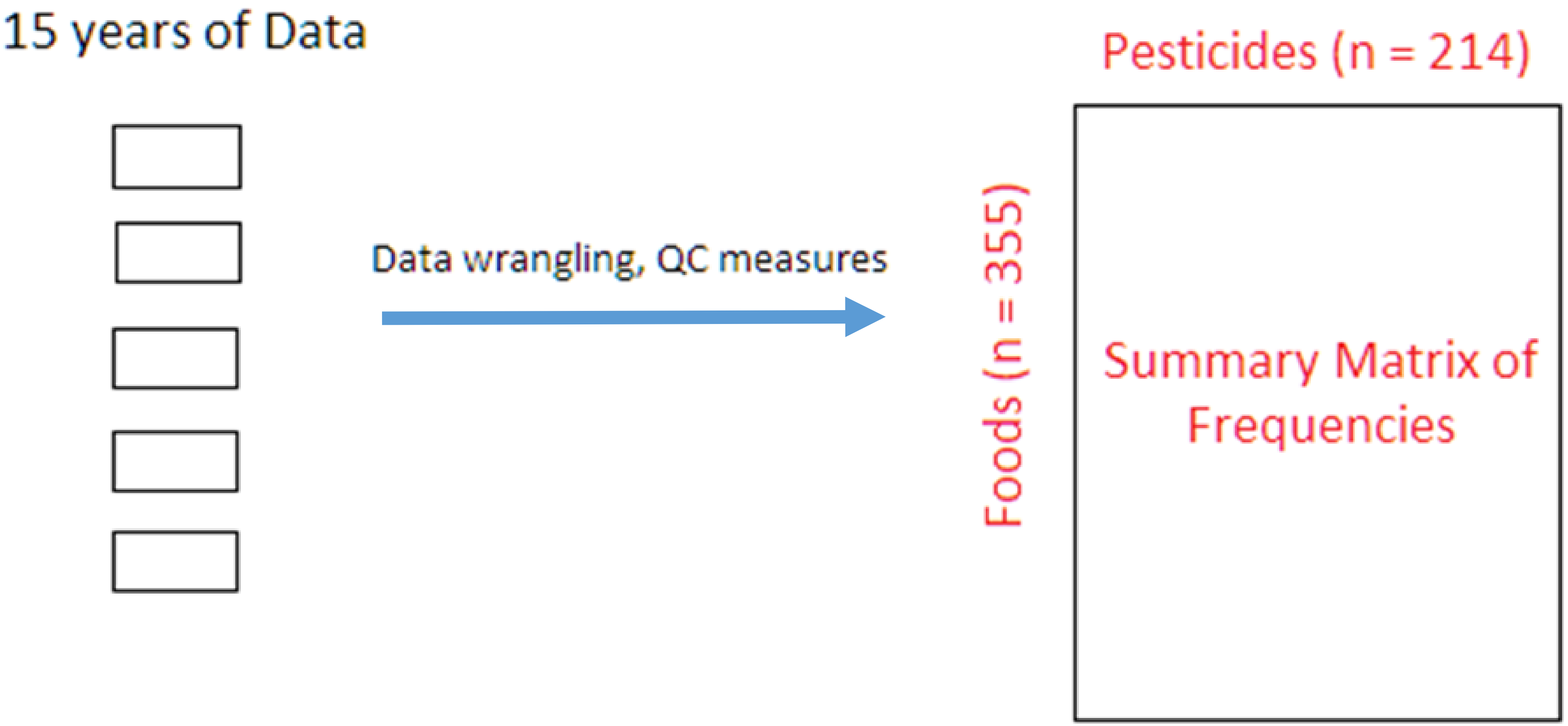


Shiny

by RStudio

Made a little app to visualize pesticides in food over time

Preparing raw data for clustering



15 years of Data

Data wrangling, QC measures

Pesticides (n = 214)

Foods (n = 355)

