Study of Worldwide Food Consumption Pattern

Yuwei Xiong, *Computer Engineering, Columbia University, yx2385@columbia.edu*
Yilan Ji, *Electrical Engineering, Columbia University, yj2425@columbia.edu*
Zhuo Kong, *Computer Engineering, Columbia University, zk2202@columbia.edu*

***Abstract*—Food balance sheet shows the a brief picture of the pattern of the food supply of a country. Based on food balance sheet, our project analyzes the dietary pattern of countries through 50 years, clusters and builds models which showed relationship with the GDP per capita and IBM index of countries and regions.**

***Index Terms*—Food Balance Sheets, Diet Pattern, K-means Clustering, Linear Regression, PCA, Echarts**

I. INTRODUCTION

The Food Balance Sheet (FBS) is a data set collected by the Food and Agriculture Organization (FAO) which contains data regarding national food supply. The FBS presents a comprehensive picture of the pattern of a countrys food supply during a specified reference period by showing sources of supply and their utilization. These data are widely used and cited by public health officials and aid or development agencies[8]. The FBS data have been an important contributor to analysis and policy creation. The FBS comprise a very large set of information about food supply and use. The files include more than 180 countries (and territories) and about 100 different food commodities. The FBS include estimates of the nutritional characteristics of food available, including calories/capita/day, fat/capita/day, and protein/capita/day. These are obtained by applying nutritional tables and related information to the amount of each commodity estimated for food use and dividing by eatimates of national population.

The FAO suggests various uses for their FBS[9], including a) observe a country's food supply and its trends, b) compare food supply to nutritional requirements for healthy diets, c) estimate supply/shortage measures, d) evaluate food and nutrition policies, e) measure the degree of chronic undernutrition, e) examine changes in diet patterns, f) investigate relationships between food supplies, famine, and malnutrition, g) calculate self-sufficiency and import-dependency ratios, h) set goals for trade and production, and project future supply and demand. In this report, we focus on the usages related to diet patterns, including finding food consumption patterns that lead to healthy diets, and examing changes in diet patterns.

There may be a relationship between lifestyle including food consumption and potentially lowering the risk of cancer or other chronic diseases. For example, a diet high in fruits and vegetables appears to decrease the risk of cardiovascular disease and death.[5]An unhealthy diet is a major risk factor for a number of chronic diseases including: high blood pressure, diabetes, abnormal blood lipids, overweight/obesity, cardiovascular diseases, and cancer. [6] The WHO estimates that 2.7 million deaths are attributable to a diet low in fruits and vegetables every year.[6] Food consumption pattern varies by country as a result of diversity in food culture, agriculture and geography, economy, and population.

The World Health Organization (WHO) makes the following 5 recommendations with respect to both populations and individuals[7], including a) Eat roughly the same amount of calories that your body is using and maintain a healthy weight.b) Limit intake of fats, and prefer unsaturated fats to saturated fats and trans fats. c) Increase consumption of plant foods, particularly fruits, vegetables, legumes, whole grains and nuts. d)Limit the intake of sugar. A 2003 report recommends less than 10% of calorie intake from simple sugars.e) Limit salt / sodium consumption from all sources and ensure that salt is iodized.

II. RELATED WORKS

Big data analytic is a popular method in health care studying.[1] Kromhout[10] had done study on food consumption patterns in the 1960s based on the FBS data. Smith[2] and Svedberg[3], [4] detail how the FAO uses FBS figures for daily per person caloric availability as the mean of a log-normal distribution of each countrys caloric availability from which is determined a citizen of that countrys probability of not meeting a minimum dietary energy requirement. (The spread of the distribution is determined by the variability in dietary intake over a countrys population that is determined by a household survey.) Smith argues that under this methodology the prevalence of under-nutrition is more or less determined by the figure for per capita daily energy supply which is taken from the FBS. Svedberg argues that such estimates of undernutrition are overly sensitive to the daily food energy supply. Pinstrup-Andersen[11] compiled a list of all countries with daily per capita caloric intake less that 2200 calories a day as a loose index of undernourished countries. However, there may be enough error in the underlying data that one must use any ranking with extreme care. Svedbergs sensitivity analysis, which demonstrated notable changes in prevalence of undernutrition constructed using FBS caloric figures, suggests that errors well with a plausible range could easily change rankings substantially. Clearly, whenever household consumption or anthropometric data are available, it may be more appropriate to base estimates of under-nutrition on these data rather than food balance sheet data.

