Joy of Cooking

Wisam Barkho

Contents

The book *Joy of Cooking* is one of the most popular books in the United States. It has provided recipes for many American-favorite dishes since 1936. Each edition, eight in total, has seen recipes come and go, yet some recipes appear in each of the eight editions with minor changes, if any at all.

However, controversy brewed when in 2009, Brian Wansink of Cornell University published "The Joy of Cooking Too Much: 70 Years of Calorie Increases in Classic Recipes". In this publication, the authors conclude that "calorie density and serving sizes in recipes from *The Joy of Cooking* have increased since 1936". In 2018, this paper was retracted when an investigation found that academic misconduct had taken place.[4]

Problem Introduction

This school project was composed of two components. The first part required each student to create data tables for two assigned recipes from the book *Joy of Cooking*, each appearing in 1936 and 2006 (a total of four data files for each student). Each student was then to work on their own to merge each individual recipe into a single data table.

The second part consisted of conducting statistical analysis of the student's choosing. I chose to compare methods of calculating calories and included a subset of the Wansink data to determine if any bias exist. Method 1 calculates calories using the 4-4-9 method which used factors of 4 for calories from carbohydrates (CHO), 4 for calories from proteins (PRO), and 9 for calories from fat (FAT)[2]. Method 2 uses the Atwater factors from the USDA database, which are known to be more accurate.

Programming Languages/Software: R, RStudio

Skills Used: Text Processing and Parsing, Data Retrieval, Data Wrangling, Data Aggregation

Github Location: https://github.com/wisamb/JoyOfCooking

Part I: Data PreProcessing

Recipes from each student were exported as a tab delimited file with columns *Amount*, *Measure*, and *Ingredient*. *Ingredient* names were to match values from the USDA database (found on the USDA National Agricultural Library website - https://data.nal.usda.gov/search/type/dataset) and an *NDB_No* from the USDA database was included to identify each ingredient.

Importing the Data

The first challenge in this dataset is that each student had different naming conventions. Examples of differences encountered are different naming conventions for the required columns (ie, *Unit* rather than *Measure*), not having the minimum columns of *Amount*, *Measure*, and *Ingredient*, having additional columns besides *Amount*, *Measure*, and *Ingredient*, and some files were duplicates.

A *for-loop* is used to read in the data from all the files. Using the %in% operator, I was able to convert any nonconforming columns to specification. The recipe name and year are parsed from the file name per naming convention. Files that did not have the minimum three columns needed to be dropped. Finally, the recipe name and year are parsed from each file name per naming convention.

There are 2 versions of Hungarian Goulash so I removed one of them.

A total of 1250 ingredients were retrieved.

Matching *Ingredient* Entries

In order to calculate calories for each ingredient and recipe, the *NDB_No* value is matched with that in the USDA table. However, many files did not contain *NDB_No* values, so I retrieved these values from the USDA database using a custom function that matches the ingredient to that in the USDA database. Furthermore, the first letter of each entry in the USDA database is capitalized so I ensured each *Ingredient* entry in the merged table is also capitalized using a custom capitalization function. Finally, To continue filling NA values, I used another custom function to tokenize the *Ingredient* column and searched the USDA database for the closest match.

Using a *for-loop*, I check remaining NA values in the *NDB_No* column.

```
## [[1]]
## [1] "308"
##
## [[2]]
## [1] "Demi glace sauce concentrate, Master's Touch, 1263, food service"
##
## [[1]]
## [1] "310"
##
## [[2]]
## [1] "Sherry, cooking"
##
## [[1]]
## [1] "1040"
##
## [[2]]
## [1] "Brandy,90 proof"
##
## [[1]]
## [1] "1049"
##
## [[2]]
## [1] "Sherry,cooking"
##
## [[1]]
## [1] "1050"
##
## [[2]]
## [1] "Brandy,90 proof"
```

There are three results in five occurrences in the Ingredients column that do not have equivalents in the USDA database table. These ingredients are removed but not the recipe.

A new total of 1245 ingredients were retrieved.

Later on, if too many ingredients have been dropped for a particular recipe (>40%), then the recipe *for both years* will be dropped. Therefore, I count the number of ingredients in each recipe now.

There is a total of 188 recipes.

Matching *Measure* Entries

In order to find the correct values for the calories calculation, the entries in the *Measure* column will also need to match the USDA database. A custom function was used to standardize and correct various *Measure* entries, such as "tbs" or "Tbsp" to "tbsp".

Data Retrieval for Calorie Count

I then searched the USDA table for *Measure* entries to return the matching *Gm_Wgt* value, which is used for the calorie conversion. The *Amount* column is multiplied by the *Gm_Wgt* column to get *Grams* for each ingredient entry.

There are 349 remaining NA values. Although this will reduce the data by 28%, I removed these entries due to time constraints.

A total of 896 ingredients remain.

Data Retrieval for 4-4-9 Method

In order to calculate the total calories of each ingredient, the *Grams* column was multiplied by the nutrient content (a percent) to get grams of carbohydrate, protein, and fat. These weights were multiplied by the 4-4-9 factor to convert to calories. All three are added to determine total calories of each ingredient, and all ingredient calories are added to determine calories of each recipe.

