Connecting Diets to Disease

Using Data-Mining to Find Links between Food Consumption and Chronic Diseases

A Senior Thesis submitted in partial fulfillment of
The requirements for the degree of

Bachelor of Science
With Departmental Honors

Computer Science & Engineering
University of Washington

June 2007

Presentation of Work Given on: May 18th, 2007 at the Undergraduate Research Symposium

Thesis and Presentation Approved by:

Date:

Table of Contents

Abstract

The intersection between food and disease provides insight into the fundamental interaction between societies and their environments. Societies, economies, and public health trends are tied together through the complex interactions between people, nature, technology, and food. While many studies have explored the health outcomes of consuming specific foods or diets, we explore a more general framework to evaluate the effects of changing food consumption on disease trends. This framework, built on a combination of a query-central and statistics-central view of World Health Organization and Food and Agriculture Organization data, is a useful tool for nutritionists, epidemiologists, and public health researchers to look for insight into food and disease trends.

Introduction

Food is perhaps the most direct and unbreakable interaction between the natural environment and human societies. It is a major component of culture, and is a vital component of health and security discussions. Food not only plays a role in the personal health of every person, but also plays a vital role in defining civilizations.

In 1997, Jared Diamond wrote at great length in his book "Guns, Germs and Steel", regarding the role that various foods played in determining the outcomes of civilizations. Domesticated plants and animals are the key to building a complex civilization. Easily domesticated plants allow for a quick excess of food supplies, while domesticated animals provide many services, among which are diseases. Most of the major modern diseases, including smallpox, influenza, the common cold, and many others, are adapted forms of diseases originally found in domesticated animals. The proliferation of these diseases in certain civilizations but not others led to large differences in immunity. The most prominent consequence of this occurred when the Europeans visited the Americas, introducing smallpox, which historians believe killed roughly 95% of the Native American population that had existed before the explorers came [Diamond 1997].

Modern excess of food has led to a unique problem. While some of the world still suffers from malnutrition and low food availability, most of the modern world now is able to select from among a wide array of foods. Instead of food availability being the key driver behind food consumption trends, the wider availability of a large variety of foods has led to dietary transitions on an international scale. These transitions usually move cultures away from traditional, complex diets towards modern, high-calorie diets [Albala 2001]. The spread of high calorie diets have turned the reversed the problem of malnutrition, making the modern food problem one of over-consumption. Over-consumption has led to increases in diseases like diabetes, obesity, cardiovascular diseases, and a host of other chronic diseases, which now characterize the "malnutrition" of the modern era.

The literature regarding food consumption trends and their effects on human health is rich with studies reviewing connections between specific diets or foods and their effects on health. Examples include a study analyzing the nutritional characteristics of an Italian population [Barbagallo 2002], the benefits of the Greek diet [Simopoulos 2001], and nutritional transition in

Chile [Albala 2001]. These kinds of studies focus on a specific population and diet or dietary transition, analyzing the effects of specific food groups on specific disease outcomes. While these studies are valuable and scientifically important, they do not compose an efficient method for testing for significance among the millions of food-disease combinations.

Elisabet Helsing established the groundwork for connecting global food data with global disease data through her work that compares World Health Organization (WHO) data to Food and Agriculture Organization (FAO) data for Mediterranean countries between the 1960's and the 1990s. Her article, "Traditional diets and disease patterns of the Mediterranean, circa 1960" stops short of developing a formal model for comparing food and disease trends, and for determining significance of any trends that would emerge [Helsing 1995].

We approached the question of how to find meaningful links between food and disease on a global scale through the use of a broader statistical framework. We developed a query-oriented and statistics oriented database that provides a more meaningful view of world food and disease data, allowing us to explore a variety of questions. We hope to obtain meaningful results for nutritionists and epidemiologists to explore in the data.

Methods

We identified two different kinds of questions that we wanted to answer with new frameworks:

1. Specific Queries – Questions about the data that pertained to a small set of countries, years, diseases, or foods, for the purpose of discovering a specific trend or dataset. For example "Compare the diabetes levels to sugar consumption in Thailand between 1960 and 1990" In order to answer these questions, we chose to develop a set of Query Oriented database tables.
2. Statistical Relationships – Broader questions regarding statistical trends that encompassed a large number of countries, years, foods and diseases. For example, "What is the impact of increased sugar consumption on diabetes?" In order to answer these questions, we chose to develop a set of statistics oriented database tables.

Data Sources

Data came primarily from two sources, both under the larger branch of the United Nations. Because we wanted to compile valid information across many countries and years, few data sources existed that could be compiled consistently.

Mortality

Originally our intention was to track chronic disease incidence rates directly, however, chronic disease incidence tracking across nations is not reliable. We deferred to the use of mortality data as a rough proxy for disease incidence. While this introduces bias, mainly because of uneven access to health care, it is a reasonable approximation in light of the scale of the data.

The United Nations makes worldwide mortality data freely available to the public through the World Health Organization (WHO). The WHO data is already in CSV form, but required heavy

formatting to get into a query oriented form. The scope of the data is large, comprising roughly 50 million data-rows, separated by 219 countries, over 10,000 causes of death, gender, age, and list.

The backbone of the WHO Mortality data is the International Classification of Diseases. The ICD is used to as a set of universal identifiers for diseases. The ICD has evolved over time, forming several different ICD lists. There are several major lists, but we have merged them into several smaller lists for coherency. The lists are presented in Table 1.

Table 1: International Classification of Diseases Lists

List	List of causes (condensed/detailed) from the revisions of the International Classification of Diseases
07A	ICD 7th revision, List A (condensed)
07B	ICD 7th revision, List B (condensed)
08A	ICD 8th revision, List A (condensed)
08B	ICD 8th revision, List B (condensed)
09A,09B	ICD 9th revision, Basic Tabulation List (condensed)
09N	ICD 9th revision, Special List of causes (condensed) as reported by some countries of the newly independent States of former USSR
09C	ICD 9th revision, Special List of causes (condensed) as reported by China
101	ICD10 Mortality Tabulation List 1(condensed)
103	ICD10 3 (detailed) character list
104	ICD10 4 (detailed) character list
10M	ICD10 3 and 4 (detailed) character list. Pls note that when a 4th character code is given it is therefore not included in a 3 character code. All records are mutually exclusive.

In order to find useful data over long stretches of time, a person must connect ICD CauseID's from one list to the corresponding ICD CauseID from another list. This is not straightforward, as the CauseIDs have changed significantly from list to list, and occasionally are split or merged as the definitions of diseases change. Even within lists, there are often inconsistencies regarding how a single disease is represented. If a person wants to find out trends regarding a single disease, they must group together all the CauseID's pertaining to that disease.

Food

The United Nations makes its Food Balance Sheets available to the general public through the Food and Agriculture Organization (FAO). The FAO recently revamped their website to form an online database system. While the data is licensed for non-profit use, the amount of data freely available to the public is limited by data-row, and custom queries are not allowed on the website. We used a combination of downloading and data processing to obtain a subset of this data.

The food data we obtained is not as extensive as the mortality data. There are only about 124 different food groups tracked, labeled by country and year.

Data Restructuring

Query Oriented Database Tables

The Query Oriented Database had to be restructured so that a common backbone could be established to connect the food and mortality data. We developed a common table of Countries, standardizing spellings and using a RegionID code as a universal identifier. This RegionID serves to connect the food and mortality tables, which are also connected by year.

Statistics Oriented Database Tables

For the purpose of performing linear regression, the food groups and disease causes need to be connected directly on years as well as regions. Pairing these fields allows linear regression to proceed directly for food-disease pairs without further processing. The data tables from the Query Oriented Database tables were reorganized into a set of Statistics Oriented Database tables to allow for simpler processing of statistics information.

Because the Statistics Oriented Database tables are built off of the same core data, they are kept consistent. The information in each set of tables is the same, the only difference is the view from which the data is presented.

Figure 1 shows the table schemas for both the query and statistics oriented tables. The tables "FinalFood" and "FinalMort" are the Query Oriented Database tables, while the tables "FoodExportOutput" and "MortExportOutput" are the Statistics Oriented Database tables.