Additionally, medical research has used FBS data to investigate connections between diet and health, especially cardiac health and cancers [12], [13], [14]. Medical researchers have also evaluated the usability and relevance of FBS data. For example, Sasaki and Kesteloot examined correlations between FAO data and data from multiple surveys in 19 countries and

deemed the FBS data usable and valuable. It should be noted however, that the majority of the studies are for developed countries, which may have clearer data and methodology.

III. SYSTEM OVERVIEW

We used data from newest FBS collected by FAO to evaluate food consumption based on per capita food supply in terms of calories. All our results were visualized as a website for viewers' convenience.

Different commodities are classified into 5 categories, and the percentage contributions of each category's energy supply were taken as features indicating different diet patterns. We clustered over 180 countries in a 4 dimensional parameter space, generating 5 clusters of diet pattern. The distribution of clusters by country each year was marked on a world map.

We also collected properties of obesity rate (based on the Global Database on Body Mass Index (BMI) data set from World Health Organization) and country's economic performance (based on the World Development Indicators data set from World Band Group) in each cluster to illustrate a connection between diet pattern and health as well as diet pattern and economy. By comparing diet pattern and obesity rate of different clusters, we suggested some features of a healthy diet.

To examine the changes in diet pattern over the 50 years' period from year 1961 to year 2011, we selected several typical countries and displayed lined graphs on every year's energy supply.

IV. ALGORITHM

A. Preprocess dataset

In order analyze the food consumption problem, we use Pig/Hadoop and Spark to do query, getting data for further use including:

- Data includes food supply (kcal/capita/day) of various kinds of food in countries in several years.
- Data includes food supply (kcal/capita/day) of selected countries from 1961 to 2013.
- Data includes BMI overweight index of countries and regions in several years.
- Data includes GDP per capita of countries and regions several years.

B. Analyze data

According to WHO, various kinds of food can be generally categorized into 5 kinds: {Cereals, Veg&Fruits, Roots, Protein, Sugar&Oil }. Using pySpark to organize detailed kinds of food and cluster countries and regions into 5 clusters based on 5 dimension feature space above. Each cluster center represents the similar dietary structure of countries categorize in this cluster.

Then, wondering whether the dietary structure is related to GDP per capita in each country, we calculated average GDP per capita of countries in each cluster. The relationship between dietary structure and GDP per capita in each country is obscure, as it is nothing more than random guess based

```
In [238]: runfile('/Users/Jillian/Desktop/fbs_cluster_gdp.py', wdir='/Users/Jillian/Desktop')
/Users/Jillian/Desktop/fbs_cluster_gdp.py:36: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-
view-versus-copy
  f[col] = [float(i) for i in f[col]]
[[ 0.44458877  0.06510286  0.15691969  0.06550885  0.19003443]
 [ 0.43796173  0.05932367  0.02795207  0.16640374  0.24140623]
 [ 0.60856564  0.04336103  0.04164461  0.08396516  0.17105412]
 [ 0.28655155  0.06221835  0.04182344  0.22827165  0.29903184]
 [ 0.2981781   0.08038426  0.31976164  0.05843269  0.16748683]]
Average GDP in each cluster:
 [8392.9156263310251, 12477.783038231932, 9330.377569804552, 13291.75182246234, 12762.085247085879]
Coefficients:
 [ 0.00098593 -0.01610046  0.00222731  0.01453852 -0.00027764]
Mean predict GDP:  13.6801736755
Mean variance of error: 16.11
Variance score: -0.05
Coefficients:
 [-0.00144232  0.00065897  0.00072417  0.00323101  0.00176979]
Mean predict GDP:  12.0498815686
Mean variance of error: 1.12
Variance score: 0.60
Coefficients:
 [ 0.00256733  0.04088927 -0.00568866 -0.00437098  0.01735164]
Mean predict BMI(>25)%:  46.5569247711
Mean variance of error: 14.81
Variance score: 0.02
explained_variance_ratio_= [ 0.46810454  0.2487622   0.07796175]
get parameters = {'random_state': None, 'whiten': False, 'iterated_power': 'auto', 'n_components': 3,
'tol': 0.0, 'copy': True, 'svd_solver': 'auto'}
```

