AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

NOWCASTING OBESITY IN THE U.S. USING GOOGLE SEARCH VOLUME DATA

Sercan Sarigul, University of Rochester, sercan.sarigul@simon.rochester.edu

Huaxia Rui, University of Rochester, huaxia.rui@simon.rochester.edu

Selected paper prepared for presentation at the Agricultural & Applied Economics Associations'

2014 AAEA/EAAE/CAES Joint Symposium, Montreal, QC, May 28-30, 2014.

Copyright 2014 by [Sercan Sarigul, Huaxia Rui]. All rights reserved. Readers may make verbatim copies of this document for non-commercial purposes by any means, provided that this copyright notice appears on all such copies.

NOWCASTING OBESITY IN THE U.S. USING GOOGLE SEARCH VOLUME DATA

Sercan Sarigul, Huaxia Rui

ABSTRACT

"Googling" is now ubiquitous in our society. We typically start searching on Google before we make purchase decisions or when we have interest in certain topics. Aggregating these search data can provide us with real-time and possibly accurate information on people's behavior. In fact, Google keeps tracks of all the search queries and has accumulated a tremendous amount of information about people's interest at the society level. It currently provides search volume data of keywords for different regions and time intervals on its free and public service of Google Trends. An interesting and hot research area is how to exploit the Google Search volume data in innovative ways to benefit our society. This paper aims to reveal the connection between obesity prevalence and people's online search behavior in the United States by combining data from Google Trends and data from Behavioral Risk Factor Surveillance System (BRFSS) which is published by the Centers for Disease Control and Prevention (CDC) annually. We first hand-selected keywords that are associated to people's life style and used panel data model to study association between search pattern and obesity level. We found significant correlation power of those keywords with Body Mass Index (BMI) level and results suggest great promise of the idea of obesity monitoring through real-time Google Trends data. We believe this is an important finding and is particularly attractive for government health institutions and private businesses such as insurance companies etc.

I. INTRODUCTION

Google Trends is a public web facility of Google Inc., based on Google Search, that shows how frequent a particular search-term is searched relative to the search volume of all keywords.[1] It provides a line chart which has normalized and relative search volume frequency on vertical axis and time line on horizontal axis.* There are 210 metro areas, defined by Google, in the U.S., however Google Trends allows to compare at most 5 metro areas at a time. We used some conversion method to make indices comparable as details provided in Section-IV. Google has not still released official API for its Google Trends project, however data can be downloaded in the .csv format manually.

We focus on the United States for two main reasons. First, the vast majority of population in the U.S. has access to internet and use of Google Search feature is a daily routine. As a result, Google can provide regional level information of search volumes for the U.S., while it does not –and most probably because it cannot- for most of the other countries due to low search volumes. Secondly, obesity is one of the most prominent diseases in the U.S.[2]

According to the World Health Organization (WHO), the U.S. government spent $8608 on health care per capita in 2011 and it accounts 17.9% of the U.S. GDP for that year.[3] Moreover, average U.S. citizen is expected to pay approximately $4000 annually for the health insurance under the new Affordable Care Act.[4] The U.S. is usually listed in the lowest ranks in terms of health care quality among similar developed countries, albeit high cost figures. It is an indicator of inefficiencies in health care operations and each year government is investing millions of dollars to health care research on different areas like developing new treatments or medicines, educating health care providers, publishing health care data and so forth to mitigate these inefficiencies.

* Although default settings of Google Trends turns worldwide search volume indices from 2004 through the date of inquiry, one can easily change it to a particular country or particular metro area. Similarly, time interval can be adjusted based on user preference. (Appendix-1 and Appendix-2) Google Trends provides search volume indices which are scaled between 0 and 100, so index value depends also on other keywords' search volumes. In that sense, having two different Google Trends datasets for two different metro areas with same index value for same year and same keyword does not mean that the keyword was searched same number of times in both metro areas. Nevertheless, it means that relative search volumes within the metro areas are same. This relativity would cause a big problem of comparability, if Google did not provide online tool to compare indices of multiple metro areas. (Appendix-3)

According to the OECD Health Data 2012, obesity rates have increased substantially all over the world during past 20 years and are highest in the U.S. compared to other OECD nations. (Appendix-4) The work of International Association for the Study of Obesity shows that over one-third of children in the U.S. are overweight or obese. (Appendix-5) It is an upsetting sign that obesity will continue to be one of the biggest health issues in the U.S. for the next few decades. Research on that field could help to increase life quality of people and save billions of dollars for government. Finkelstein[5] et al. (2009) estimated that medical care costs of obesity in the U.S. totaled about $147 billion in constant 2008 dollars.

We used data from Behavioral Risk Factor Surveillance System[6] (BRFSS) which is released by Centers for Disease Control and Prevention (CDC) to see whether Google Trends data is useful to monitor obesity prevalence in different metro areas.

Our analysis indicated that search pattern of keywords of “mcdonalds”, “farmers market”, “games”, “movies”, “weather”, “pain” and “weight loss” have significant power to monitor BMI average of a metro area.

The rest of the paper is organized as follows: We reviewed relevant literature from different perspectives in Section-II. In Section-III, hypothesis was developed and in Section-IV, data collection process is illustrated. Section-V summarizes analyses and Section-VI concludes the paper.

II. LITERATURE REVIEW

There are plenty of papers that use Google Trends data to predict some real world phenomena. The idea of investigating relations between prevalence of some illness and search volume indices of some related keywords is also not new. Below we review how Google Trends data has been used by many scholars in various academic fields. We also review studies on economic implications of obesity, epidemiologic nature of obesity, monitoring obesity and geographic distribution of obesity.

a) Google Trends

In this section, the papers which use Google Search volume index to explain some real world phenomena, other than health related ones, are discussed. Researchers have been using Google Trends to nowcast and/or forecast unemployment rates, private consumption, consumer sentiment, inflation, box-office revenues, election results etc.

Probably the first published paper which suggested the idea of using web search data to nowcast some real world phenomena is Ettredge[7] et al. (2005), which developed a regression model to nowcast unemployment rate. Besides, Choi and Varian[8] (2012) can be considered as first seminal paper in the field as Google Trends data was successfully used to forecast near-term values of economic indicators such as automobile sales, unemployment claims and consumer confidence index.