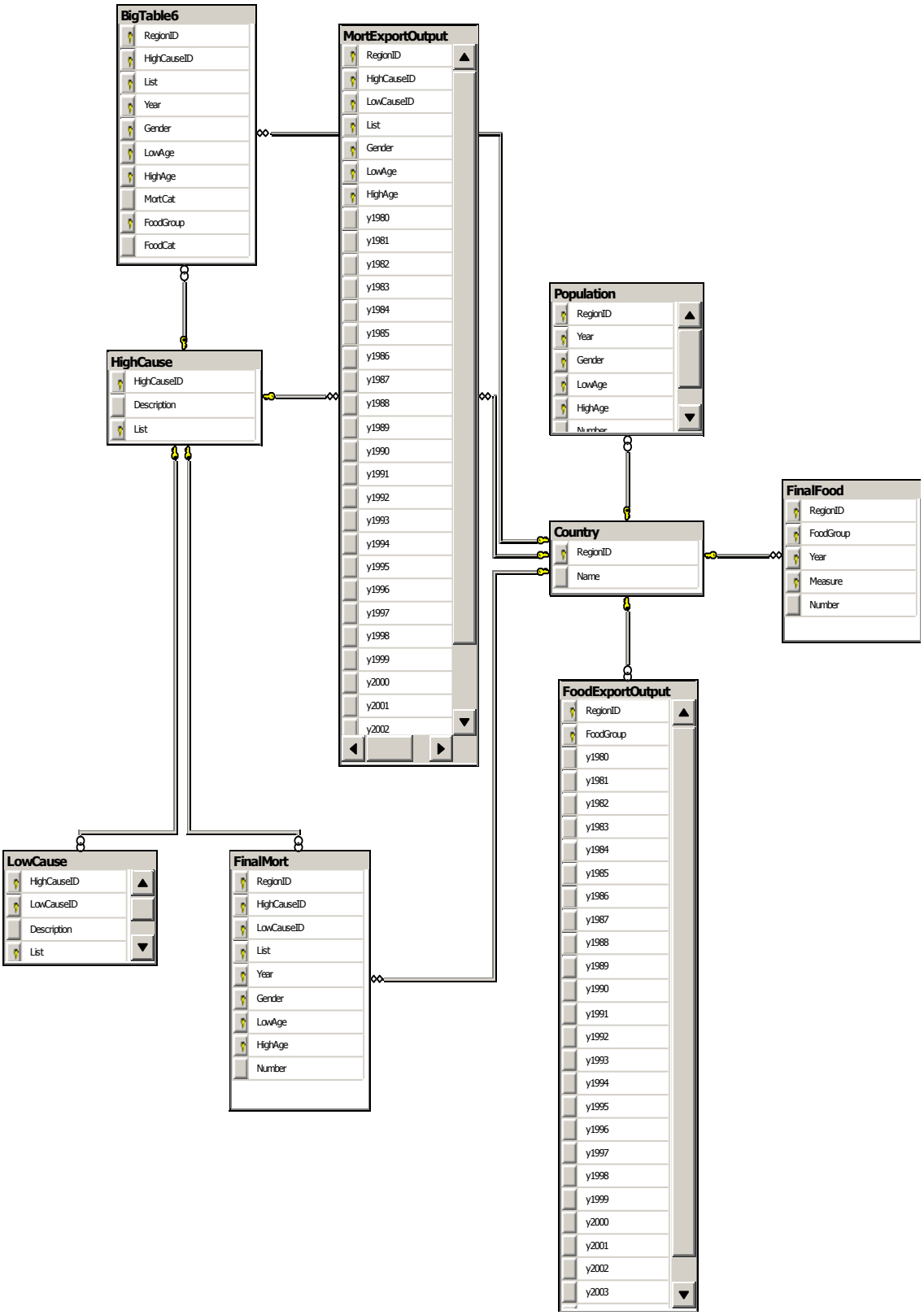

Figure 1: Database Schemas

Data Cleaning

The data contained several anomalies which had to be cleaned out before proceeding. Many of these anomalies were due to issues such as bad reporting or political problems in the countries reporting, leading to inconsistencies in labeling of diseases from year to year or lack of complete data. We suspect that other errors were introduced through data formatting. However, the vast majority of data conformed to our database format. In each case, we took the smallest subset of data that we deemed consistent. Even given the data removed for inconsistencies, large amounts of data still remained.

Query Results

Both the Query Oriented Database Tables and the Statistics Oriented Database tables yielded interesting results. The main exploration using the Query Oriented Database tables was to graph the health outcomes of several Asian countries compared to food consumption changes. Several graphs are shown below, giving an idea of what kinds of trends can be discovered through the Query Oriented Database tables.