Fig. 1. Analyze data

on the results of linear regression model. Nevertheless, the relationship between dietary structure of each country and the average GDP per capita of the cluster which the country belongs to based on the results of linear regression model. Results indicate that each cluster can be identified by average GDP.

Furthermore, dietary structure is related with BMI index as common sense. We calculated the average BMI distribution of countries in each cluster. Comparing the average BMI distribution and dietary structure, we can get insight as mentioned in discussion. To further understand the relationship between various kind of food and specific BMI index(Overweight), we build a linear regression model to indicate the quantitative relation and use PCA model to select significant kinds of food related to the overweight index.

C. Visualize results

We use Matlab to dump the 5 cluster center points into pie chart, as well as average BMI structure in each cluster. Then we analyze the develop tendency of the dietary structure of specific countries for 42 years. Finally, we use Apache+PHP+Mysql+Echarts to build a website and show the distrbution of 5 clusters on a world map while showing the results of analysis on GDP and BMI index at the same time.

V. SOFTWARE PACKAGE DESCRIPTION

Machine learning tools: scikit learn in python.

- KMeans:
  http://scikit-learn.org/stable/modules
  /generated/sklearn.cluster.KMeans.html
- LinearModel:
  http://scikit-learn.org/stable/modules/
  generated/sklearn.linear_model.LinearRegression.html
- PCA:
  http://scikit-learn.org/stable/modules/
  generated/sklearn.decomposition.PCA.html

VI. EXPERIMENT RESULTS

The percentage contribution of different categories of commodities was visualized for each cluster center as shown in the left halves in Fig 2, 3, 4, 5, 6 (the data of year 2010 was selected). It could be observed that for all clusters, cereals play an important role in energy supply. Sugar and oil is also

a main source of energy, according to the result of year 2010, it takes up 36% (rank first) in cluster 5, and ranks top 3 for other clusters. Roots contributes dominantly in cluster 1, but plays a minor role in other clusters.

As shown in Fig 7, countries in the world map were colored with 5 different colors illustrating different food consumption patterns they belong to. We observed that many geographic similar regions were distributed to the same food pattern cluster, which is reasonable considering that the neighboring countries are likely to have similar cultural background and agricultural conditions. For example, many countries in Southeast Asia are in either orange or gray, i.e. cluster 2 and cluster 4. Cereals are the ranking first energy source for both clusters, coincidentally, Southeast Asia countries are famous for a variety of rice featured menu. Geographically, the climate in Southeast Asia is humid and mild, suitable for crops planting.

The average BMI of each cluster was shown in the right half together with the food consumption pattern. We compared the diet pattern of each cluster with BMI distribution. For cluster 2 and cluster 4, the combined contribution of protein and sugar&oil takes up similar percentage of total energy supply and the percentage is rather small. As shown in BMI graph, the contributions of obesity in these two clusters are relatively low, both below 10%. In cluster 1, cereals take up as much as 63% of total supply, ranking first in all clusters. Accordingly, the percentage of normal and underweight BMI is the highest for cluster 1 among all clusters. By comparing cluster 3 and cluster 5, it could be observed that both clusters share a similar percentage of sugar&oil, veg&fruit and roots, while the cereals-to-protein ratio of cluster 3 is much higher than cluster 5. Accordingly, the percentage of pre-obsessed is higher in cluster 5. Compare cluster 4 and cluster 5. Both clusters share a similar percentage of veg&fruit, roots and cereals, while the sugar&oil-to-protein ratio of cluster 4 is much higher than cluster 5. Accordingly, the percentage of obsessed and pre-obsessed is higher in cluster 4.