For many people, it is quite attractable to reach statistics of economic indicators before government published them. In that sense, many papers focused on the relationship between economic indicators and search volume indices. D'Amuri and Marcucci[9] (2009) showed that U.S. unemployment rate can be predicted based on Google Trends and also showed that Google Trends data increases predictive power of traditional forecasting models substantially. Predicting unemployment rate using Google Trends took lots of attraction, as it is an important public affair and people want to know about it regularly. Askitas and Zimmermann[10] (2009), D'Amuri[11] (2009) and Anvik and Gjelstad[12] (2010) are similar works to predict unemployment rates in Germany, Italy and Norway, respectively.

There are many other papers on, again, links between Google Trends and economic indicators, other than unemployment rate. Preis[13] et al. (2010) and Da[14] et al. (2011) showed that Google Trends data can be used to capture investor attentions to the sample of Russell 3000 stocks and to monitor weekly transaction volumes of S&P 500 companies, respectively. Vosen and Schmidt[15] (2011) compared performance of Google Trends in private consumption forecast with performances of University of Michigan Consumer Sentiment Index and Consumer Board Consumer Confidence Index and stated that indicators developed using search volume data outperforms the others. Notable sample from many others can be listed as Huang and Penna[16] (2009), which used search volume data to predict consumer sentiments, Wu and Brynjolfsson[17] (2010), that suggested prediction of house sales volume

and prices using search volumes, Goel[18] et al. (2010), which focused on predicting box office revenue of films before they released, Guzman[19] (2011) that developed a model based on Google Trends to predict inflation and Lindberg[20] (2011), which nowcasted volume of retail sales in Sweden.

Some other papers investigated the relationship between search queries and election results. The findings of notable papers, however, showed that search volume data is not a good predictor for election results. Lui[21] et al. (2011) concluded that Google Trends was not actually good predictor of both 2008 and 2010 elections of the U.S. Congress and Metaxas[22] et al. (2011) verified this conclusion. On the other hand, Reilly[23] et al. (2012) showed that Google searches for ballot measures' names one week before the 2008 Presidential election correlate with actual participation on those ballot measures.

b) Google Trends and Health

In this section, we review papers which connects Google Search volume index with some health-related real world phenomena.

Epidemic diseases, especially influenza like illnesses, captured most of the attention of researchers. Influenza is very common disease that most of the population experience frequently, so lots of data is available. Secondly, predicting possible outbreaks of an epidemic disease -or nowcasting current activity level- is very helpful in terms of health care management. In fact, Google is providing a special Google Trends service for flu and dengue, named as Google Flu Trends.[24] Basic idea is to use large number of Google Search queries to reveal, if there is any, presence of flu or dengue.

Although papers on influenza like illnesses are more common, first paper in this field, Cooper[25] et al. (2005), is actually on correlation between some search volume data, associated with specific cancers, and estimated incidence rates of these cancer types. On the other hand, pioneering papers on influenza like illnesses are Polgreen[26] et al. (2008) and Ginsberg[27] et al. (2009). Former one worked on the influenza level data from 2004 through 2008 and developed a model that allows prediction of influenza outbreaks in the U.S. two weeks in advance and prediction of mortalities attributable to influenza five weeks in

advance, whereas latter one developed a model that can estimate weekly influenza activity in each region of the U.S. with a reporting lag of about one day.

There have been four streams of papers after these two. One stream is based on applying already developed flu activity nowcasting/forecasting models for some other countries. Wilson and Brownstein[28] (2009) used Google Trends data to predict influenza activity in Canada and Mason[29] et al. (2009) worked on H1N1 influenza epidemics in New Zealand. The question of whether such tools could be applicable for non-English speaking countries or not was answered by Turbelin[30] et al. (2009) and Valdivia and Monge-Corella[31] (2010), where influenza activities in France and Spain nowcasted, respectively.

The other main stream is to implement similar idea to predict or nowcast activity level of other illnesses than influenza. Turbelin[30] et al. (2009) worked on gastroenteritis and chickenpox and showed that their activity level can also be estimated by search volume indices. Ari[32] et al. (2010) worked on another epidemic disease, lyme, and concluded that search volumes can be used to approximate certain trends previously identified in the epidemiology of that disease. Chan[33] et al. (2011) came up with a tool to analyze Google Trends data for an early detection and monitoring of dengue epidemics.

The third set of papers focused on improving current models to monitor and forecast influenza activity more accurately. Doornik[34] (2009) suggested two improvements on Google Flu Trends such that improved model can detect influenza activities that are limited in short time periods, e.g. a few days, and can make better forecasts about the activity level of current influenza prevalence. Liu[35] et al. (2012) stated that dynamic query set of keywords, rather than static, is more accurate in influenza forecast. Recently, Olson[36] et al. (2013) assessed the performance of Google Flu Trends algorithms and concluded that they are not reliable anymore as they missed two important influenza activities in last few years. The reasoning behind is explained as the changes in internet search behavior and age distribution of the epidemics between the periods of Google Flu Trends model fitting and prospective use.

The fourth and last main stream is exploiting some other online data than Google Trends to monitor flu activity. Corley[37] et al. (2009) used blog posts that discussed influenza and Hulth[38] et al. (2009) used a medical website operating in Sweden to monitor flu activity in this country. Both concluded that data from such sources is accurate for syndromic

surveillance. Hassan[39] et al. (2010) stated that online visual platforms for patients to share their experiences such as PatientsLikeMe[40] and PlanetCancer[41] can be also utilized by research on health foresight.

Papers in this research area successfully proved that using Google Trends data matters and can be helpful for stakeholders of health industry.

c) Obesity

In this section, we review papers which focused on economic implications of obesity, epidemiologic nature of obesity, efforts to monitor obesity and geographic distribution of obesity.

Obesity puts substantial amount of economic burden on governments and public. Results in Wolf[42] et al. (2008) indicated that mean healthcare cost for an obese member of U.S. population is twice as much as the mean health care cost for a non-obese member. Parks[43] et al. (2012) stated that 1 unit increase in BMI for every adult in the U.S. would increase annual public medical expenditures by $38.7 billion. Moreover, they concluded that there would be yearly savings of $173.7 billion (in constant 2008$) in public medical expenditures, if there was no obese in the U.S. These two papers suggest that along with its adverse effects on health, obesity is also a big problem for the economy. Indirect costs of obesity should also be considered along with its direct, medical costs. Wolf and Colditz[44] (1998) estimated that each year number of restricted activity days attributable to obesity is increasing 6% on average. It was also estimated that each year, on average, number of bed-days and number of work-lost days attributable to obesity are increasing at rates of 4.5% and 10%, respectively.