Legend

Symbol	Country
T	Thailand
B	Brazil
K	Korea (South)
J	Japan

Mortality Plots

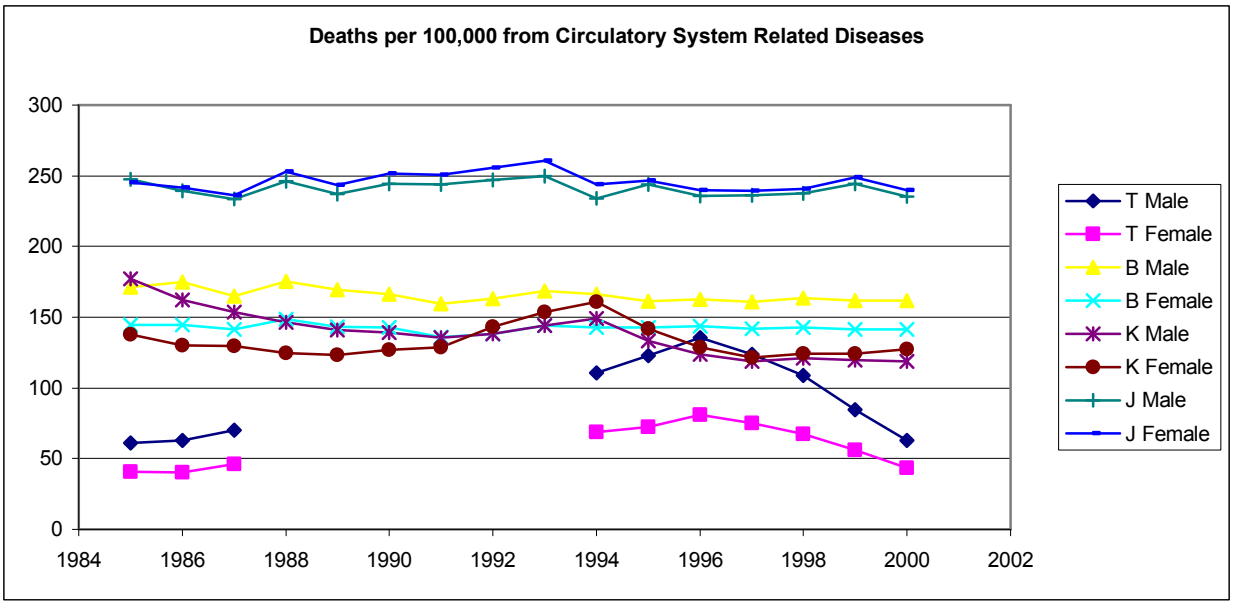

Figure 2: Deaths per 100,000 from Circulatory System Related Diseases

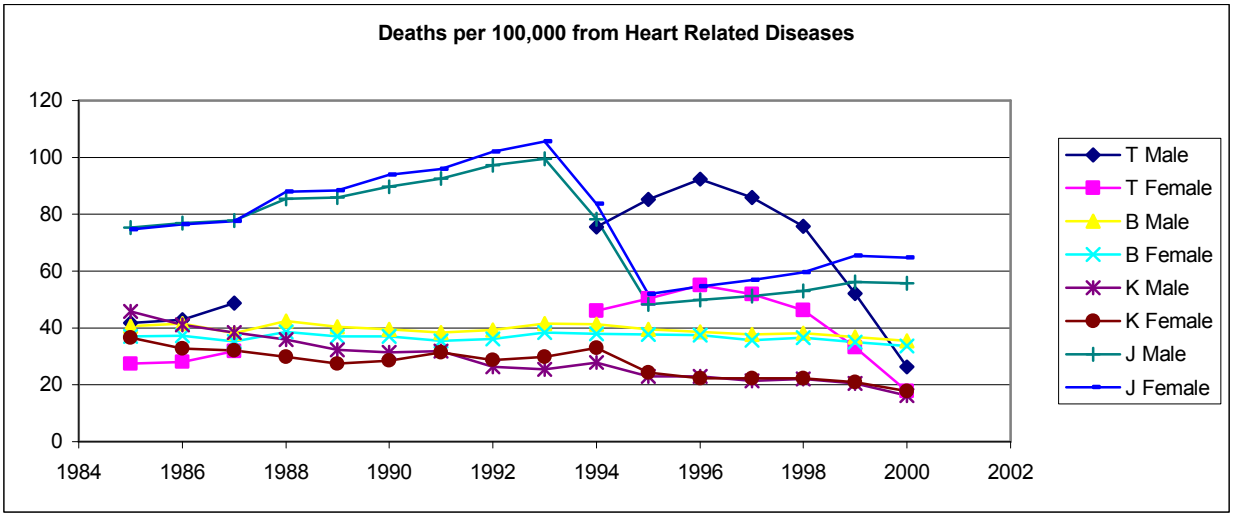

Figure 3: Deaths per 100,000 from Heart Disease

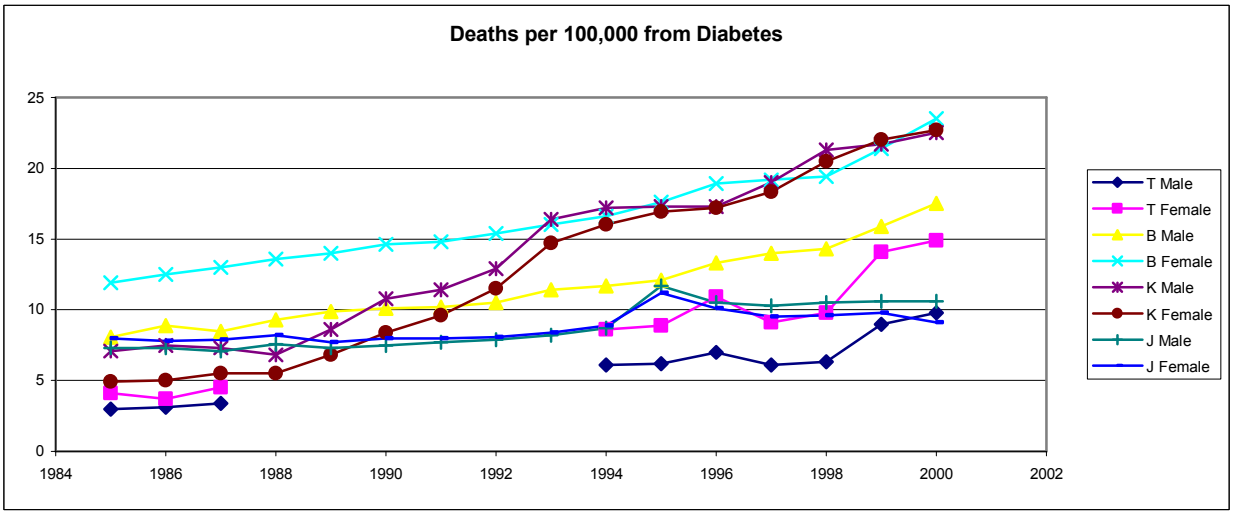

Figure 4: Deaths per 100,000 from Diabetes

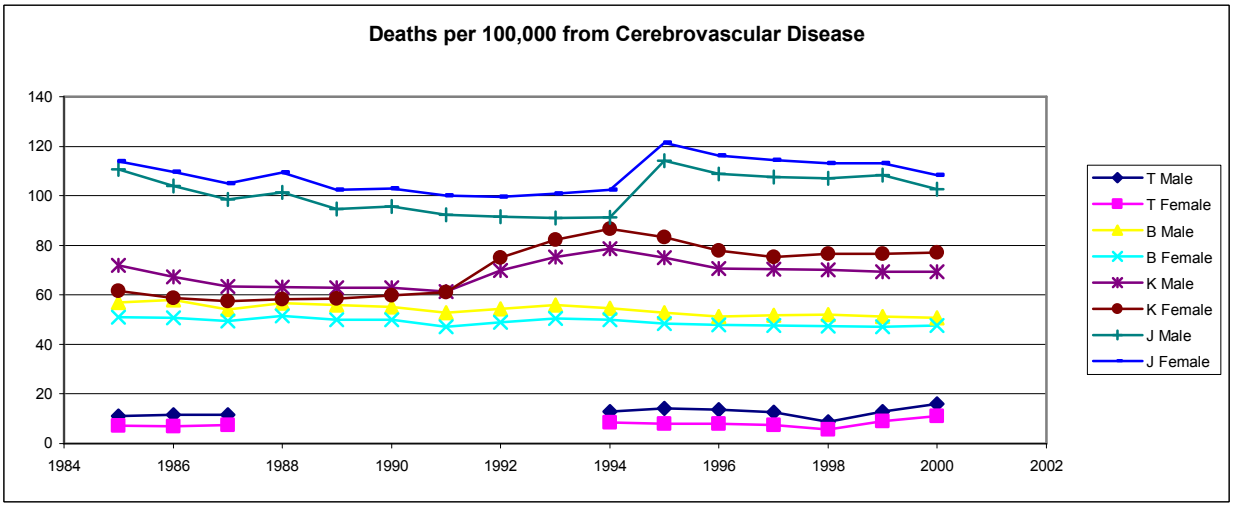

Figure 5: Deaths per 100,000 from Cerebrovascular Disease

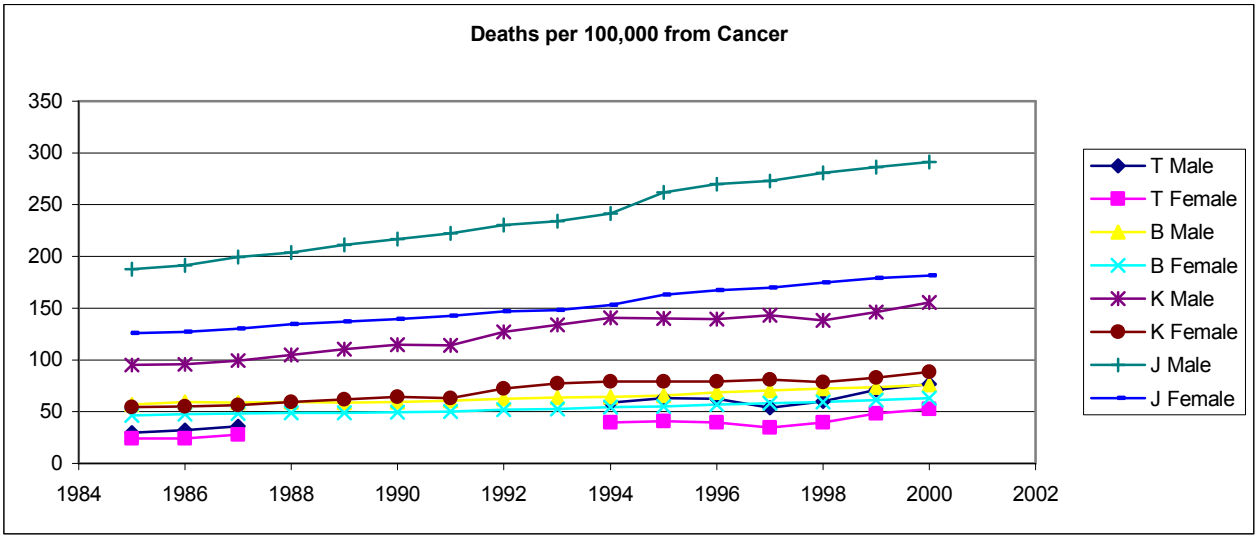

Figure 6: Deaths per 100,000 from Cancer

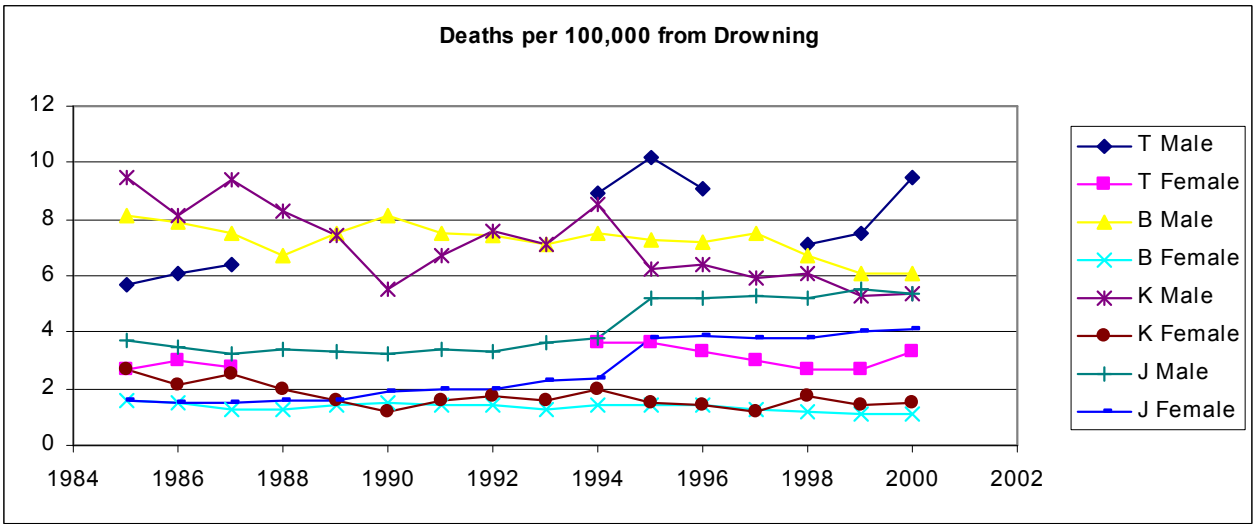

Figure 7: Deaths per 100,000 from Drowning

Food Plots

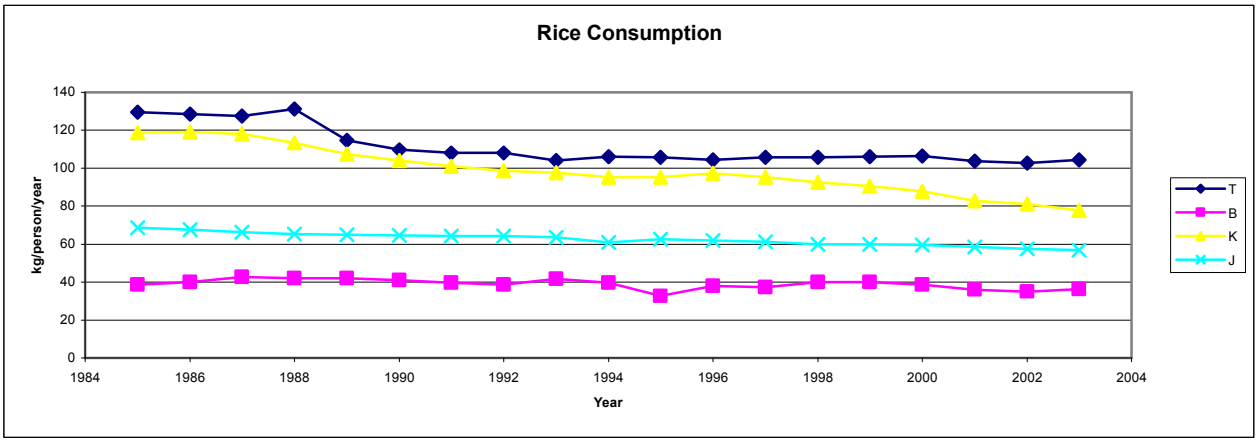

Figure 8: Rice Consumption (kg available / person /year)