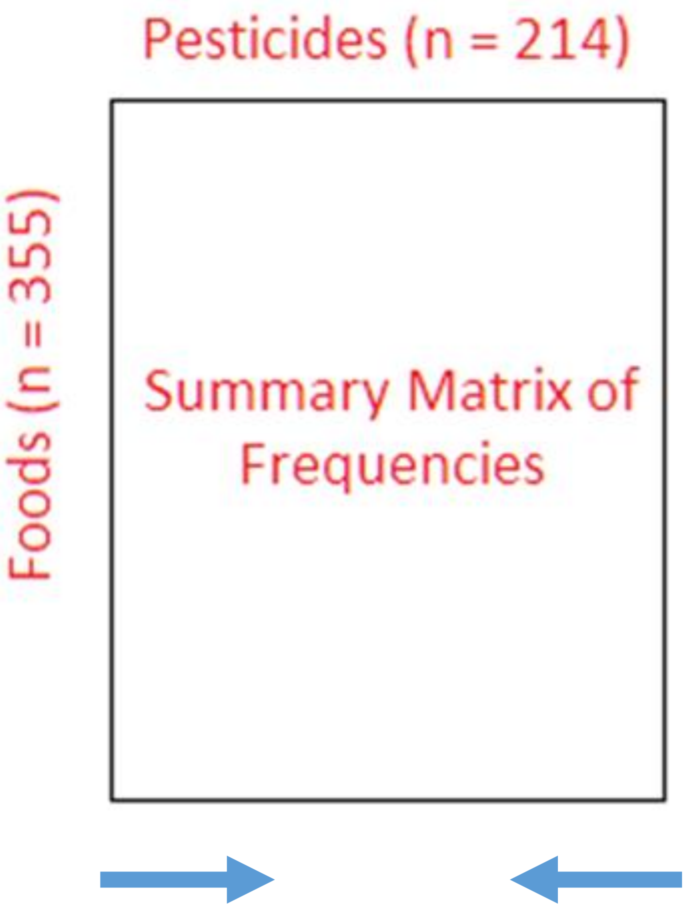
Summary Matrix of Frequencies

Initially we wanted to use concentrations as the response variable.

Which unsupervised clustering algorithm?

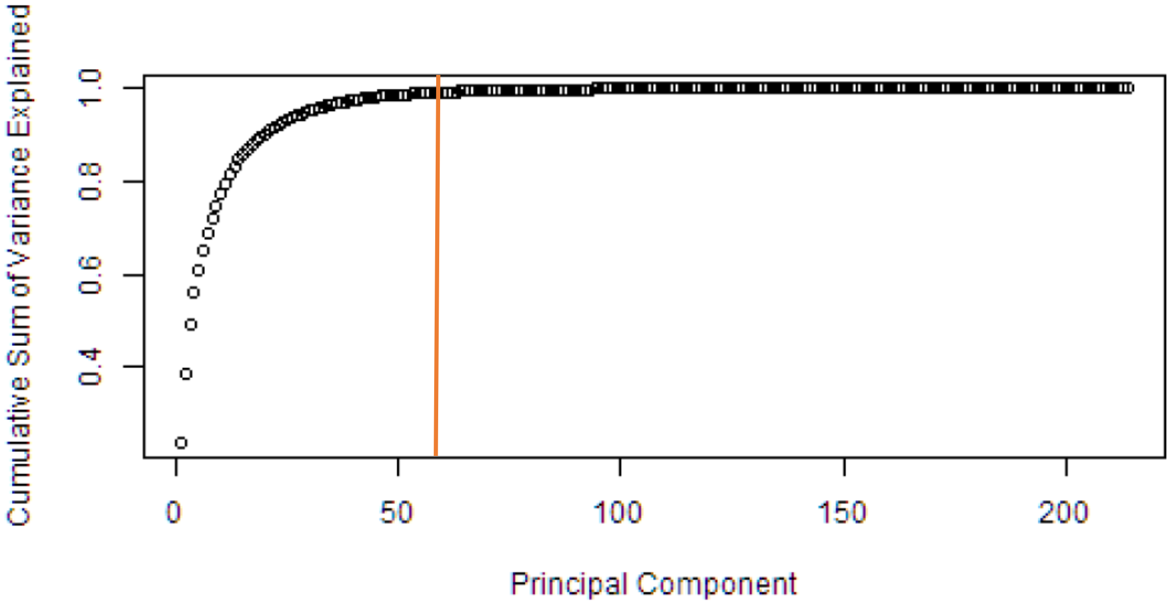
- K-means vs hierarchical clustering
- Went with **hierarchical clustering**:
  1. K-means assigns initial nodes randomly, so in a massive, sparse matrix the algorithm can easily get caught in a local optimum rather than a global one.
  2. HC is deterministic.
  3. O(n^2) complexity algorithm, so compute time increases exponentially with dataset.