In 1997, WHO recognized obesity as global epidemics. [Cabellero[45] (2007)] There have been many studies on verification of epidemic nature of obesity and the importance of monitoring obesity. Christakis and Fowler[46] (2007) proved the epidemiologic nature of obesity in a large social network by examining its spread over 32 years. Gollust[47] et al. (2012) concluded that images that accompany obesity related news coverage can shape public understanding about the social epidemiology of that condition. Li[48] et al. (2013) studied the role of social networks and the use of social media in child obesity. Finkelstein[49] et al. (2005) investigated the underlying economic causes and consequences of obesity epidemic.

Another stream of obesity research focused on monitoring obesity. BMI was frequently used as numerical measure of obesity, since it is easy to estimate and internationally recognized. Espinel and King[50] (2012) stated that monitoring of population weight status is valuable in order to track changes and identify likely causes and implications, and to adjust health policy and program priorities. Lacy[51] et al. (2012) noted that childhood obesity monitoring is fundamental component of obesity prevention. Recently, efforts to predict obesity prevalence have also increased. Majer[52] et al. (2013) developed a model to forecast probability distribution of BMI for different combinations of age and gender in 2020. Similarly, Kim and Basu[53] (2013) developed a statistical model to generate transition probabilities between BMI categories and then used it to forecast future BMI probability distributions among children.

While obesity is seen as an inevitable result of unnecessarily high energy intake, studies on geographic distribution of obesity, focused on two main causes of this high energy intake: economic and socio-cultural. Probably, first paper is Al-Nuaim[54] (1997), which investigated the geographic distribution of obesity. It noted that lifestyle, nutritional habits and socio-cultural beliefs have substantial effect on obesity prevalence in different regions, then gave examples on how economic development and culture of region can affect rates of obesity. Onis and Blossner[55] (2000), Wang and Baydoun[56] (2007) and Ji and Cheng[57] (2008) proved that obesity prevalence in urban, economically strong, and industrialized regions are higher compared to rural, low economic activity regions. Stronger economy brings more wealth and more wealth brings more fast food restaurants, increase in automobile usage, increase in TV/video watching and energy-dense diet. Zhang[58] et al. (2011) and Wang and Lim[59] (2012) concluded similar results for child obesity and noted that rapid economic grow comes with an increase in overweight and obesity prevalence among children and adolescents.

Many publications verified the epidemiologic nature of obesity, so it is appropriate to use search volume indices as an indicator of economic and socio-cultural characteristics of some region and to monitor obesity prevalence, like influenza activity. Any study on monitoring obesity could help to increase life quality of people and to control the cost of health care for the government.

Fast-changing world and paradigms in it make use of innovative methods mandatory to reach more accurate conclusions in a shorter amount of time. Best of our knowledge, the idea of nowcasting obesity in the U.S using Google Search volume data has not been utilized yet.

III. HYPOTHESIS

Data collection and analyses conducted rely on two links that are assumed to be strong in cognitive science. One of them is between physical condition of person and his/her mind set and second one is between online search pattern of person and his/her mind set. For example, if a person is obese, then s/he will think about losing weight and s/he will search keyword "weight loss" more, compared to non-obese fellow.

Obesity develops when energy intake energy exceeds energy expenditure over a prolonged period of time.[60] Body weight is also the result of genes, metabolism, behavior, environment, culture and socioeconomic status.[61] However, Karam and McFarlane[62] (2007) noted that only some small portion of obesity prevalence can be attributed to the secondary causes rather than high calorie intake and sedentary lifestyle.

In that manner, we constructed our hypothesis on five different groups of keywords. While first four, high energy intake positive/negative and sedentary life style positive/negative, are causes of obesity, third group covers consequences of obesity. For example, keyword of "weather" belongs to the group of sedentary life style negative and its low search volume is actually indicator of sedentary life style, so contributes to higher obesity prevalence for a region in question. Figure.1 depicts the scheme of keyword groups and their relations with obesity prevalence.

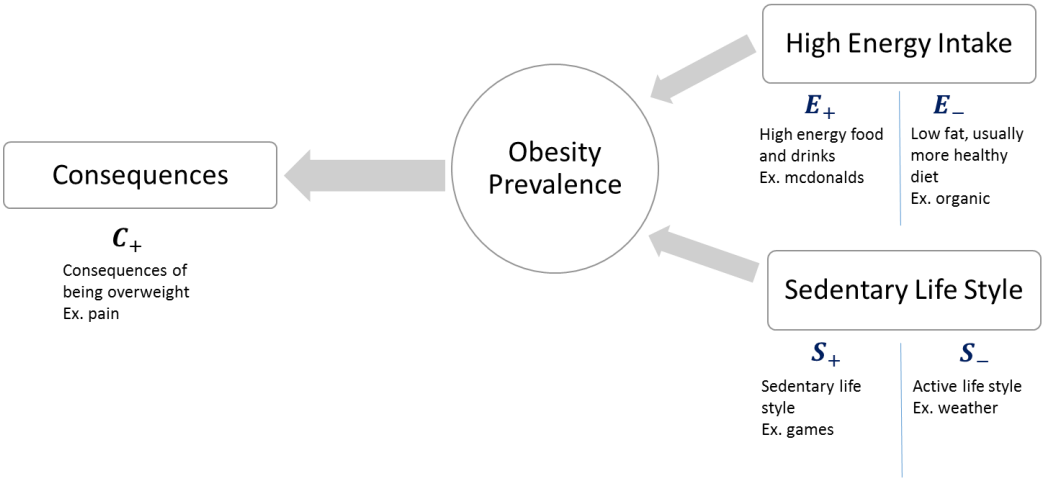


Figure.1: Keyword groups with respect to their effects

According to the U.S. Department of Health and Human Services[63], a lack of energy balance is the most often cause of overweight and obesity. When the amount of energy or calories that have been gotten from food and drinks is more than the amount of energy used by human body for breathing, digesting, moving etc., this energy unbalance causes the weight gain. Muntel[64] (2012) stated that fast food can lead to weight gain. In that manner, we hypothesized that (Hypothesis-1a) high search volume of keywords in the E_+ group is indicator of high obesity prevalence. On the opposite site of the story, we know that being careful with daily diet and having more healthy food can help people to avoid obesity or to mitigate its severity[65] [66]. That being said, we propose that (Hypothesis-1b) high search volume of keywords in the E_- group is indicator of low obesity prevalence.

