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

Giovanni Pagano[1] and Weida Zhu[2]

Advisor: Michela Taufer

[1] Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE 19711

[2] Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19711

Motivations

- One of the most fundamental aspects of a healthy diet is intake of the three “macronutrients”—proteins, carbohydrates and fats
- A diet with too much or too little of any macronutrient can have a significant effect on short-term and long-term health
- Understanding how each of these macronutrients contribute to the total “energy balance” can allow people to correct deficiencies and customize their diet to their lifestyles
- Not everyone can identify or make changes--lack of nutrition education, insufficient financial resources to purchase higher-quality food, or a lack of access to healthcare
- People with low income or people with no college degree may be more more challenged to maintain a balanced diet

Our Key Question:

Can we predict possible macronutrient intake for an individual based on the available demographics information?

Our Approach

- How does nutrient and energy intake change over time? Is there a correlation between macronutrient intake and total energy intake?
- How does macronutrient intake vary across demographics like gender, age, race, education level and income level? Are education level and poverty level strong predictors of macronutrient intake?

Data of Interest: National Health and Nutrition Examination Survey

- Biannual survey of US children and adults to gauge diet and health status
- Combination of questionnaires, dietary recall and physical examinations
- 2013-2014: 10,175 total participants

Data Preprocessing

- Conversion of files to CSV format with tools provided by Michael Wyatt
- Focus on the demographic and dietary (day 1 intake) data files:
  - SEQN ID, Age, Education Level, Ratio of family income to poverty, nutrient and calorie intakes
- Filtering of dataset / removal of missing data
  - Focused on adults, people 20 and older as defined in NHANES
  - Removed people with no nutrient data

Statistical Testing to Determine Trends

- Unpaired two sample t-tests for equal means; equal variances assumed
- Correlation calculated for each nutrient versus caloric intake