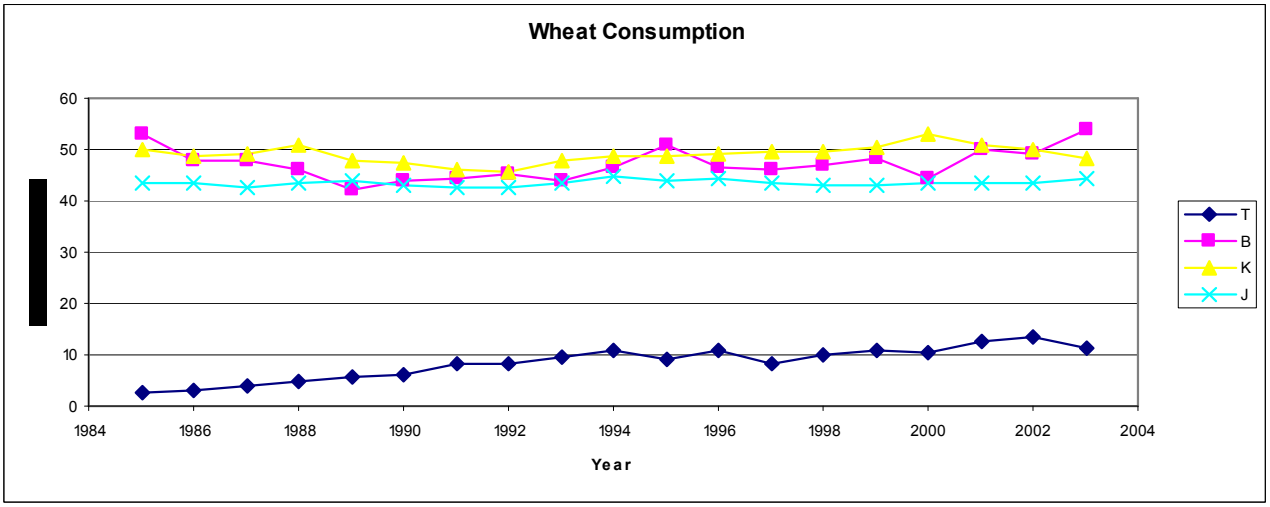

Figure 9: Wheat Consumption (kg available / person /year)

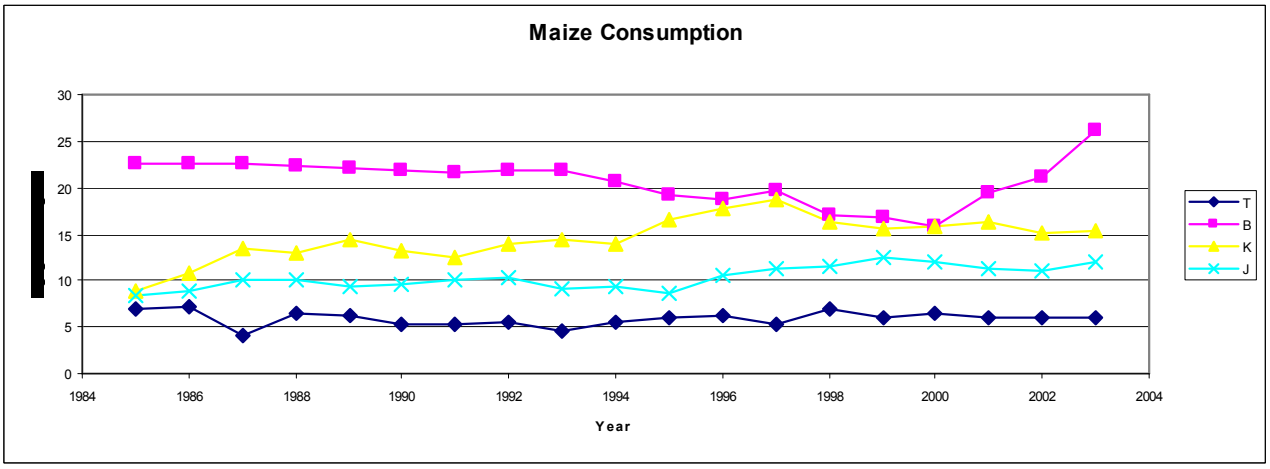

Figure 10: Maize Consumption (kg available / person /year)

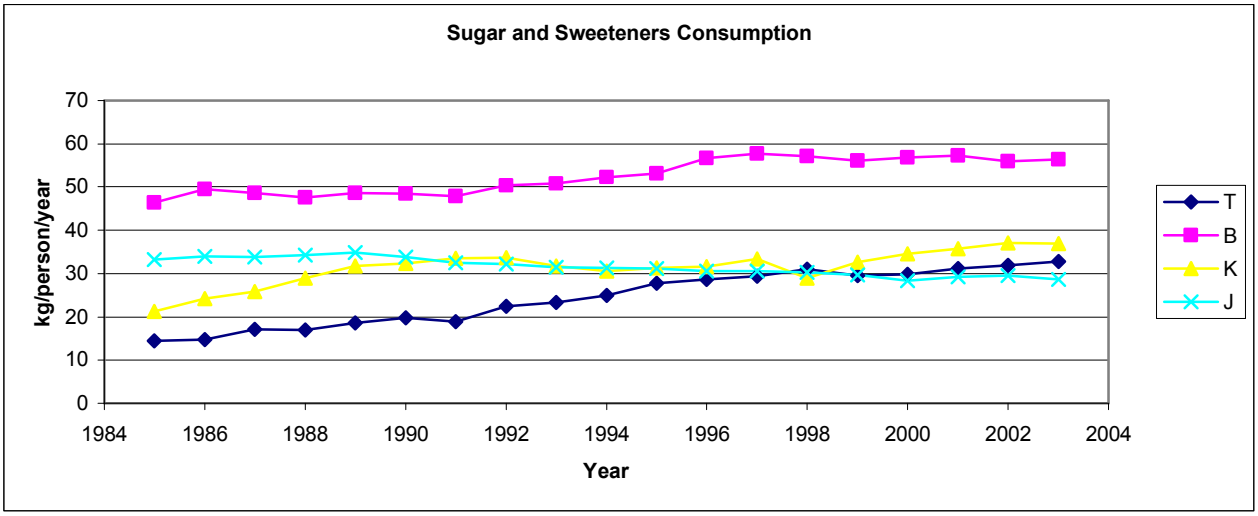

Figure 11: Sugar and Sweeteners Consumption (kg available / person /year)

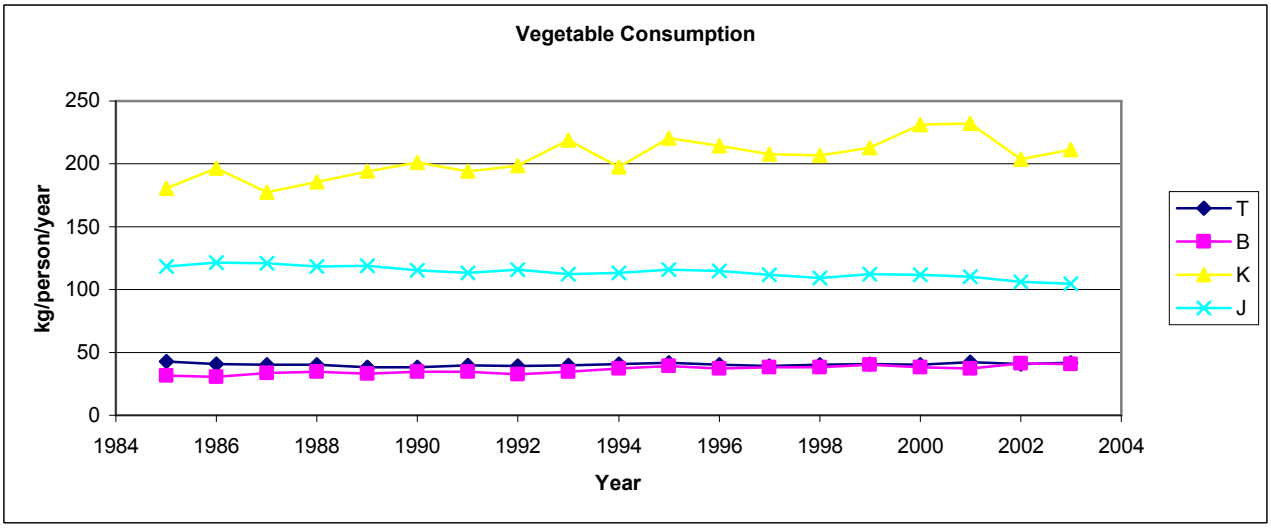

Figure 12: Vegetable Consumption (kg available / person /year)