While food and drinks constitutes one side of the energy balance equation mentioned above, physical activity is responsible for the other side. U.S. Department of Health and Human Services[65] recommends being active as one of the few ways of preventing obesity. Based on the fact that Fairnardi[67] et al. (2009) stated that watching TV, playing video games and reading a book are sedentary lifestyle patterns, we determine our related two hypothesis as (Hypothesis-2a) high search volume of keywords in the S_+ group is indicator of high obesity prevalence and (Hypothesis-2b) high search volume of keywords in the S_- group is indicator of low obesity prevalence.

While energy balance and active/sedentary lifestyle can be interpreted as choices leading to being overweight or obese, there are also implications of obesity that might possibly be observed by checking search pattern of appropriate keywords. Stone and Broderick[68] (2012) reported that higher BMI comes with higher rates of pain. With the intuitive assumption that very high portion of overweight/obese people would like to lose weight, we set our Hypothesis-3 as high search volume of keywords in the C_+ group is indicator of high obesity prevalence.

We chose to start with many number of keywords to stay as flexible as possible. Then we eliminated some of them to get the final set of categorized keywords which performed relatively better in terms of explaining variation of BMI.Table.1 lists keywords that have been included in the final set with respect to their groups.

Table.1: Categorized keywords

High Energy Intake		**Sedentary Life Style**		**Consequences**
+	**-**	**+**	**-**	**+**
mcdonalds	organic	books	fitness	pain
dominos	farmers market	games	parks	weight loss
pizza	diet	movies	weather	
			bar	

IV. DATA

We worked with two sets of data. In first subsection, BRFSS data and its collection process is discussed and then handling of Google Trends data is explained.

a) BRFSS Data

Each year BRFSS conducts telephone based surveys with approximately 400,000 U.S. residents from different metro areas of the country and collects data on their health. BRFSS annually releases the results of this survey; however, these releases are typically only available with a reporting lag of approximately 10 months and are often revised a few years later.

BRFSS data is readily available without any prior approval of usage so anybody can go online and download the data. It keeps track of hundreds of health related indicators of thousands (Appendix-6) of U.S. residents so it is one of the most representative and detailed datasets on health of the U.S. population. BRFSS representatives call thousands of people from different states/counties and ask them some demographic questions such as "Number of adults in the household" or "County that has been lived in" etc. and also some health related questions such as "Whether respondent had any stroke" or "weight and height of respondent" etc. Results are compiled at the end of each year and published in a single dataset in next year. A codebook accompanies the dataset to list estimated variables and their categories. (Appendix-7) We used SAS program and PERL codes to process data and make it ready for analysis.

In BRFSS, respondents' state and county information coded in ANSI codes, which were defined by U.S. Census Bureau[69] data is available in county level. On the other hand, Google Trends provides data for metro areas defined by Nielsen[70]. Each of the 3143 counties listed in survey data were assigned to one of the 210 metro areas defined by Google Trends.

We used a computer program to first read survey data and county-metro area assignments and then to estimate average BMI for each of the 210 metro areas. (Appendix-8) Figure.2 depicts frequency histogram of BMI averages of metro areas for some selected years.

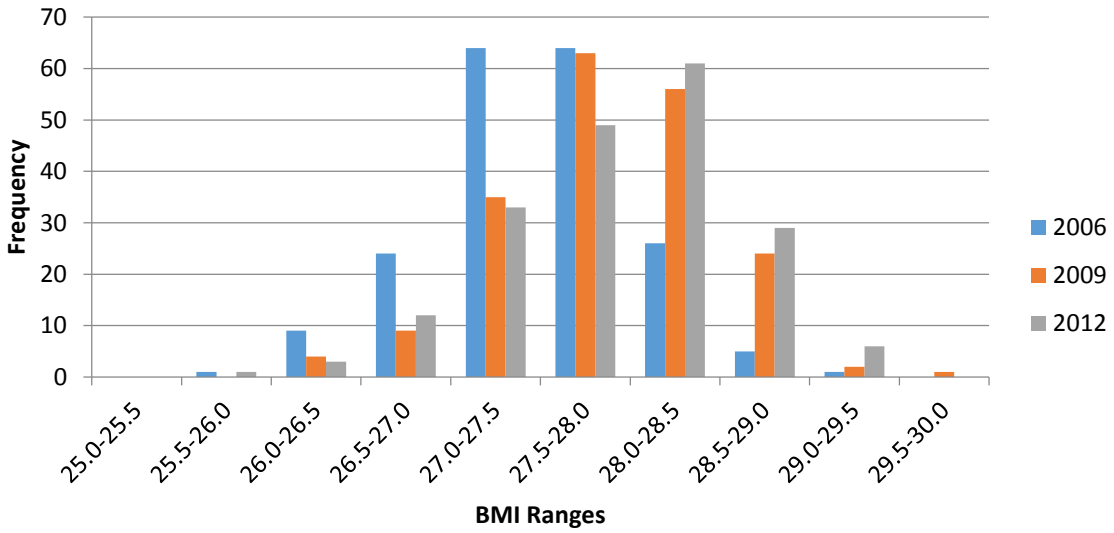


Figure.2: Frequency histogram of BMI averages of metro areas for selected years of 2006, 2009 and 2012

We chose to restrict our study between the years of 2006 and 2012, since results of BRFSS 2013 have not been published yet and Google Trends data is available for most of the metro areas after 2006. Nevertheless, it provides search volume indices for many metro areas and compensates the rather short time series available.

According to the internationally recognized BMI categorization, person with BMI value of less than 18.5 is considered as underweight, person with BMI value from 18.5 to 25 is considered as normal (healthy weight), person with BMI value from 25 to 30 is considered overweight and person with BMI value of greater than 30 is considered obese.

b) Google Trends Data

Google Trends data is provided by Google and available starting from 2004 through today. There are two challenges affiliated with its collection process. Researchers should download data in .csv files separately for each of the metro area and keyword combinations. For instance, should search volume index of "diet" needed, then 210 separate downloads have to be made. Secondly, Google Trends rescales the search volume and publishes only search volume indices, not absolute search volumes, so indices of metro areas are not directly comparable. These index values are estimated by Google by dividing total query volume for search term in a given geographic region divided by the total number of queries in that region at a point in time.* Then these shares are normalized so that to take values between 0 and 100.