Combining the diet pattern with the average GDP of each cluster, it shows that higher the contribution of protein, higher the GDP. It could also be observed that the combined contribution of cereals and roots is higher in countries with lower GDP.

In conclusion, a diet high in cereals, efficient in protein and relatively low in sugar&oil may help to drive you away from obesity and remain a fit shape.

The line graphs as Fig 8,9,10 showed the trends and changes of the pattern of the per capita food supplies of selected countries, during a 50 years period from year 1961 to 2011.

For many developed countries, such as USA, the grand total of energy consumption increases a little during the first half of the period and remain stable for the second half. The diet structure is also stable in those countries.

For some developing countries who see a major development in economy during the period, such as China, the graph shows large increasing in grand total as well as the percentage contribution of meat.

For some countries who depend mainly on domestic agricultural resources, such as North Korea, the line graph of energy supplies shows sharp changes during the period, where major reductions coincident with natural disasters.

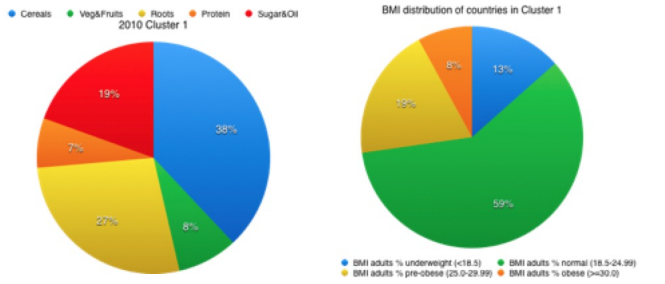

Fig. 2. Food consumption pattern (left) and average BMI percentage contribution (right) for cluster 1 in year 2010.

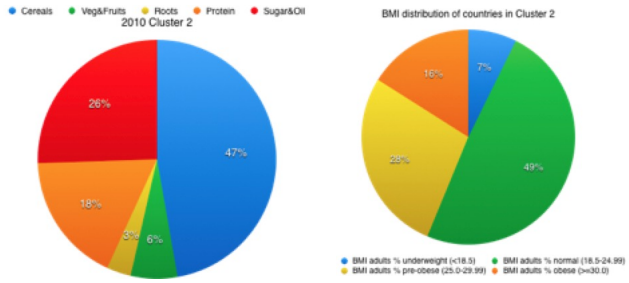

Fig. 3. Food consumption pattern (left) and average BMI percentage contribution (right) for cluster 2 in year 2010.

VII. CONCLUSIONS

A. Limitation

The data set has some limitation itself considering our object. Even if FBS are taken as approximation of per capita consumption, the amount of food actually consumed may be lower than the quantity shown, depending on the magnitude of wastage and losses of food in the household, e,g, during storage, in preparation and cooking, thrown or given away; They do not give any indication of the differences that may exist in the diet consumed by different population groups, e.g. people of different socio-economic groups, ecological zones or geographical areas within a country; Neither do they provide information on seasonal variations in the total food supply.

The average GDP is computed after clustering by diet pattern, so the data is not able to imply the affect of economic performance on diet pattern. The BMI distribution could have other factors such as total fat consumption, considering merely diet pattern is imperfect.

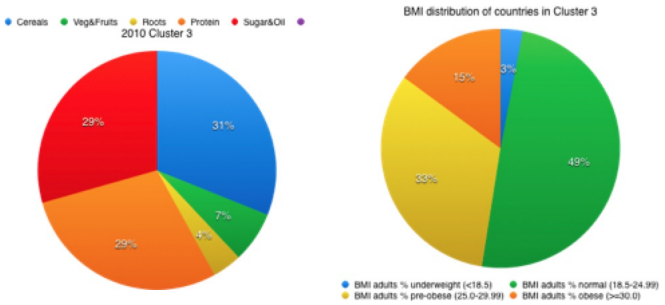

Fig. 4. Food consumption pattern (left) and average BMI percentage contribution (right) for cluster 3 in year 2010.

