Tracing Trends in Macronutrient Intake and Energy Balance Across Demographics with Statistics and Machine Learning

Weida Zhu, Giovanni Pagano, Michela Taufer
(zhuweida,gpagano,taufer)@udel.edu
University of Delaware; Newark, DE 19711

ABSTRACT

Proteins, carbohydrates and fat are the three major contributors to energy intake in the human body. Demographic factors such as income and education level may influence what foods people eat, and lead to changes in nutrient and energy balance. In this project, we explored the relationship between macronutrient and calorie intake, using data from the NHANES database. We also set out to build a model to predict macronutrient intake based on demographic data. We used a model based on decision trees, which classify data into groups based on a given set of features. Due to time limitations, we oversimplified the model, leading to high error rates. We suggest ways in which our modeling process could be improved, such as more robust feature selection and alternative classification schemes. These improvements may increase the accuracy of our model, and allow us to explore which demographic features contribute most to the dietary habits of Americans.

MOTIVATION

While there are many components to a healthy and balanced diet, one of the most fundamental aspects is the intake of the three "macronutrients" proteins, carbohydrates and fats. Each is essential for the body to properly function, and a diet that has too much or too little of any of the macronutrients can have a significant effect on the short-term and long-term health of an individual.

Understanding how each contributes to the total "energy balance" can allow individuals and their medical providers to customize a diet for a person's lifestyle, as well as correct any imbalances that may lead to complications like diabetes or heart failure. However, people may not necessarily have the opportunity or ability to identify and make these changes. Reasons may include a lack of nutrition education, insufficient financial resources to purchase higher-quality food, or a lack of access to healthcare services. As a result, people with lower incomes or people without higher education may be more likely to have issues with maintaining a balanced diet.

If there are strong enough correlations between these demographic factors and their macronutrient intake, it may be possible to predict macronutrient intake for different populations in the US. To test this hypothesis, we built a system to evaluate the strength of education level and income level as predictors of macronutrient intake.

METHODOLOGY

Dataset of interest: National Health and Nutrition Examination Survey (NHANES)

NHANES is a biannual survey of American children and adults. The survey combines questionnaires and physical examinations to gauge the diet and health of respondents. Our statistical analysis focused on the datasets from 2003-2004 to 2013-2014, while the decision tree methodology was carried out on the 2013-2014 data.

Data download

Data was downloaded from the NHANES website with tools provided by Michael Wyatt. We also used these tools to convert the raw data from .sas format to .csv format [1].

Data preprocessing

Columns with data of interest (e.g. age, dietary recall, nutrient intake, education level, and poverty) were manually extracted from the .csv files. We removed all people with insufficient dietary recall, as they had no intake data to work with. To focus on the adults in the sample, we also filtered out people aged 19 and younger.

Statistical analysis

For our statistical analysis, we first calculated the sample mean and standard deviation of nutrient and calorie intake for each year to see how consumption varied over time. We also calculated the Pearson correlation coefficient for each macronutrient versus calories to quantify the strength of their relationships.

To determine if there were differences in deitary habits between previous years and the most recent survey (2014), we performed unpaired, two-sample t-tests for equal means. Based on the similarity of the variances between each sample, along with the fact that both samples were drawn from the US population, we assumed equal variances in our testing. We calculated two-tailed p-values with an alpha value of 0.05.

Decision tree classification

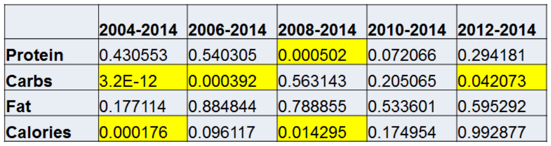
We focused on poverty ratio and education level in the demographic data for our prototype. We reclassified the data based on the preprocessed data set, using K-means clustering to calculate the boundaries for continuous variables (nutrient intake and poverty ratio). Categorical variables like education level were translated into a scheme with fewer classes.

We manually rearranged the .csv file and used a script from the phraug toolkit to convert the data into libsvm format[2]. The training and testing of the decision tree was performed using MLlib for each nutrient[3]. The data was randomly divided 70-30 into training and testing data by MLlib, and the error rate of classification was used to evaluate our models.

RESULTS

Statistical analysis

We observed some significant differences in mean carbohydrate, calorie and protein intake for the sample pairs tested (see figure 1). While the testing assumed equal variances, it is possible the US population changed enough over time to affect the variance. Other statistical tests like the F-test could be used to compare the variances and validate this assumption.