Former challenge was overcome by coding URLs and submitting tens of them at a time, which allows downloading multiple .csv files. (Appendix-9) Latter one was solved by downloading comparative data for U.S. and each metro areas (Appendix-3) rather than downloading data for only one single metro area (Appendix-2). Given that absolute search volumes resulted U.S. search volume indices are same for each metro area, we used them to

* For detailed work on calculation of Google Trends, please see Choi and Varian[71] (2012). Let S_{rtk} represents the search volume of keyword k during time interval of t for the region r. Moreover, let $S_{rt} = \sum_k S_{rtk}$ denotes the total number of search queries made from region r during time interval of t. First, Google estimates the shares as $\frac{S_{rtk}}{S_{rt}}$. Then, shares are normalized between 0 and 100. Note that although $\frac{S_{rtk}}{S_{rt}}$ is constant, normalized value is subject to change with respect to selected time interval, such as from 2006 through 2010 or just year of 2012, metro areas compared at a time, if any, and keywords compared at a time, if any.

rescale data and make it comparable across metro areas and as a result of some benchmarking procedure we got index values which are comparable across regions and time.

Datasets comparable across regions and time:

In order to successfully implement benchmarking idea here, it has been chosen to download data for each region along with some benchmark region that has enough number of search volumes for every keyword and time combination. In our model, we found it appropriate to use the U.S as our benchmark region.

Once we download data for one metro area's search volume indices relative to the U.S., we are getting a table of indices where vertical axis stands for time, weekly intervals, t=1,2,….,T and horizontal axis stands for regions, in this case we have two of them as the metro area in question, r=1,2,….,R and benchmark region, namely the U.S. Note that for each keyword, k=1,2,….,K that we are interested in separate set of downloads should be made.

Let I_{rtk} represents search volume index of time period t for the download made for region r and keyword k and I^*_{rtk} represents search volume index of time period t for the benchmark region, the U.S., relative to the region in question and again for keyword k. For one single keyword and for R number of regions, R separate downloads should be made and each download will contain tables with T rows and 2 columns filled with I_{rtk} and I^*_{rtk} values.

In order to normalize search volume indices so that they will be comparable across regions and time, one region should be picked as base and all others' indices should be adjusted accordingly. An educated selection is made based on performance measure of $\frac{\sum_{t=1}^{T} I_{rtk}}{\sum_{t=1}^{T} I^*_{rtk}}$ and region $r_k^{\#}$ is selected as base region where $r_k^{\#} = argmax_r \frac{\sum_{t=1}^{T} I_{rtk}}{\sum_{t=1}^{T} I^*_{rtk}}$ for each k=1,2,…K

Let I'_{rtk} denotes adjusted index values. Base region's index values will remain unchanged $I'_{r_k^{\#}\,tk} = I_{r_k^{\#}\,tk}$ and others will be updated as $I'_{rtk} = \frac{I_{rtk} * I^*_{r_k^{\#}tk}}{I^*_{rtk}}$ for all k=1, 2, …, K

Appendix-10 provides small part of output file. Although, I'_{rtk} which is comparable across time and region is sufficient as qualified input to our model, we further dug to get index values which are comparable across all of the three dimensions, regions, keywords and time.

(Appendix-11) In order to achieve that, it is necessary to download another stream of data from Google Trends: comparable across keywords and time. After this step, some reconciliation between two available data sets should be conducted to succeed comparability across all three dimensions.

It is usually possible to get a particular metro area's search volume index for a given keyword and for a given year but exceptions may apply. There were not enough search volume for some keyword and year combinations.

Note that BRFSS data is compiled from survey with thousands of respondents, while Google Trends data is an outcome of millions of people's billions of search queries. Therefore, monitoring obesity using Google Trends data should yield more representative results, compared to BRFSS data.

V. ANALYSIS

Analyses conducted in this section are to test five hypothesis listed above. A crucial decision along with keyword selection is whether BMI average of metro area, or percentage of obesity (or overweight and obesity) prevalence in metro area should be used as dependent variable. Our analyses clearly showed that working with percentages rather than BMI average causes some lost in precision of data. For instance, BMI of a person can go up from 21 to 24.5 where s/he will still be considered as not obese. Therefore, we worked with BMI averages of metro areas rather than percentages. The rest of the section includes details of our panel data setup and nowcasting efforts.

Both BRFSS and Google Trends data have two dimensions, temporal and cross-sectional. Time range is from 2006 through 2011, and cross-sections are 210 geographic regions, metro areas, defined by Google Trends. We left data for 2012 untouched to be utilized evaluating the performance of models.

Panel data analysis can be conducted to gain deeper insights into variation within (fixed effect) and between (between effect) metro areas. Hausman robust tests suggested that our data is appropriate and consistent for fixed effect, so fixed effect estimator was used.

In panel data analysis, we run 216 different unbalanced panel data models, each being different combinations of keywords from categories listed above. In each model, there is exactly one keyword from each category and we chose to opt out using combination of keywords as a representative of some category as determining appropriate weights is not quite plausible without any bias involved. As a result, for each model constant term, coefficients of keywords and coefficient of dichotomous variables for regions were estimated.

Using the estimated coefficients and constants terms for the models and corresponding Google Trends data for year of 2012, BMI averages of regions for 2012 were nowcasted as if data has not been published yet.

Mean Absolute Percentage Error (MAPE) was used as a performance measure of models, where benchmark model imitates nowcasting BMI averages of regions for 2012 without any Google Trends data. MAPE is a measure of accuracy for fitted statistical model like ours. It is basically average of absolute values of differences between actual value and nowcast value divided by actual value over all possible nowcasts for a given model. We found it appropriate to report top five models with respect to the percentage improvement they provide over benchmark model. Table.2 lists summary statistics for those models. Coefficients of regional dummies are not displayed due to the size of the table for display.