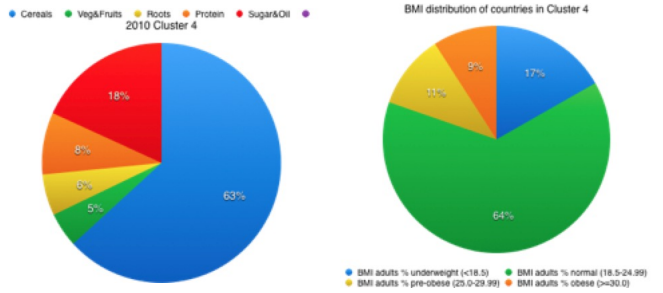

Fig. 5. Food consumption pattern (left) and average BMI percentage contribution (right) for cluster 4 in year 2010.

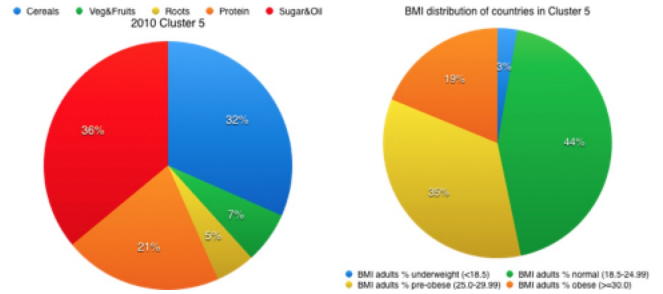

Fig. 6. Food consumption pattern (left) and average BMI percentage contribution (right) for cluster 5 in year 2010.

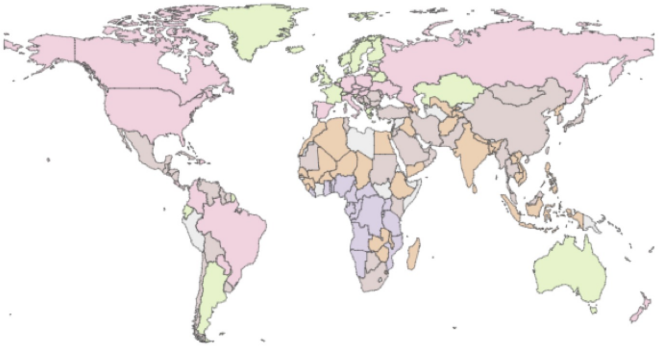

Fig. 7. Distribution of food consumption pattern by country in year 2010.

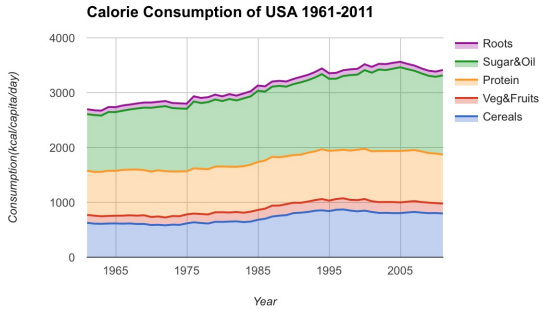

Fig. 8. Changes of the pattern of the per capita food supplies of USA from 1961 to 2011.

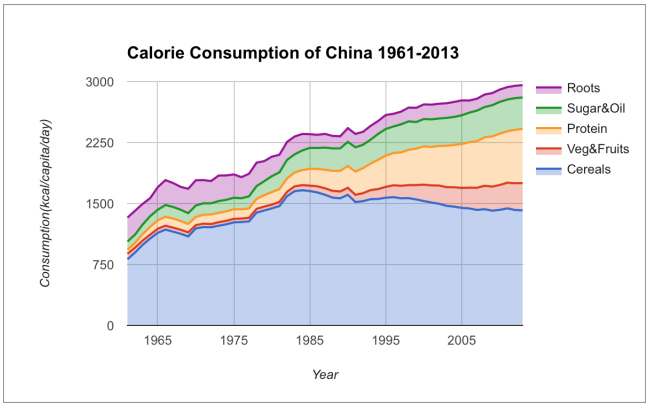

Fig. 9. Changes of the pattern of the per capita food supplies of China from 1961 to 2013.