	2004-2014	2006-2014	2008-2014	2010-2014	2012-2014
Protein	0.430553	0.540305	0.000502	0.072066	0.294181
Carbs	3.2E-12	0.000392	0.563143	0.205065	0.042073
Fat	0.177114	0.884844	0.788855	0.533601	0.595292
Calories	0.000176	0.096117	0.014295	0.174954	0.992877

Figure 1: P-values for two-sample t-testing. Yellow boxes indicate statistically significant differences.

All three nutrients had a highly positive correlation with calorie intake for most years (see figure 2). This means that as nutrient intake increased, caloric intake increased as well. Interestingly, the correlation between protein and calories was weaker than carbohydrates and calories across all years, perhaps indicating a skew in the diets of the sample towards carbohydrates.

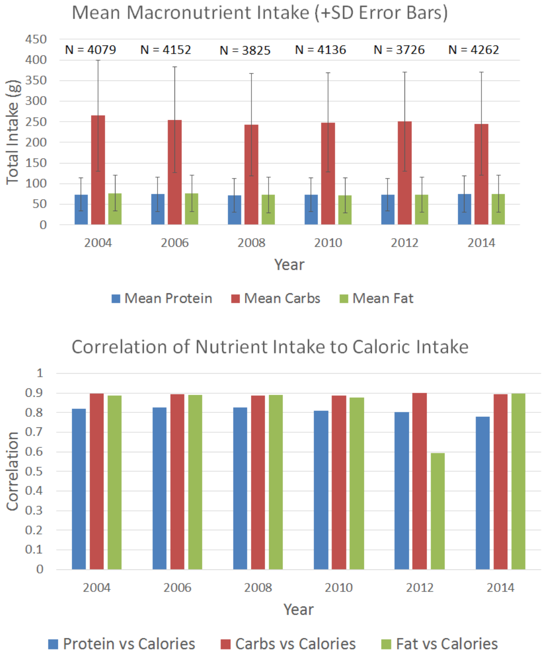


Figure 2: Summary statistics and correlation for each macronutrient, from 2004 to 2014.

Decision tree

Our classification model produced decision trees with error rates around 0.5 or 50 percent for each macronutrient. A simplified example of a decision tree model is shown in figure 3.

We believe there are two key factors that account for the high error rate of our decision tree model. First, we likely oversimplified our model, and should have also considered additional demographic categories like age or race. The categories we chose may also not be strongly correlated to nutrient intake, and are thus poor features to use for classification. Incorporating methods such as lasso regression into our analysis to evaluate the strength of demographic features as predictors would allow us to select an optimal set of features for our model.

Second, the assumptions we made when deciding how to reclassify the data could be flawed, and may not reflect the realities of how demographics relate to diet. Transitioning to a more coarse-grained classification scheme likely led to the loss of information that could provide more precise classifications. In addition, using K-means to define our cutoffs may not be the best method; determining boundaries based on our statistical analysis or literature review might be a more fruitful approach.

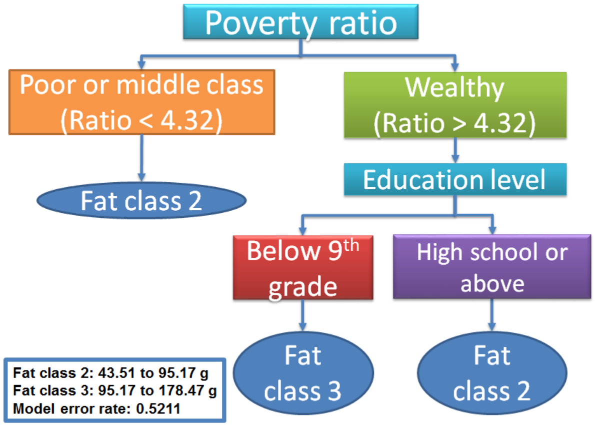


Figure 3: A sample decision tree using fat intake.

CONCLUSIONS AND FUTURE WORK

Future analyses could dig into the subcategories of macronutrients (e.g. saturated vs unsaturated fat) or layer on micronutrient data to get a better picture of the American diet. While we ran into problems with our decision tree prototype, the method is still practical for this type of work. With the changes we suggest above, it should be possible to increase the quality of the model and probe the connection between diet and demographics.

REFERENCES

[1]Github repository of mrwyattii/NHANES-Downloader URL: https://github.com/mrwyattii/NHANES-Downloader

[2] GitHub repository of zygmuntz/phraug URL:https://github.com/zygmuntz/phraug/

[3] Apache Spark URL:http://github.com/apache/spark