In all of the five models we have more than 50% improvement with respect to the benchmark models' MAPE and average number of observations is around 260, which means that there are more than 40 regions involved for each one of the years from 2006 through 2011. Signs of the coefficients are as expected except coefficients of "farmers market" and "books" in the third best model. On the other hand, as they are being very small there is no prominent effect on the nowcasting power of the model.

Results in Table.2 provides supporting evidence to the five hypothesis listed above. Hypothesis-1a proposes that coefficients of "mcdonalds", "dominos" and "pizza" are positive which is clearly a valid argument. On the other hand, Hypothesis-1b suggests to have negative coefficient for keyword of "farmers market", which is still valid but not as strong as the first hypothesis. Keywords that are predicted to take positive coefficients by Hypotheses 2a, have all positive coefficients with one exception. Hypothesis-2b and 3 are possibly have

the strongest evidence as coefficients of all corresponding keywords to these hypotheses are as expected and moreover, they are all statistically significant.

Table.2: Summary statistics of best five models

Rank	Benchmark	1	2	3	4	5
Number of observations	1225	263	265	257	263	257
R-square	0.788	0.938	0.934	0.929	0.936	0.935
MAPE	0.0099	0.0046	0.0047	0.0047	0.0048	0.0048
Improvement	-	53.74%	52.89%	52.28%	52.24%	51.84%
VARIABLES						
mcdonalds				0.004		0.006*
				(0.003)		(0.003)
dominos		0.002				
		(0.003)				
pizza			0		0.006**	
			(0.002)		(0.002)	
farmers market		-0.002	0	0.001	-0.003	-0.001
		(0.002)	(0.002)	(0.003)	(0.002)	(0.003)
books				-0.001		
				(0.005)		
games		0.019***			0.025***	
		(0.006)			(0.006)	
movies			0.012**			0.013***
			(0.005)			(0.004)
fitness				-0.010**		
				(0.004)		
weather		-0.014**	-0.011*		-0.013**	-0.014**
		(0.006)	(0.007)		(0.006)	(0.006)
pain		0.010**	0.009*			
		(0.004)	(0.005)			
weight loss				0.012***	0.009***	0.008**
				(0.004)	(0.003)	(0.003)
Constant	27.661***	26.707***	26.656***	27.086***	26.619***	26.778***
	(0.201)	(0.213)	(0.232)	(0.189)	(0.212)	(0.242)
Robust standard errors in parentheses				*** p<0.01, ** p<0.05, * p<0.1		

VI. CONCLUSION

Obesity is a tremendously huge problem for U.S. in terms of the life quality of individuals, and also in terms of the government expenditures. U.S. Health System is carrying undue burden of billions of dollars due to recently increased rates of obesity. Each year the government agencies fund obesity research regarding its causes, its consequences, its treatment methods, and its traceability.

While there are thousands of factors potentially related to obesity, only a small portion of them can be conceptualized, and even a smaller fraction of can be observed with relatively shorter latencies. In a striking contrast, here we show that people's use of the Google Search, so Google Trends data, can provide an almost real-time window into the general trends of obesity across the populations of different regions, potentially opening ways to new insights into the disease of obesity.

In this paper, analyses have been conducted with the purpose of shedding light on the relationship between search volume indices of some keywords, e.g. "weight loss", and obesity prevalence. Based on panel data analyses, conducted with fixed effect estimator it has been concluded that search volume indices play important role in explaining "within" BMI average variation of metro areas.

Our analyses suggest that Google Search volume index can be useful to monitor/nowcast obesity prevalence across different metro areas of the U.S. It is a low cost tool with easy implementation and fairly reliable.

REFERENCES

[1] How Trends Data is normalized – Trends Help
https://support.google.com/trends/answer/4365533?hl=en&ref_topic=4365599
(last access on 05/01/2014)

[2] Ogden, C.L., Carroll, M.D., McDowell, M.A. and Flegal, K.M. (2007) Obesity among adults in the United States-no change since 2003-2004. *NCHS data brief no 1. Hyattsville, MD: National Center for Health Statistics.*

[3] WHO | United States of America, Statistics, http://www.who.int/countries/usa/en/
(last access on 05/01/2014)

[4] Obamacare's Average Monthly Cost Across US: $328,
http://news.msn.com/us/obamacares-average-monthly-cost-across-us-dollar328?stay=1v
(last access 05/01/2014)

[5] Finkelstein, E.A., Trogdon, J.G., Cohen, J.W., and Dietz, W. (2009) Annual medical spending attributable to obesity: Payer- and service-specific estimates. *Health Affairs*.

[6] Behavioral Risk Factor Surveillance System, http://www.cdc.gov/brfss/
(last access on 05/01/2014)

[7] Ettredge, M., Gerdes, J. and Karuga, G. (2005) Using Web-based Search Data to Predict Macroeconomic Statistics. *Communications of ACM.*

[8] Choi, H. and Varian, H. (2012) Predicting the Present with Google Trends. *Economic Record*.

[9] D'Amuri F. and Marcucci, J. (2009) Google it! Forecasting the US unemployment. *MPRA Paper18248, University Library of Munich, Germany.*

[10] Askitas, N. and Zimmermann, K.F. (2009) Google econometrics and unemployment forecasting. *IZA Discussion Paper No 4201*.

[11] D'Amuri, F. (2009) Predicting unemployment in short samples with internet job search query data. *Munich Personal RePEc Archive.*

[12] Anvik, C. and Gjelstad, K. (2010) "Just Google it." Forecasting Norwegian unemployment figures with web queries. *Center for Research in Economics and Management Publications.*

[13] Preis, T., Reith, D. and Stanley, H.E. (2010) Complex dynamics of our economic life on different scales: insights from search engine query data. *Philosophical Transactions of the Royal Society A.*

[14] Da, Z., Engelberg, J. and Gao, P. (2011) In Search of Attention. *The Journal of Finance.*

[15] Vosen, S. and Schmidt, T. (2011) Forecasting private consumption: survey-based indicators vs. Google Trends. *Journal of Forecasting.*

[16] Huang, H. and Penna, N.D. (2009) Constructing Consumer Sentiment Index for U.S Using Google Searches. *University of Alberta, Department of Economics.*

[17] Wu, L. and Brynjolfsson, E. (2010) The future of prediction: how Google searches foreshadow housing prices and sales. *NBER Conference Technological Progress & Productivity Measurement.*