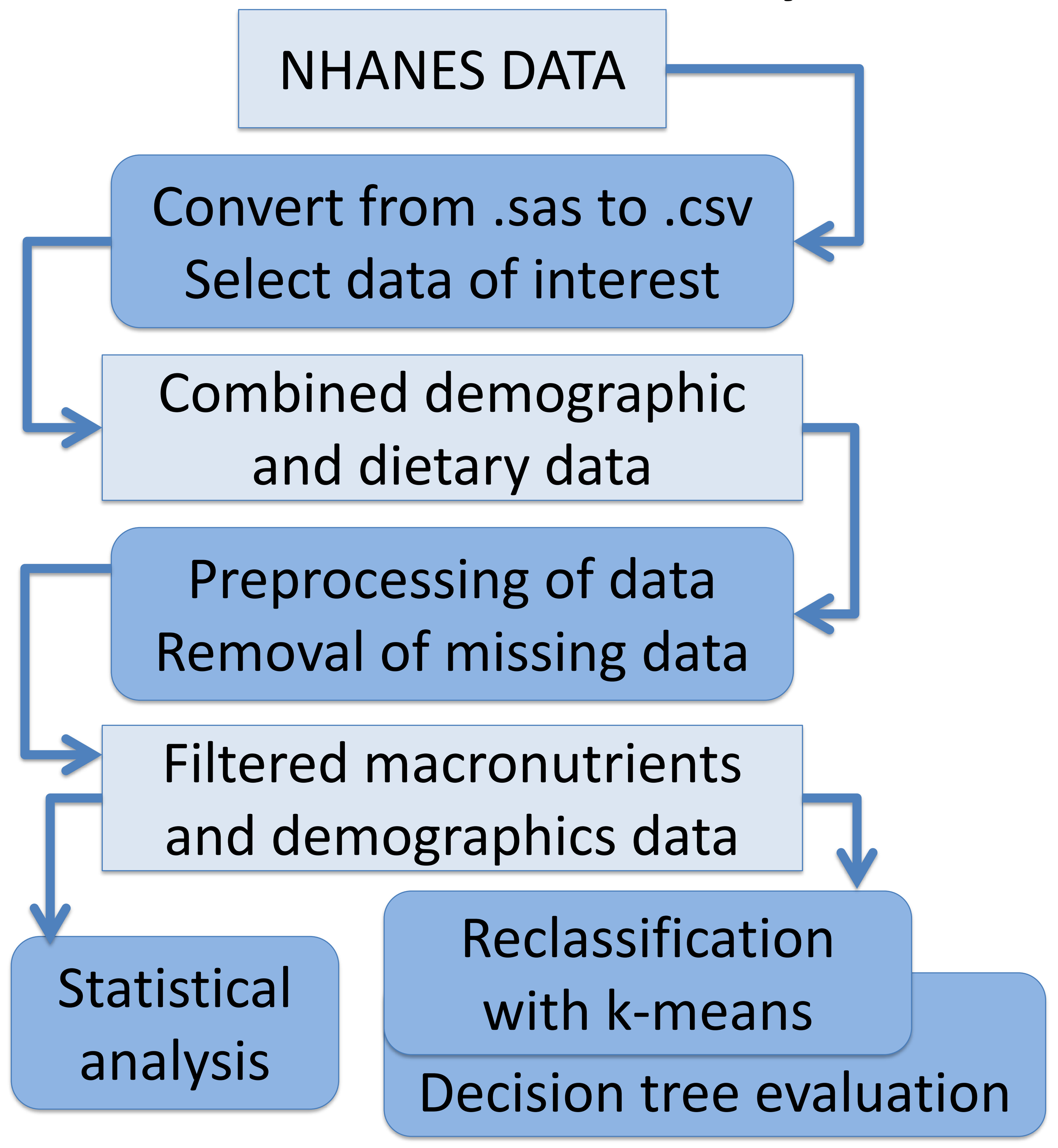
Reclassification of Demographic Data

- Simplify our testing for classification by converting continuous variables to categorical variables
- Used K-Means to establish boundaries—centroids used as cutoffs