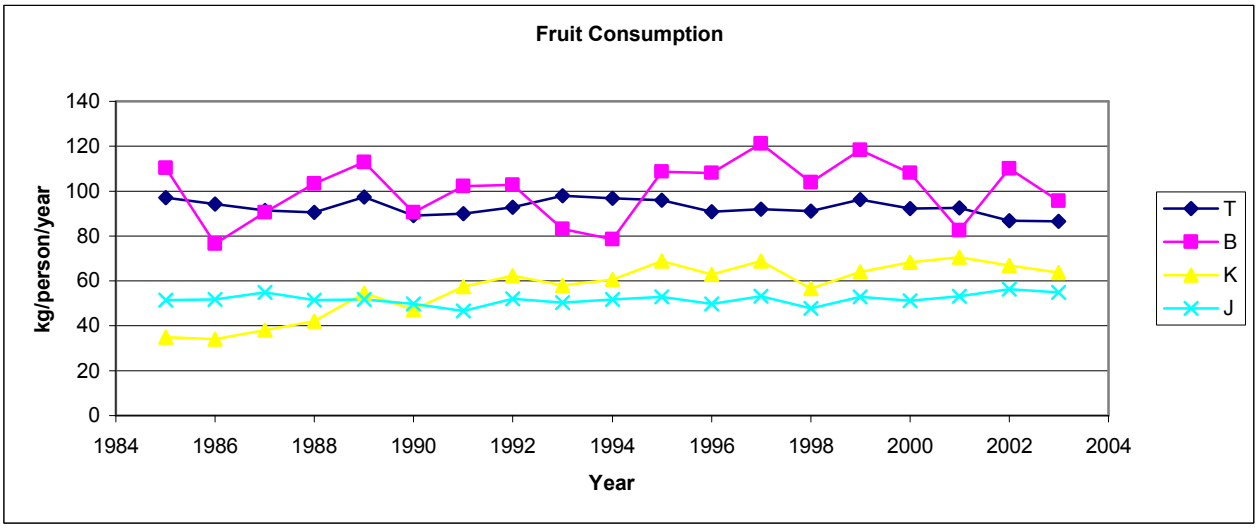

Figure 13: Fruit Consumption (kg available / person /year)

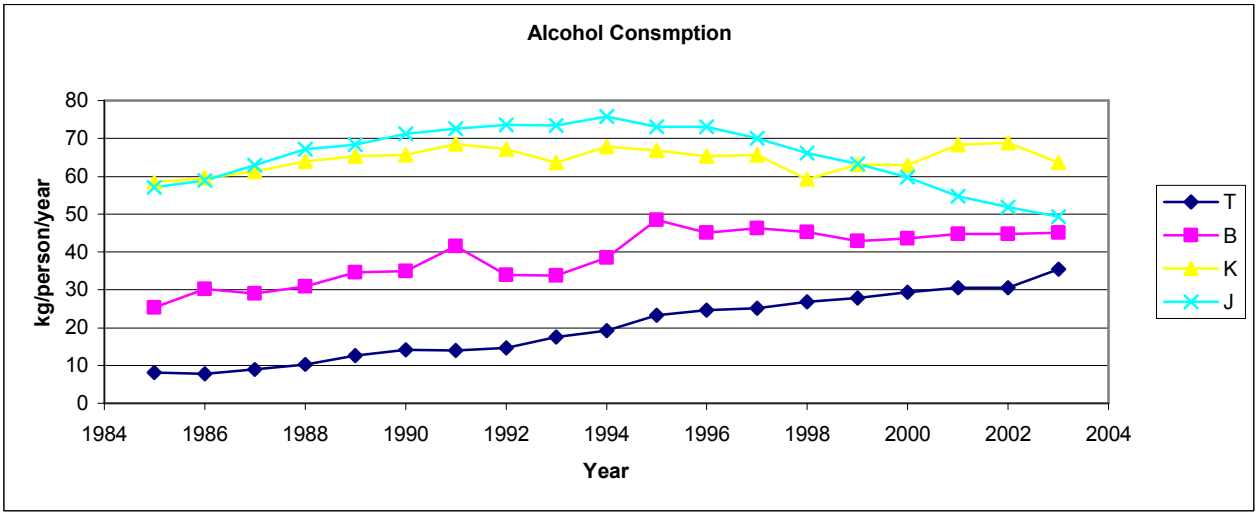

Figure 14: Alcohol Consumption (kg available / person /year)

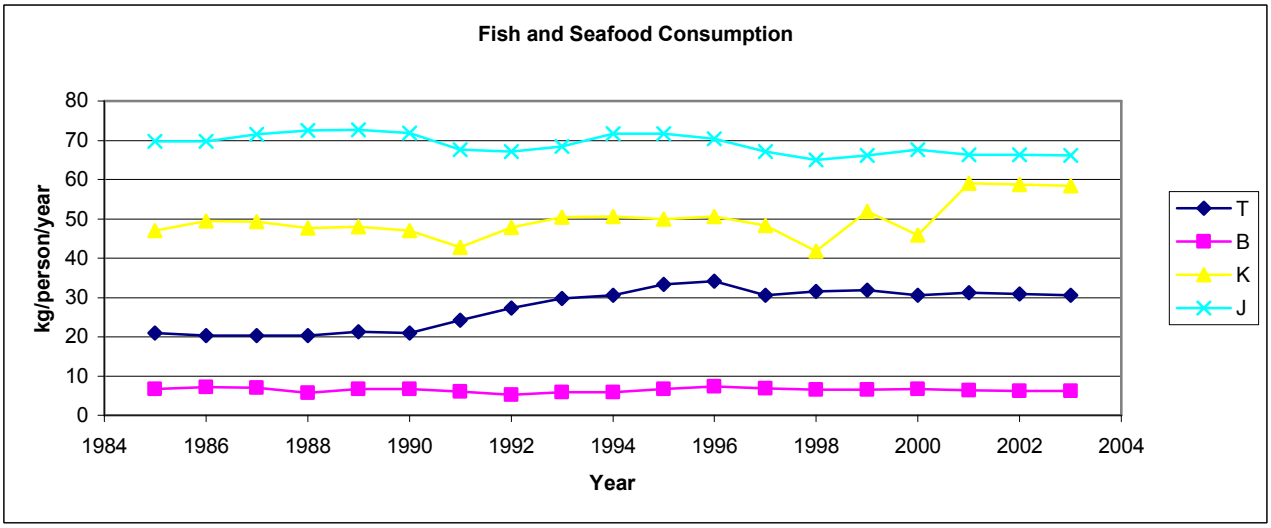

Figure 15: Fish and Seafood Consumption (kg available / person /year)

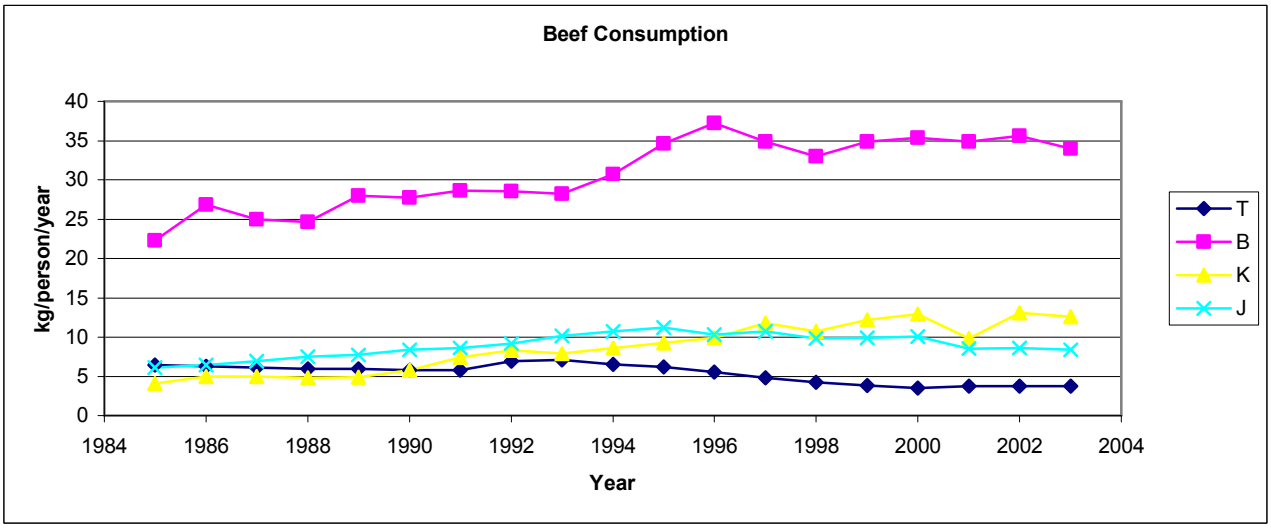

Figure 16: Beef Consumption (kg available / person /year)