[18] Goel, S., Hofman, J.M., Lahaie, S., Pennock, D.M. and Watts, D.J. (2010) Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences of the United States of America.*

[19] Guzman, G. (2011) Internet search behavior as an economic forecasting tool: The case of inflation expectations. *Journal of Economic and Social Measurement.*

[20] Lindberg, F. (2011) Nowcasting Swedish Retail Sales with Google Search Query Data. *Stockholm University.*

[21] Lui, C., Metexas, P.T. and Mustafaraj, E. (2011) On the predictability of the U.S. elections through search volume activity. *Predicting with Social Media.*

[22] Metaxas, P.T., Mustafaraj, E. and Gayo-Avello, D. (2011) How (Not) to Predict Elections. *Privacy, Security, Risk and Trust.*

[23] Reilly, S., Richey, S. and Taylor, B. (2012) Using Google search data for state politics research: an empirical validity test using roll-off data. *State Politics & Policy Quarterly.*

[24] Explore flu trends – United States, http://www.google.org/flutrends/us/#US (last access on 05/01/2014)

[25] Cooper, C.P., Mallon, K.P., Leadbetter, S., Pollack, L.A., and Peipins, L.A. (2005) Cancer internet search activity on a major search engine. *Journal of Medical Internet Search.*

[26] Polgreen, P.M., Chen, Y., Pennock, D.M. and Nelson, F.D. (2008) Using internet searches for influenza surveillance. *Clinical Infectious Diseases.*

[27] Ginsberg, J., Mohebbi. M.H., Patel, R.S., Brammer, L., Smolinski, M.S. and Brilliant L. (2009) Detecting influenza epidemics using search engine query data. *Nature 457.*

[28] Wilson, K. and Brownstein, J. (2009) Early detection of disease outbreaks using the Internet. *Canadian Medical Association Journal.*

[29] Mason, K., Tobias, M., Peacey, M., Huang, Q.S. and Baker, M. (2009) Interpreting Google Flu Trends data for pandemic H1N1 influenza: the New Zealand experience. *Eurosurvelliance.*

[30] Turbelin, C., Pelat, C., Bar-Hen, A., Flahault, A., and Valleron, A. (2009) More disease tracked by using Google Trends. *Emerging Infectious Diseases.*

[31] Valdivia, A. and Monge-Corella, S. (2010) Diseases tracked by using Google trends. *Emerging Infectious Diseases.*

[32] Ari, S., Alison, S., Geis, K. and Aucott, J. (2010) The utility of "Google Trends" for epidemiological research: Lyme disease as an example. *Geospatial Health.*

[33] Chan, E.H., Sahai, V., Conrad, C. and Brownstein, J.S. (2011) Using Web Search Query Data to Monitor Dengue Epidemics: A New Model for Neglected Tropical Disease Surveillance. *PLOS Neglected Tropical Diseases.*

[34] Doornik, J.A. (2009) Improving the Timeliness of Data on Influenza-like Illnesses using Google Search Data. *University of Oxford.*

[35] Liu, F., Lv, B., Peng, G. and Li, X. (2012) Influenza Epidemics Detection Based on Google Search Queries. *Recent Progress in Data Engineering and Internet Technology.*

[36] Olson, D.R., Konty, K.J., Paladini, M., Viboud, C. and Simonsen L. (2013) Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographical Scales. *PLOS Computational Biology.*

[37] Corley, C.D., Mikler, A.R., Singh, K.P. and Cook, D.J. (2009) Monitoring Influenza Trends through Mining Social Media. *International Conference on Bioinformatics and Computational Biology.*

[38] Hulth, A., Rydevik, G. and Linde, A. (2009) Web Queries as a Source for Syndromic Surveillance. *PLOS One.*

[39] Hassan, M., Joddy, R. and Peter, A.S. (2010) Five promising methods for health foresight. *The Journal of Future Studies, Strategic Thinking and Policy.*

[40] Live Better, Together | PatientsLikeMe, http://www.patientslikeme.com/
(last accessed on 05/01/2014)

[41] Planet Cancer, http://myplanet.planetcancer.org/
(last accessed on 05/01/2014)

[42] Wolf, A.M., Finer, N., Allshouse, A.A., Perdergast, K.B., Sherrill, B.H., Caterson, I., Hill, J.O., Aronne, L.J., Hauner, H., Radigue, C., Amand, C. and Despres, J.P. (2008) PROCEED: Prospective Obesity Cohort of Economic Evaluation and Determinants: baseline health and healthcare utilization of the US sample. *Diabetes, Obesity & Metabolism.*

[43] Parks, J.C., Alston, J.M. and Okrent, A.M. (2012) The Marginal External Cost of Obesity on the United States. *Agricultural and Applied Economics Association.*

[44] Wolf, A.M. and Colditz, G.A. (1998) Current estimates of the economic cost of obesity in the United States. *Obesity Research.*

[45] Cabellero, B. (2007) The Global Epidemic of Obesity: An Overview. *Epidemiologic Reviews.*

[46] Christakis, N.A. and Fowler, J.H. (2007) The Spread of Obesity in a Large Social Network over 32 Years. *The New England Journal of Medicine.*

[47] Gollust, S.E., Eboh, I. and Barry, C.L. (2012) Picturing obesity: Analyzing the social epidemiology of obesity conveyed through US news media images. *Social Science & Medicine.*

[48] Li, J.S., Tracie, C., Barnett, A. and Goodman, E. (2013) Approaches to the Prevention and Management of Childhood Obesity: The Role of Social Networks and the Use of Social Media and Related Electronic Technologies. *AHA Scientific Statement.*

[49] Finkelstein, E.A., Ruhm, C.J. and Kosa, K.M. (2005) Economic Causes and Consequences of Obesity. *Annual Review of Public Health.*

[50] Espinel, P. and King, L. (2012) A Framework For Monitoring Overweight and Obesity in NSW. *NSW Department of Health and the Physical Activity Nutrition Obesity Research Group.*

[51] Lacy, K., Kremer, P., de Silva-Sanigorski, A., Allender, S., Leslie, E., Jones, L., Fornaro, S. and Swinburn, B. (2012) The appropriateness of opt-out consent for monitoring childhood obesity in Australia. *Pediatric Obesity.*