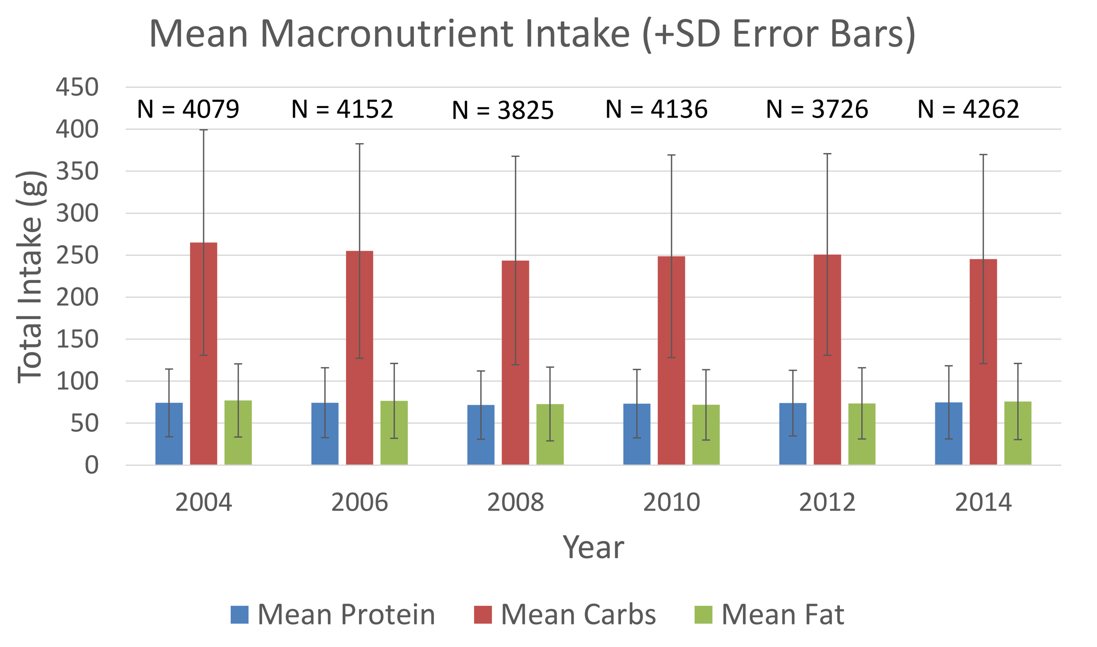
Classification with Decision Trees

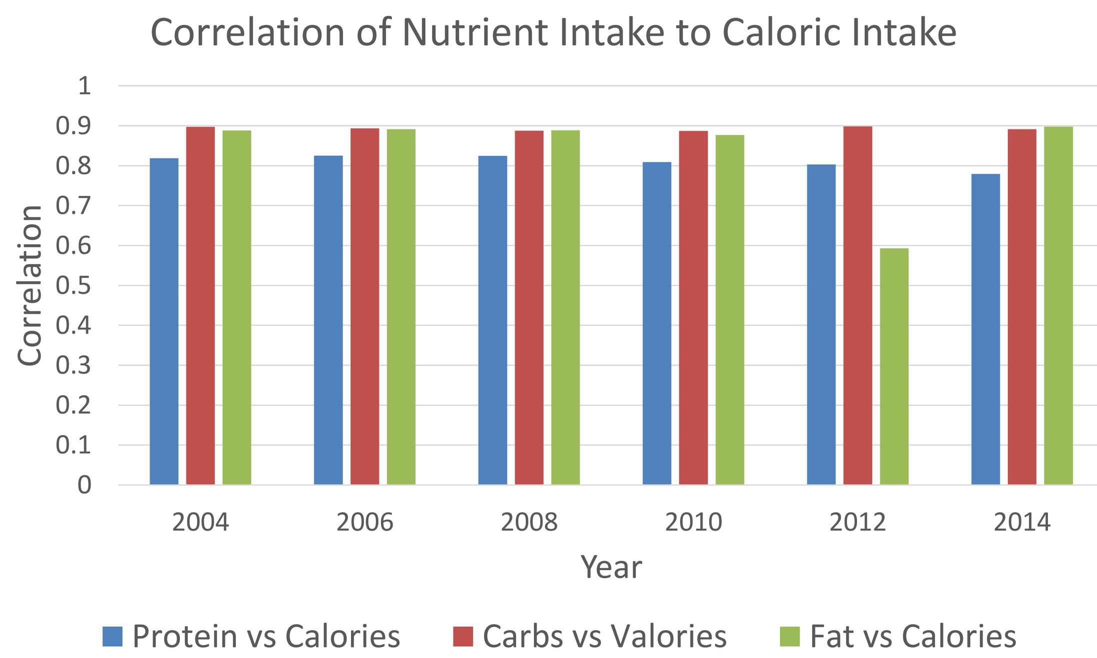
- A set of rules to classify data, based on a given set of attributes for the data
  - Class: Macronutrient intake, three separate tests
  - Attributes: Education Level and Poverty Ratio
- Randomly divide the data into training and testing sets, use the training data to build the tree, and classify the testing data using the rules
- Error rate used to assess quality of trees