PCA

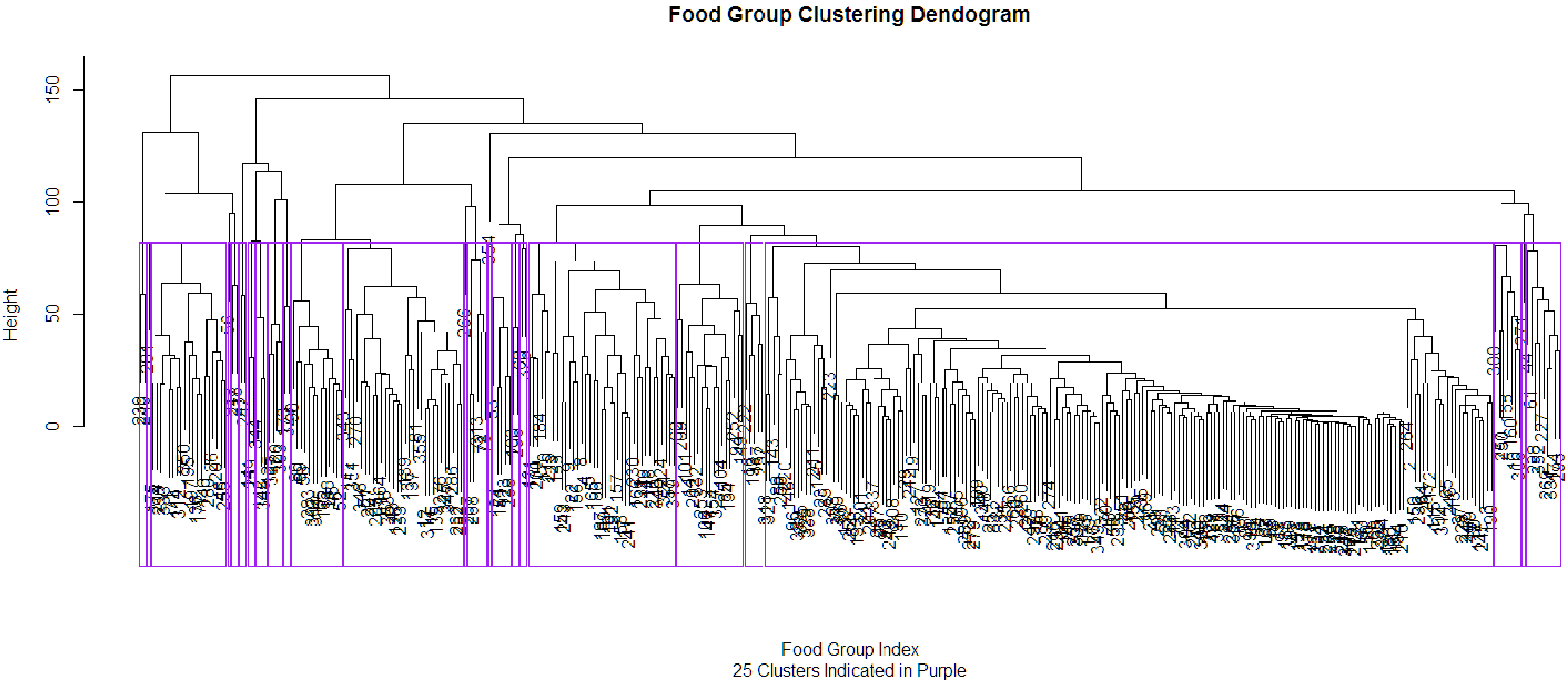


Reducing # of dimensions (pesticides) while preserving the information contained

Keep principle components with variances (eigenvalues) ≥ 1. Leaves us with 58 dimensions.



(>99% of variance explained in 58 dims)



*Purple boxes around 25 cuts

Sample of Results

```
 [1] "American, processed cheese"  "beef chuck roast, baked"     "beef, strained/junior"
 [4] "bologna, sliced"            "cheddar cheese"             "cream cheese"
 [7] "evaporated milk, canned"     "frankfurters, beef, boiled"  "ground beef, pan-cooked"
[10] "half & half cream"          "lamb chop, pan-cooked"       "lasagna with meat, homemade"
[13] "meatloaf, homemade"         "pork bacon, pan-cooked"      "salami, sliced"
[16] "sour cream"                 "Swiss cheese"               "vanilla ice cream"
[19] "whole milk, fluid"
```

dde, p,p'

```
 [1] "apple juice, bottled"                          "apple juice, strained"
 [3] "applesauce, bottled"                           "applesauce, strained/junior"
 [5] "bananas with tapioca, strained/junior"               "broccoli, fresh/frozen, boiled"
 [7] "cabbage, fresh, boiled"                        "creamed spinach, strained/junior"
 [9] "dessert, banana, apple"                        "dry table wine"
[11] "eggplant, fresh, boiled"                       "fruits, apples/applesauce w/apricots"
[13] "fruits, apricots, strained/junior"                  "fruits, bananas & pineapple"
[15] "fruits, pears and pineapple"                       "grape juice, from frozen concentrate"
[17] "grapefruit juice, from frozen concentrate"            "green beans, fresh/frozen, boiled"
[19] "green beans, strained/junior"                      "iceberg lettuce, raw"
[21] "jelly, any flavor"                             "juice, apple-banana"
[23] "juice, apple-cherry"                           "juice, apple-grape"
[25] "juice, grape"                                  "juice, mixed fruit"
[27] "lima beans, immature, frozen, boiled"                "mixed vegetables, frozen, boiled"
[29] "oatmeal, quick (1-3 min), cooked"                    "orange juice, from frozen concentrate"
[31] "orange juice, strained"                        "pear, canned in light syrup"
[33] "pears, strained/junior"                        "rice infant cereal, instant, prepared with whole
milk"
[35] "tomato juice, bottled"                         "tomato, stewed, canned"
[37] "white rice, cooked"
```

ethylenethiourea

Conclusions

- 7 of the 25 food groupings created by this method are of an interpretable size (between 5 and 40 foods), each of which has a unique “primary” pesticide contamination.
- The other 18 groups are overwhelmingly singletons, with exception to the 1 group which contains the vast majority of the foods in question (visible in the dendogram). That may be due to the sparse matrix and diverse cocktail of pesticides present in processed foods.
- Although FDA has monitored this data in cross-sections, a retrospective analysis like this one can help identify chronic exposure rather than acute alone.