Data Retrieval for Atwater Factor

This method is essentially the same as Method 1 except Atwater factors were used rather than 4-4-9. The Atwater factors were obtained from the USDA database.

Atwater factor is not provided for many ingredients and NA values must be removed to continue with analysis, resulting in another 209 ingredients being dropped.

A total of 687 ingredients remain.

Then, I removed any recipe that has lost more than 40% of its ingredients. If a recipe is dropped, it will be dropped for both years.

A total of 122 recipes remain.

Part II: Statistical Analysis

Method Comparison

I start by looking at the summary statistics for all four conditions by year.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   49.16  186.20  271.46  350.57 1738.30
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.05   70.69  199.48  282.60  374.06 1782.38
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   221.0   836.3  1189.1  1375.1  1821.0  3541.5
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   159.2   565.3  1034.6  1167.5  1595.3  3055.9
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   55.51  208.35  327.00  416.70 3681.13
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.05   53.60  203.58  306.75  407.15 3879.72
```

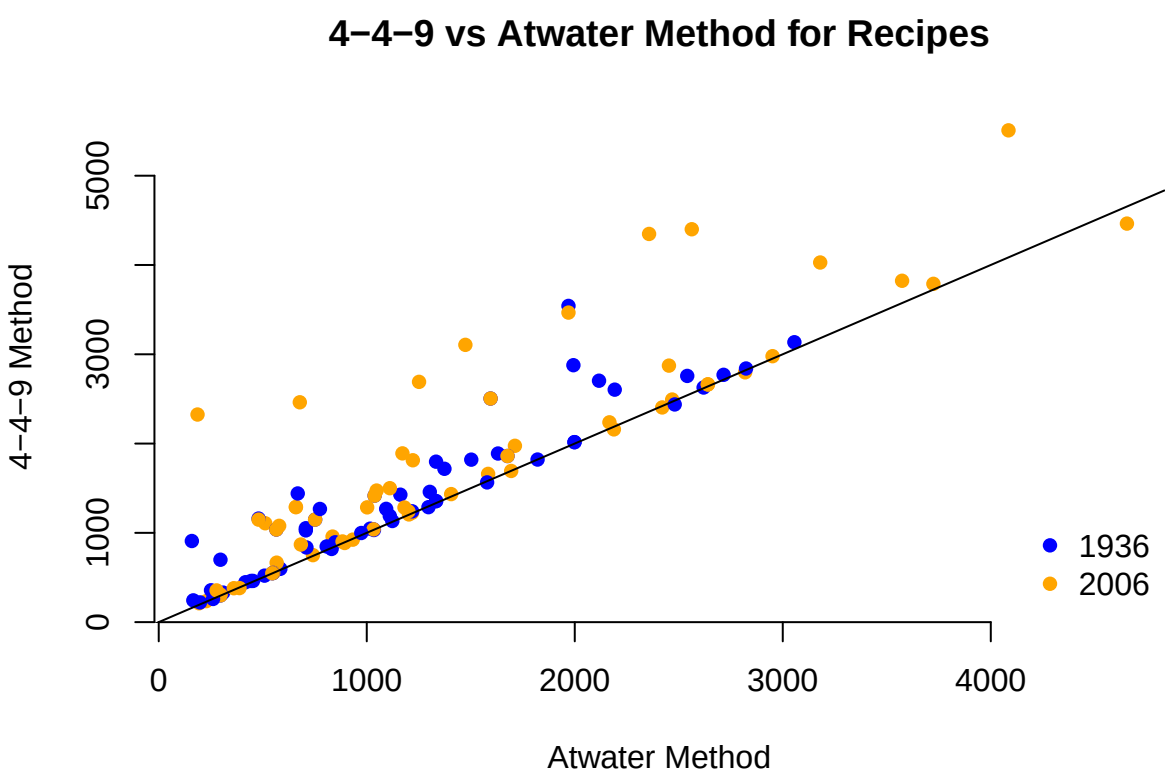
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   211.6   958.2  1433.8  1828.0  2504.5  5508.0
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   186.5   659.5  1171.6  1433.2  2166.6  4654.7
```

Mean and quantile values between the ingredient tables (for both years) are fairly close to one another (5-10% difference).

Mean and quantile values between the recipe tables differ (for both years) 15-30% with 4-4-9 values being consistently higher.

A scatterplot confirms this second finding and shows that the 4-4-9 method scores many recipes with higher calorie content than the Atwater method. Interestingly, a majority of the recipes that display the increase are from 2006.