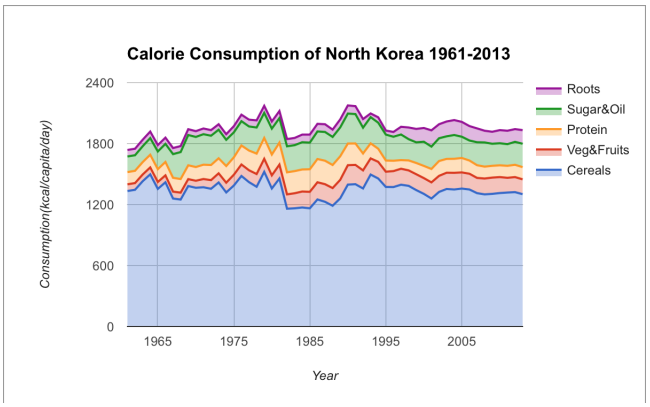

Fig. 10. Changes of the pattern of the per capita food supplies of North Korea from 1961 to 2013.

B. Improvement

To obtain a complete picture, food consumption surveys showing the distribution of the national food supply at various times of the year and among different groups of the population should be conducted. We can further study this topic by clustering GDP for countries instead of calculating the average GDP and compare it with the diet pattern clustering result.

C. Personal Contributions

Yuwei Xiong did the data preprocessing, visualized the line graphs and wrote part of the report. Yilan Ji processed and analyzed the data, using models including clustering, linear regression and PCA, then visualized the pie charts and wrote the algorithm part of the report. Zhuo Kong developed the web page project and visualized the world maps. All teammates contributed equally in determining the object and searching for data set.

ACKNOWLEDGMENT

Thanks to Prof. Lin, for his instructions on data analyzing methods and algorithms. Thanks to Eric Johnson and TA team, for their detailed instructions of using data analyzing tools.

REFERENCES

[1] Raghupathi, Wullianallur, and Viju Raghupathi. "Big data analytics in healthcare: promise and potential." Health Information Science and Systems2.1 (2014): 1.

[2] Smith, Lisa C. "Can FAO's measure of chronic undernourishment be strengthened?." Food Policy 23.5 (1998): 425-445.

[3] Svedberg, Peter. "841 million undernourished?." World Development 27.12 (1999): 2081-2098.

[4] Svedberg, Peter. "Undernutrition overestimated." Economic Development and Cultural Change 51.1 (2002): 5-36.

[5] Wang, Xia, et al. "Fruit and vegetable consumption and mortality from all causes, cardiovascular disease, and cancer: systematic review and dose-response meta-analysis of prospective cohort studies." (2014): g4490.

[6] World Health Organization (WHO). "Diet and physical activity: a public health priority. Geneva; 2016 cited 2016 Dec 22."

[7] Who, Joint, and FAO Expert Consultation. "Diet, nutrition and the prevention of chronic diseases." World Health Organ Tech Rep Ser 916.i-viii (2003).

[8] Jacobs, Krista, and Daniel A. Sumner. "The food balance sheets of the Food and Agriculture Organization: A review of potential ways to broaden the appropriate uses of the data." Food and Agriculture Organization, Rome (2002).

[9] Sheets, Food Balance. "A Handbook." Food and Agriculture Organization of the United Nations: Rome, Italy (2001).

[10] Kromhout, Daan, et al. "Food consumption patterns in the 1960s in seven countries." The American journal of clinical nutrition 49.5 (1989): 889-894.

[11] Pinstrup-Andersen, Per. "World food trends and how they may be modified." (1993).

[12] Sasaki, Shatoshi, and Hugo Kesteloot. "Value of Food and Agriculture Organization data on food-balance sheets as a data source for dietary fat intake in epidemiologic studies." The American journal of clinical nutrition 56.4 (1992): 716-723.

[13] Helsing, Elisabet. "Traditional diets and disease patterns of the Mediterranean, circa 1960." The American journal of clinical nutrition 61.6 (1995): 1329S-1337S.

[14] Kelly, A., W. Becker, and E. Helsing. "Food balance sheets." Food and health data: Their use in nutirition policy-making: WHO Regional Publications, European Series, No34 (1991): 39-48.