Workflow of the NHANES Analysis



Nutrient Intake from 2004-2014



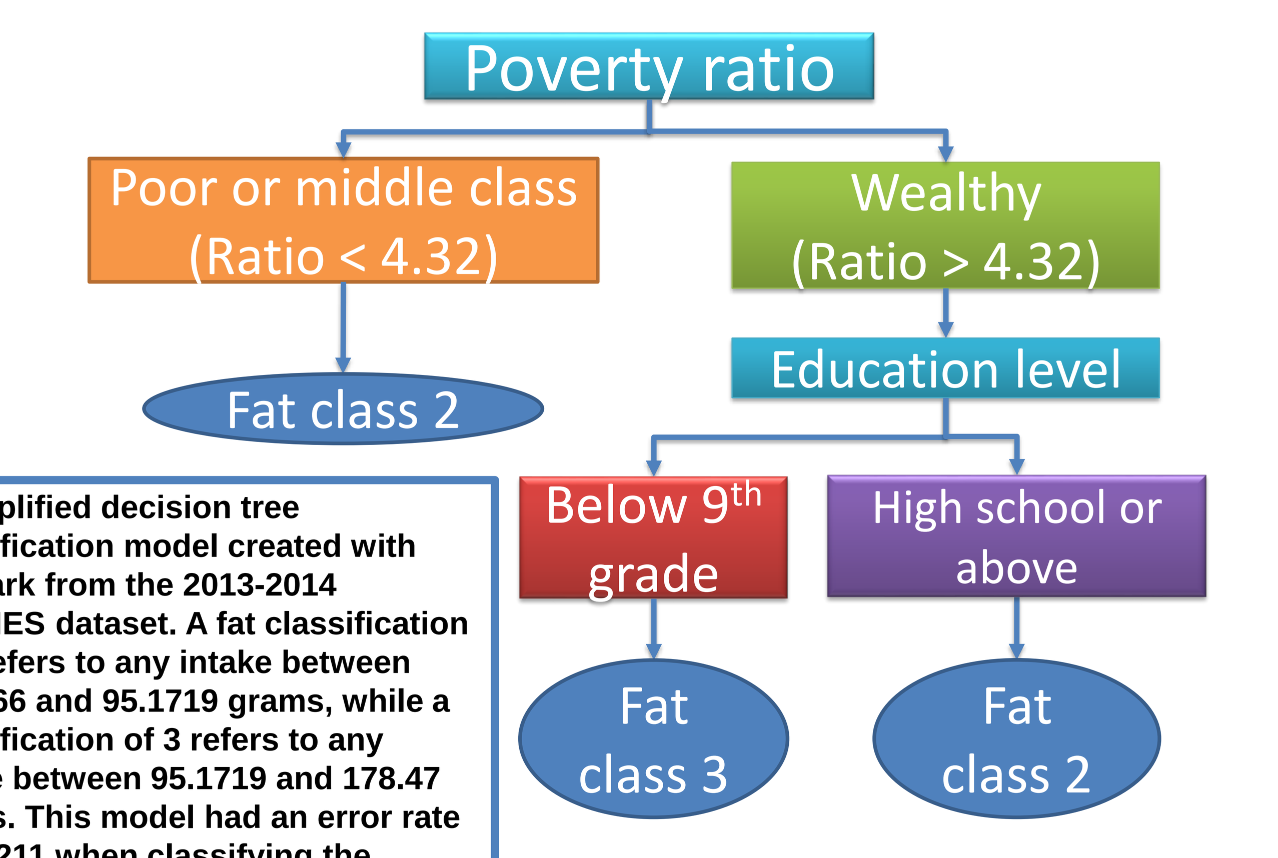


P-values for Equal Means T-Tests

	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

Results of unpaired, two-sample t-tests for equal means ($p < 0.05$) of nutrient and caloric intakes. Boxes in yellow represent combinations of years where there is a statistically significant difference between the sample means.

Sample Decision Tree for Fat Intake



A simplified decision tree classification model created with Pyspark from the 2013-2014 NHANES dataset. A fat classification of 2 refers to any intake between 43.5066 and 95.1719 grams, while a classification of 3 refers to any intake between 95.1719 and 178.47 grams. This model had an error rate of 0.5211 when classifying the testing data.

Results

- Some significant ($p<0.05$) differences observed, but not consistent for each pair
- Strong positive correlation between carbohydrate and calories, fat and calories
- Prototype of decision tree algorithm has a very high error rate for all nutrients (~50% error)
- Key reasons for decision tree errors
  - We made some assumptions when merging some classes in education level may not reflect the reality of education and nutrition knowledge
  - We considered only education level and poverty ratio; these features may be only weakly correlated to nutrient and calorie intake

Future Work

- Explore the relationship between carbohydrate intake and energy intake in depth—where are the carbohydrates/fat coming from?
- Decision trees are still practical for this type of analysis, but more time is needed to improve the error rate
- Suggestions to improve decision tree accuracy:
  - Keep encoding for education level rather than combining it, and aim for a finer-grain classification scheme
  - Take additional or other features into consideration for decision tree; use methods such as lasso regression to determine which features are the most important, and incorporate into the decision tree

Some of the software used in this project can be downloaded at: ; https://github.com/mrwyattii/NHANES-Downloader; https://github.com/zygmuntz/phraug/blob/master/csv2libsvm.py; https://hub.docker.com/r/zhuweida/docker/; https://github.com/apache/spark/blob/master/examples/src/main/python/mllib/decision_tree_classification_example.py