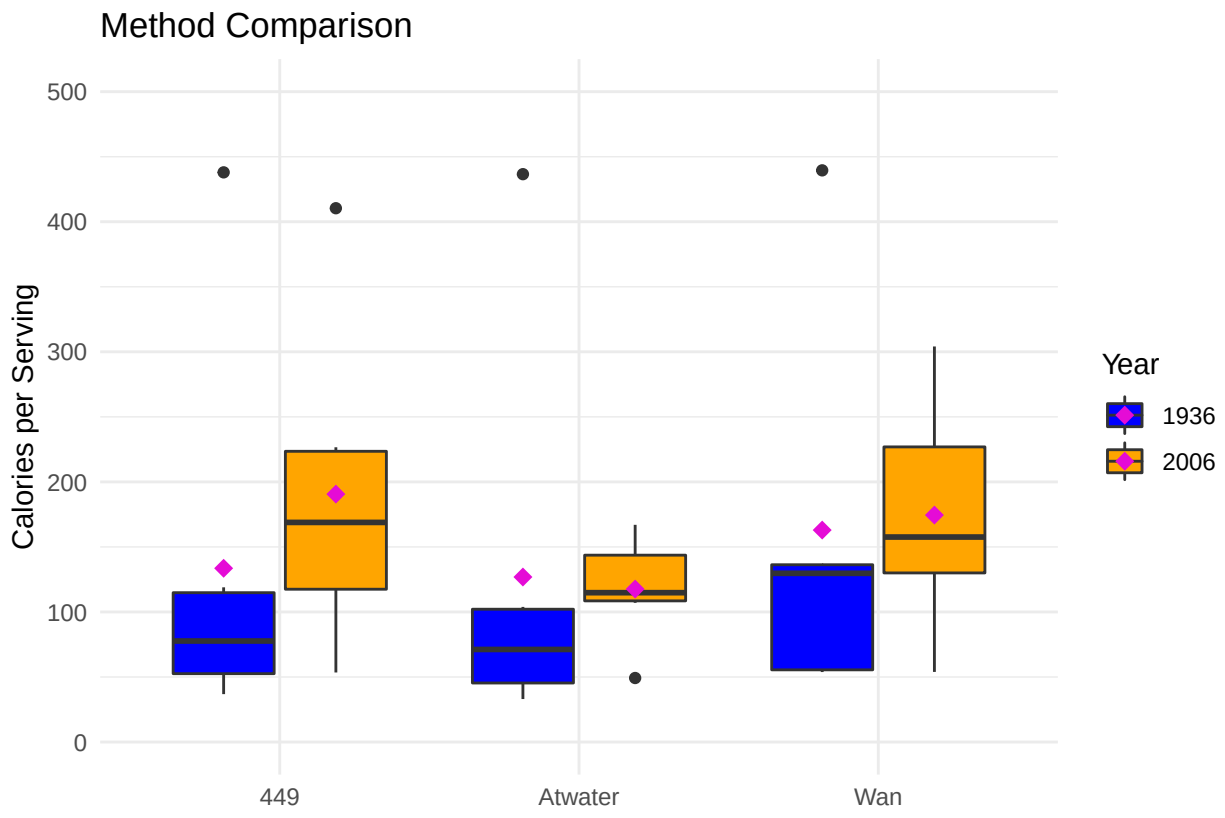
Wansink Data

I subset the Wansink data to determine if the Wansink Data uses the 4-4-9 method or the Atwater method. I find that the data contains 17 recipes although Wansink evaluated 18 recipes. Nevertheless, 18 recipes is a small sample from the 142 total recipes from *Joy of Cooking*. This small sample size introduces a bias and is a prevalent criticism of Wansink's work.

There are 6 remaining recipes that were matched between the two datasets. This is most likely due to the `NA` values for the Atwater factors. This is a small sample size and I may need to go back to collect more data.

Comparison with Wansink Data

Using boxplots, I compared the distribution of these six recipes for the 4-4-9, Atwater, and Wansink methods. This plot shows that the Wansink data matches the 4-4-9 method closer than the Atwater method. This can be a source of bias since the 4-4-9 method scores recipes with higher calorie content, namely for the year 2006. All three methods score recipes from 2006 higher than 1936. This could be due to the fact that there are only six recipes in this dataset. However, our scatterplot which has 122 recipes in the dataset shows something similar. Nevertheless, a more thorough analysis can be performed by incroporating the missing Atwater values.



References

[1] ESHA Research, Nutrition: General Database. 4-4-9. *Do you use 4-4-9 (449 or 944) to calculate Calories from the grams of carbohydrate, protein and fat?* Retrieved from https://esha.zendesk.com/hc/en-us/articles/202443626-4-4-9-Do-you-use-4-4-9-to-calculate-Calories-from-the-grams-of-carbohydrate-protein-and-fat-

[2] ESHA Research, Nutrition: General Database. *Why do I get a different amount of Calories when I use the 4-4-9 calculation?* Retrieved from https://esha.zendesk.com/hc/en-us/articles/203442937-Why-do-I-get-a-different-amount-of-Calories-when-I-use-the-4-4-9-calculation-

[3] Oransky, Ivan. "The Joy of Cooking, Vindicated: Journal Retracts Two More Brian Wansink Papers." *Retraction Watch*, 6 Dec. 2018, retractionwatch.com/2018/12/05/the-joy-of-cooking-vindicated-journal-retracts-two-more-brian-wansink-papers/.

[4] Wansink, Brian, and Collin R. Payne. "The Joy of Cooking Too Much: 70 Years of Calorie Increases in Classic Recipes." *Annals of Internal Medicine*, vol. 150, no. 4, 17 Feb. 2009, p. 291., doi:10.7326/l18-0647.