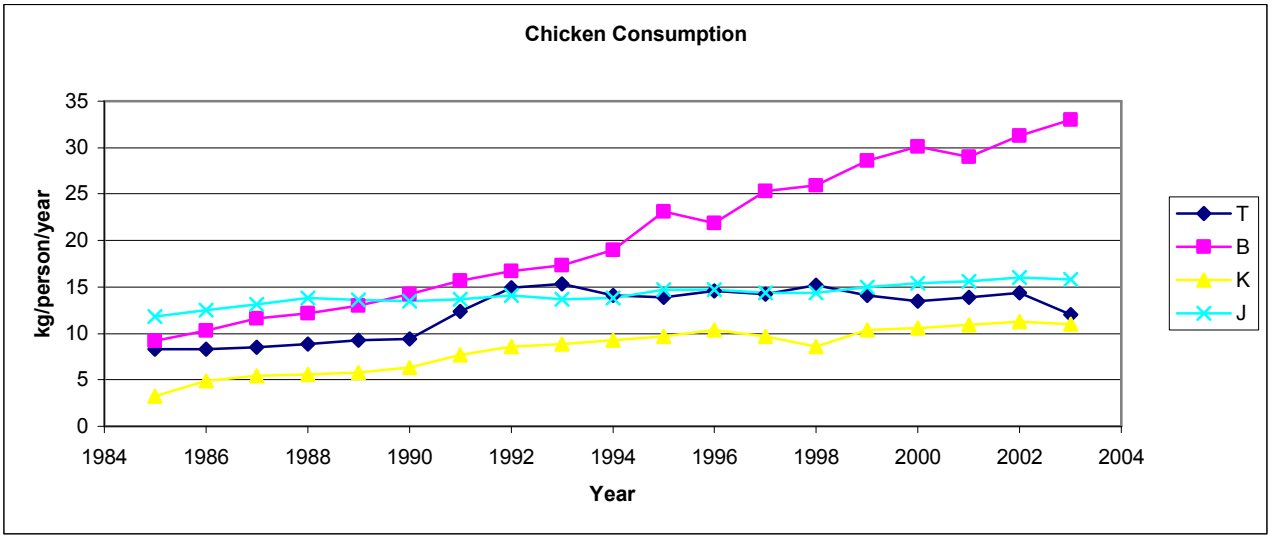

Figure 17: Chicken Consumption (kg available / person /year)

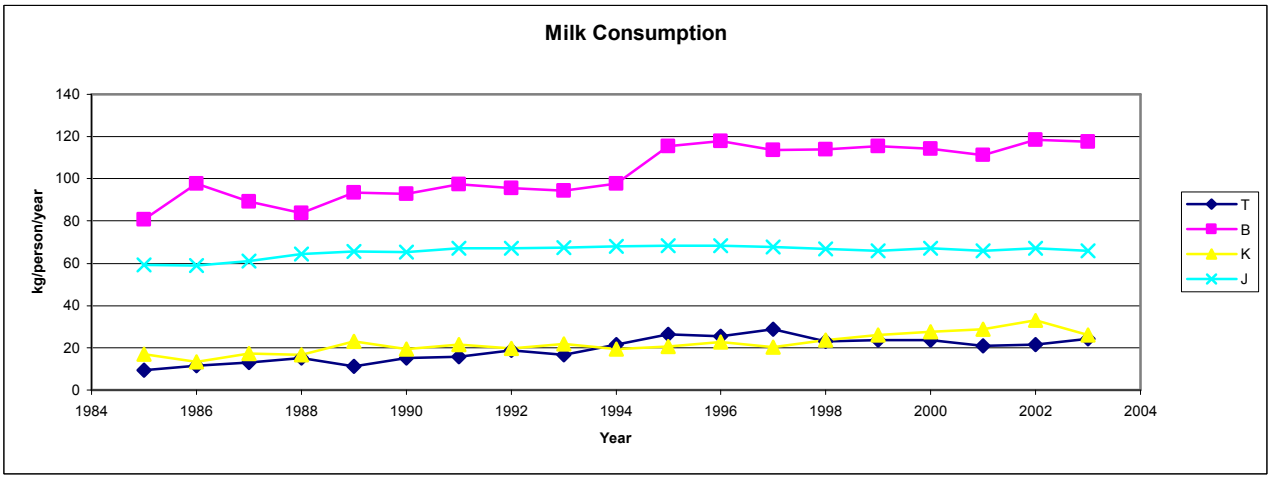

Figure 18: Milk Consumption (kg available / person /year)

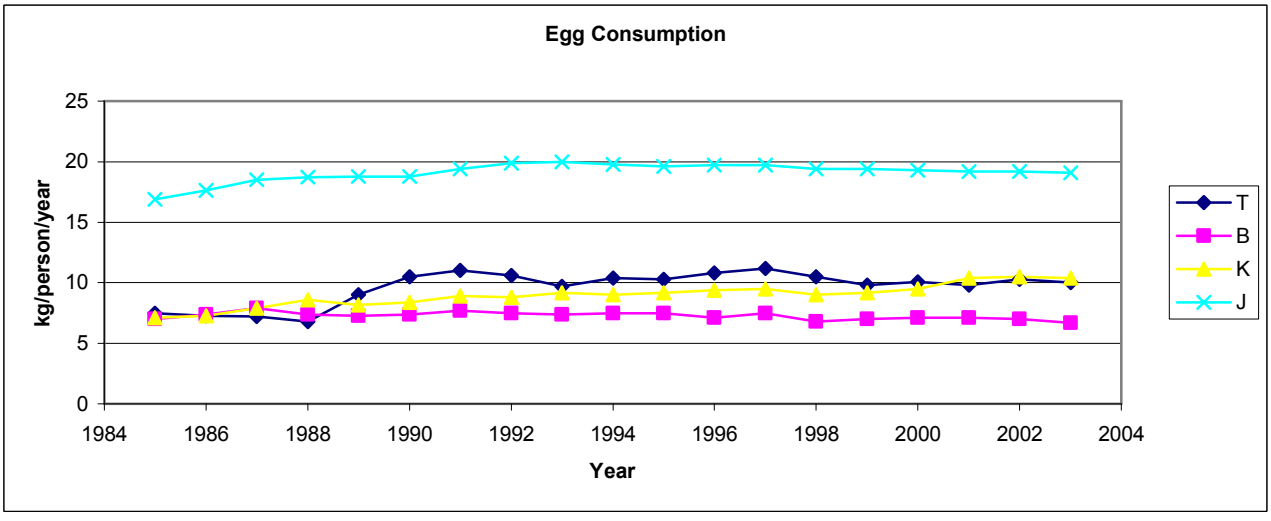

Figure 19: Egg Consumption (kg available / person /year)

Query Discussion

Several interesting results emerge from the charts of the query oriented database tables:

- Food consumption has mostly increased for countries sampled, with the exception of rice.
- Some foods and diseases have experienced drastic changes in their numbers over the period of time viewed:
    - Death per 100,000 from diabetes saw enormous increases in almost all countries sampled.
    - Deaths per 100,000 from heart disease in Japan and Thailand saw drastic declines during the 90s. This could be the result of improved health care, changes in policy, or removal of certain foods from the market. It is unlikely that this kind of change is caused by long term changes of diet because of its sudden nature.
- Several measures increase together for the entire period, suggesting that linear regression for only one country or even a small set of countries would not lead to specific connections that could be generalized to other countries, but that it could be an artifact of conditions in that specific country.
- The use of controls, such as deaths per capital from drowning, show that the data is stable and that trends that emerge are not random.
- The Query Oriented Database tables are suited for answering specific questions about food and disease trends.

Statistics Results

The Statistics Oriented Database tables yielded more general results regarding connections between specific foods and specific diseases. The linear regression was first performed on a single country, the United States. The linear regression algorithm was then extended to include all countries that had available data for a given food-disease-year. The figures below summarize the important results. In the figures below, Sector A refers to a specific subset of food-disease pairs that is used for comparing the US plot to the world plot. For the color-maps, diseases are lined up along the horizontal axis and foods are along the vertical axis.

United States Results

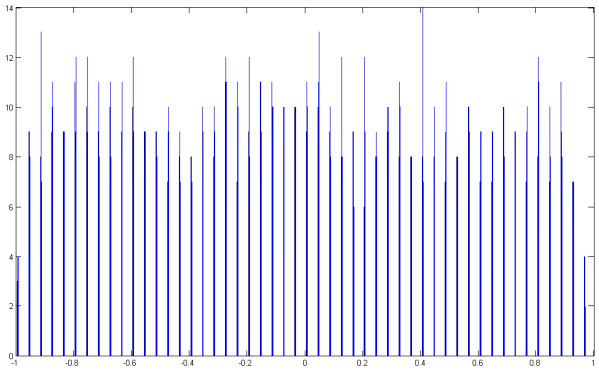

Figure 20: Histogram of Correlation Coefficients for food-disease pairs (US only)

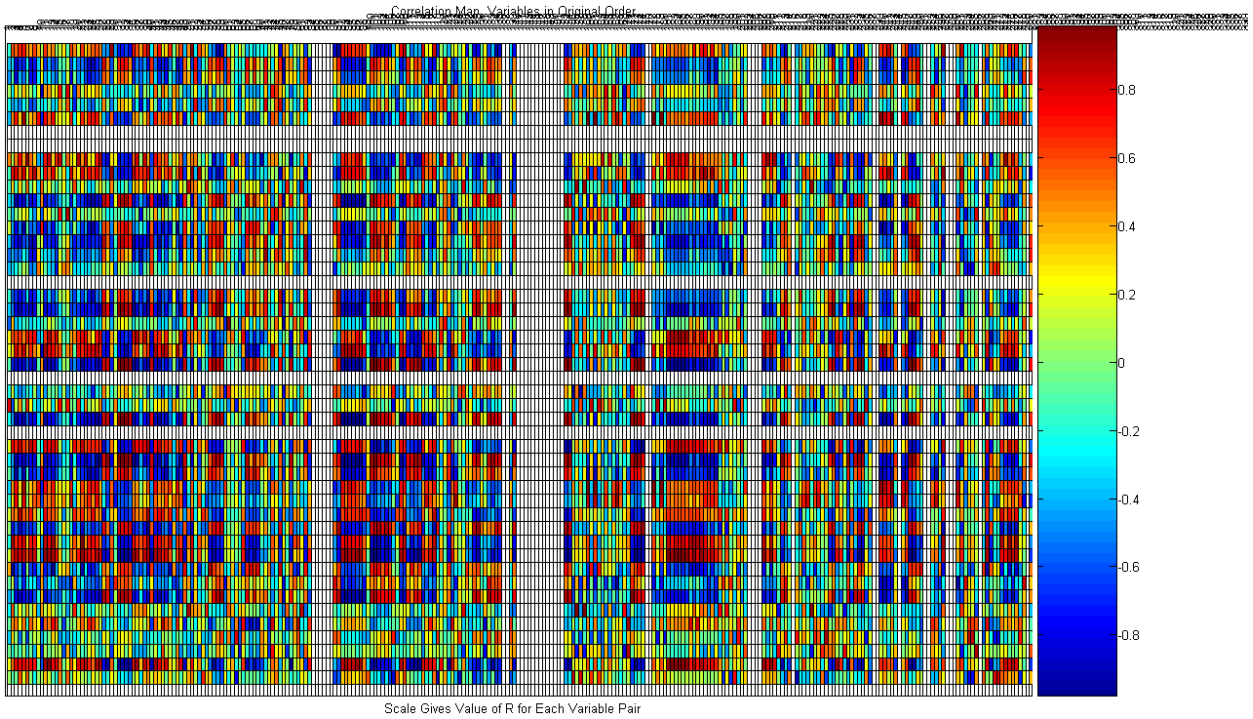

Figure 21: False colored correlation map showing strength of linear regression correlation coefficient for various food-disease pairs (US only)

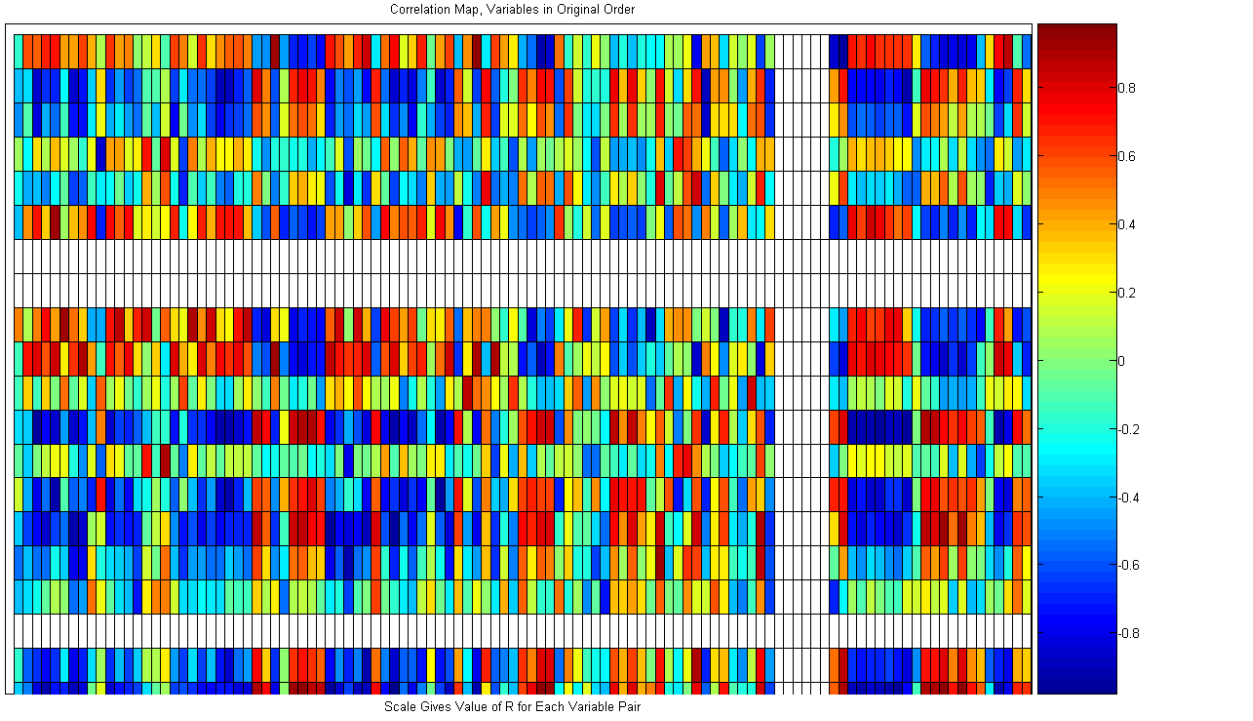

Figure 22: False colored correlation map showing strength of linear regression correlation coefficient for various food-disease pairs (US only) for sector A

World

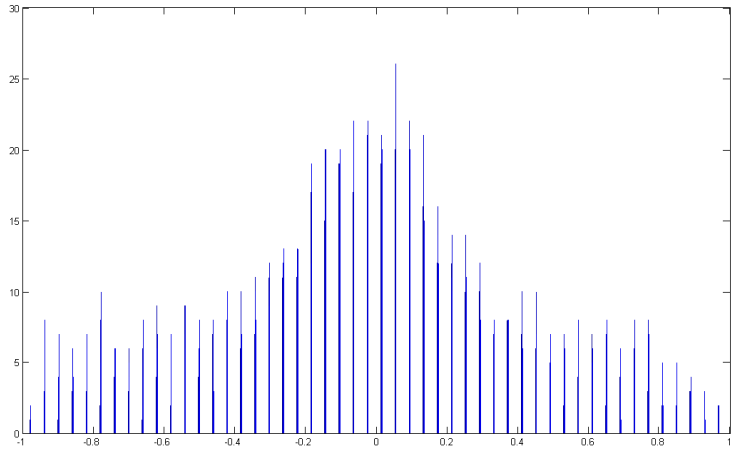

Figure 23: Histogram of Correlation Coefficients for food-disease pairs

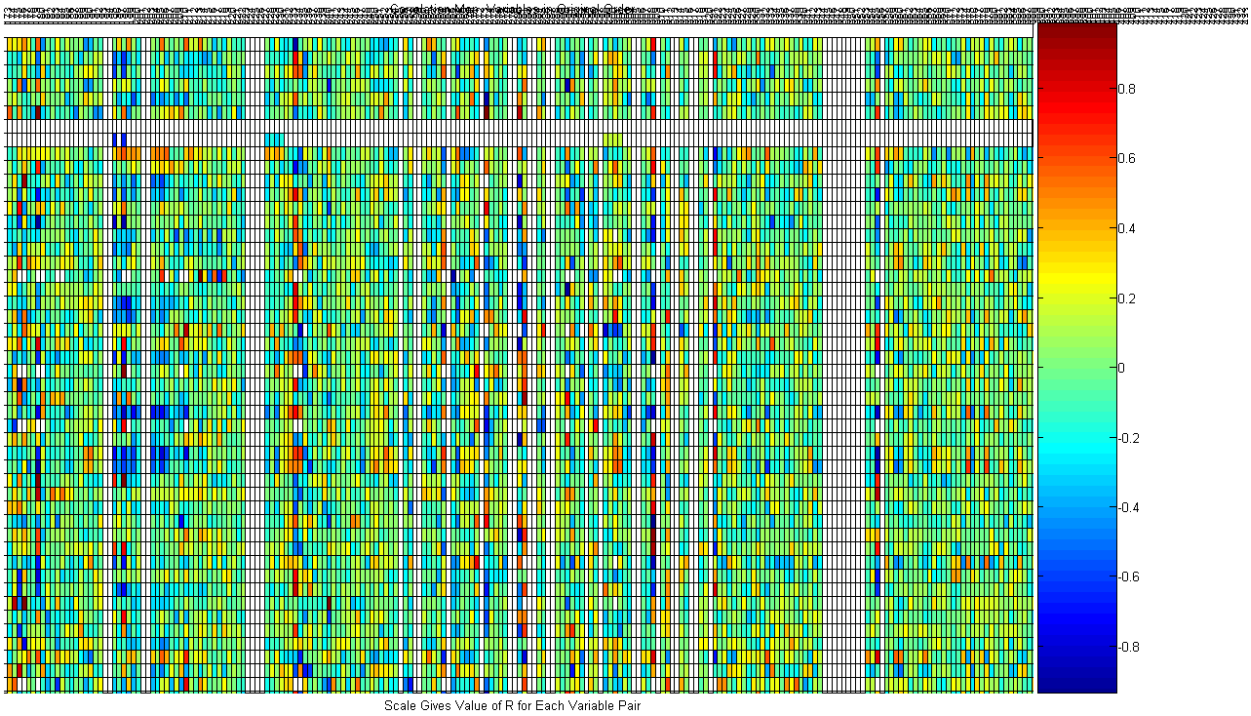

Figure 24: False colored correlation map showing strength of linear regression correlation coefficient for various food-disease pairs

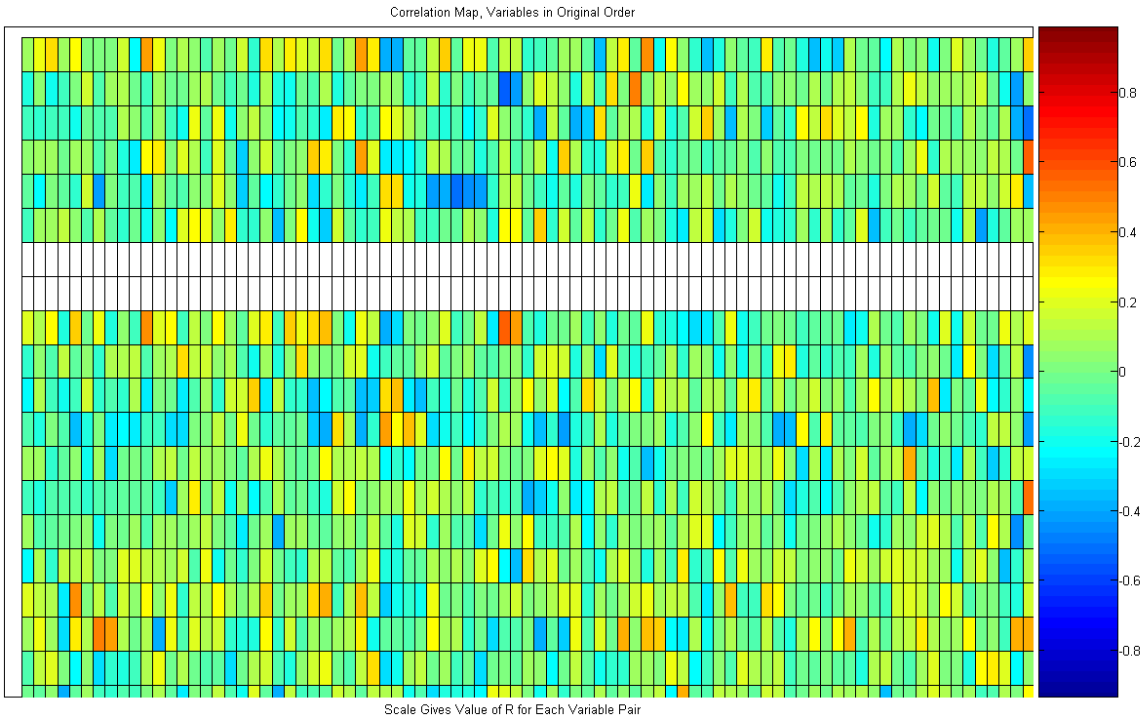

Figure 25: False colored correlation map showing strength of linear regression correlation coefficient for various food-disease pairs for sector A

Statistics Discussion

There are several notable results from the Statistics Oriented Database tables:

- The US only color-map figures would suggest several very strong correlations, however when combined with data from other countries, it is evident that those correlations are likely coincidental or applicable only to the United States.
- The histogram figures also suggest that the world plots tend to be centered around 0, whereas the US plots are more random. This suggests that most of the correlations that were derived locally could not be generalized to data from other countries.
- While many of the food-disease pairs are insignificant in the world plot, several are very significant and could yield interesting results in further studies.
- There are fewer missing food-disease correlations in the global plot, because some countries reported data that others did not.
- There are several clusters of high correlation, but they tend to be along the vertical (disease) axis. This suggests that there are certain diseases that are extremely susceptible to changes in food supply, but that food consumption is generally not as susceptible to changes in disease trends. Table 2 shows some of the diseases that were found to be especially sensitive to changes in food consumption.

Table 2: Diseases Found to be Highly Sensitive to Changes in Food Consumption

Unique Identifier	High Level Descriptor	Detailed Descriptor
197—09BTL	Secondary malignant neoplasm (cancer) of	None

	respiratory and digestive system	
198—09BTL	Secondary malignant neoplasm of other specified sites	None
199—09BTL	Malignant neoplasm without specification of site	None
B01-4-09BTL	Intestinal infectious diseases	Amoebiasis
B03-0-09BTL	Other bacterial diseases	Plague
B05-0-10M	Measles	Measles complicated by encephalitis
B07-4-09BTL	Other infectious and parasitic diseases and late effects of infectious and parasitic diseases	Filarial infection and dracontiasis
B08-2-10M	Viral infections characterized by skin and mucous membrane lesions, not elsewhere classified	Exanthema subitum [sixth disease]
B19-0-09BTL	Nutritional deficiencies	Kwashiorkor

Several of the diseases in Table 2 are suspected not to be related to food consumption, however may have experienced severe declines in the period in question, and therefore exhibit strong (negative) correlation with other foods that have similarly high increases in consumption. For example, the plague and measles are infectious diseases, with little known connection to food consumption. On the other hand, there are many food and nutrition related diseases in the list, such as cancer of the digestive system and various nutritional deficiencies. It is expected that these would be strongly correlated with food consumption. These categories thus serve as a validation of the model.

In order to find less apparent connections between foods and disease, a more detailed analysis of the food-disease pairs would have to be conducted. The diseases listed in Table 2 were highly correlated with all foods, suggesting a broader relationship, and showed up prominently in the correlation color-map. While more specific relationships would be harder to find in a color-map, the construction of a different data structure would be beneficial to the identification and validation of these connections. There are also hidden connections between groups of foods and certain diseases, such as the composition of a known diet and its connection to certain health outcomes. These known connections could potentially serve as a validation point for the model and should be explored further.

Most Significant Findings

- There are meaningful connections between food and disease that are consistent across global data. Certain diseases are found to be very highly susceptible to changes in food supply.

- In order to find more specific connections between foods or food combinations and diseases, additional work would have to go into developing a more advanced set of algorithms to identify and validate these connections.
- Local food-disease correlations tend to be stronger than global food-disease correlations. This could be because of specific conditions in local areas, different preparations of foods, coincidence, food availability, health care, or a variety of other factors. Using global food-disease correlations reduces the bias from regional factors.
- Better and more consistent tracking of food and disease data across the world would greatly improve the resolution and validity of our model, and would increase the ease with which computer learning models are constructed to analyze this data.

Possible Next Steps

Several additional efforts could be made to improve the validity, depth, and consistency of the results presented, and could yield further relevant results. These further steps include:

- Development of more advanced computer learning models to automatically perform data-mining.
- Utilization of larger spans of data, incorporating a larger time series.
- Using multiple, unevenly weighted keys, for example a known diet composition, as a predictor of disease trends.
- Weighting of the data based on reliability.
- Incorporation of disease incidence instead of mortality.
- Analyzing the effects of policy on food and disease trends.
- Validation of model with known examples.
- Analysis of specific, strong food-disease pairs.

References

Albala, Cecilia et al. "Nutritional Transition in Latin America, the Case of Chile". Nutrition Reviews. Health Module. June 2001; Vol. 59, Issue 6, pg. 170.

Barbagallo, Carlo M. "Nutritional Characteristics of a Rural Southern Italy Population: The Ventimiglia di Sicilia Project." Journal of the American College of Nutrition. 2002. Vol. 21, No. 5, pg. 523-529

Diamond, Jared. "Guns, Germs, and Steel: The Fates of Human Societies" New York: W.W. Norton & Company, 1997.

FAOSTAT. Food and Agriculture Organization of the United Nations. Core Consumption Data. http://faostat.fao.org/site/345/default.aspx,
Copyright information: http://www.fao.org/copyright_en.htm

Helsing, Elisabet. "Traditional Diets and Disease Patterns of the Mediterranean, circa 1960." The American Journal of Clinical Nutrition. Bethesda: June 1995. Vol. 61, Issue 6, pg. 1329S.

Simopoulos, Artemis P. "The Mediterranean Diets: What Is So Special about the Diet of Greece? The Scientific Evidence." American Institute for Cancer 11th Annual Research Conference on Diet, Nutrition, and Cancer. American Society for Nutritional Sciences. 2001.

United Nations Mortality Data. World Health Organization. http://www3.who.int/whosis/menu.cfm?path=whosis,mort&language=english, Copyright information: http://www.who.int/about/copyright/en/