[52] Majer, I.M., Mackenbach, J.K. and van Baal, P.H.M. (2013) Time trends and forecasts of body mass index from repeated cross-sectional data: a different approach. *Statistics in Medicine.*

[53] Kim, D.D. and Basu, A. (2013) Forecasting Body Mass Index Distributions Among the US Children Using a Longitudinal Dataset. *Society for Medical Decision Making.*

[54] Al-Nuaim, A.R. (1997) Population-Based Epidemiological Study of the Prevalence of Overweight and Obesity in Saudi Arabia, Regional Variation. *Annals of Saudi Medicine.*

[55] Onis, M. and Blossner, M. (2000) Prevalence and trends of overweight among preschool children in developing countries. *The American Journal of Clinical Nutrition.*

[56] Wang, Y. and Beydoun, M.A. (2007) The Obesity Epidemic in the United States-Gender, Age, Socioeconomic, Racial/Ethnic, and Geographic Characteristics: A Systematic Review and Meta-Regression Analysis. *Epidemiologic Reviews.*

[57] Ji, C.Y. and Cheng, T.O. (2008) Prevalence and geographic distribution of childhood obesity in China in 2005. *International Journal of Cardiology.*

[58] Zhang, J., Seo, D., Kolbe, L., Middlestadt, S. and Zhao, W. (2011) Associated Trends in Sedentary Behavior and BMI among Chinese School Children and Adolescents in Seven Diverse Chinese Provinces. *International Journal of Behavioral Medicine.*

[59] Wang, Y. and Lim, H. (2012) The global childhood obesity epidemic and the association between socio-economic status and childhood obesity. *International Review of Psychiatry.*

[60] John, P.H. and Wilding, D.M. (2001) Causes of obesity. *Practical Diabetes.*

[61] Obesity and Overweight for Professionals: Adult: Causes – DNPAO – CDC, http://www.cdc.gov/obesity/adult/causes/index.html
(last accessed on 05/01/2014)

[62] Karam, J.G. and McFarlane, S.I. (2007) Secondary causes of obesity. *Therapy.*

[63] What Causes Overweight and Obesity? – NHLBI, NIH, http://www.nhlbi.nih.gov/health/health-topics/topics/obe/causes.html
(last access on 05/15/2014)

[64] Muntel, S. (2012) FAST FOOD – Is it the enemy? *Research Articles*, *Obesity Action Coalition.*

[65] How Can Overweight and Obesity Prevented? – NHLBI, NIH, http://www.nhlbi.nih.gov/health/health-topics/topics/obe/prevention.html
(last access on 05/15/2014)

[66] Local Farmers Markets Help Reduce Obesity Rates | KUNM, http://kunm.org/post/local-farmers-markets-help-reduce-obesity-rates
(last access on 05/15/2014)

[67] Fainardi, V., Scarabello, C., Brunella, I., Errico, M.K., Mele, A., Gelmetti, C., Sponzilli, I., Chiari, G., Volta, E., Vitale, M. and Vanelli, M. (2009) Sedentary lifestyle in active children admitted to a summer sport school. *Acta Biomed.*

[68] Stone, A.A. and Broderick, J. E. (2012) Obesity and Pain Are Associated in the United States. *Obesity*.

[69] 2010 ANSI Codes for Places – Geography – U.S. Census Bureau, https://www.census.gov/geo/reference/codes/place.html
(last access on 05/01/2014)

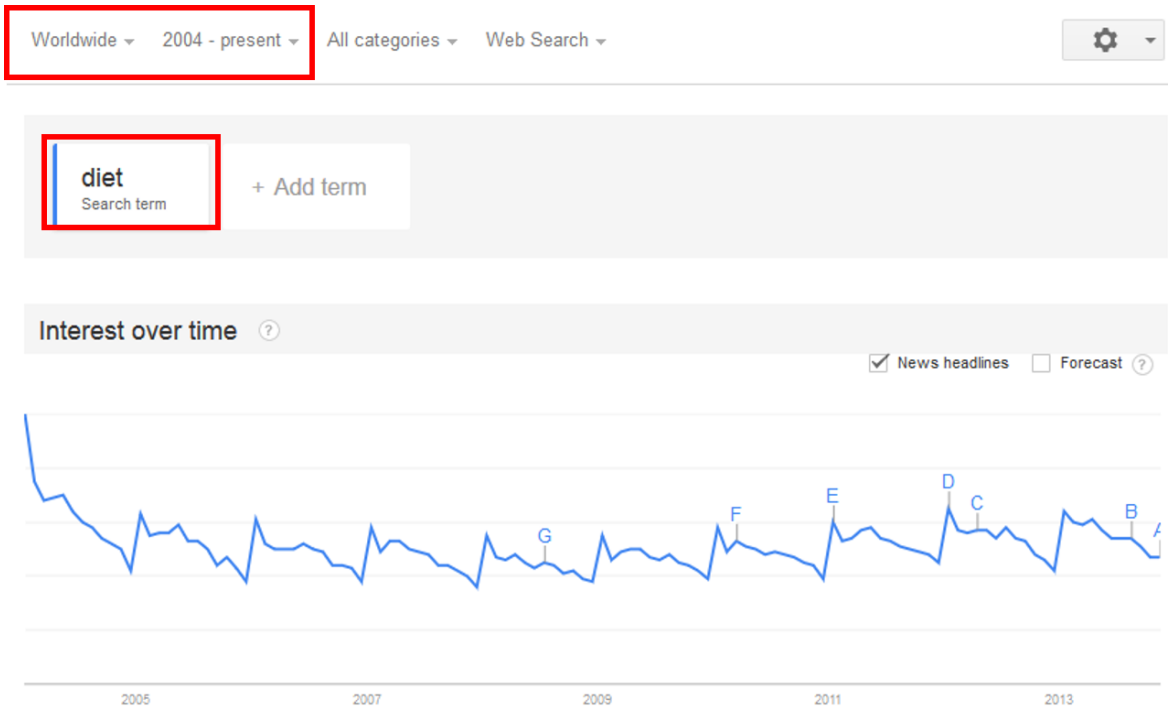
[70] Nielsen Audio, http://www.nielsen.com/us/en/nielsen-solutions/audience-measurement/nielsen-audio.html
(last access on 05/01/2014)

[71] Choi, H. and Varian, H. (2012) Predicting the Present with Google Trends. *Economic Record*.

APPENDICES

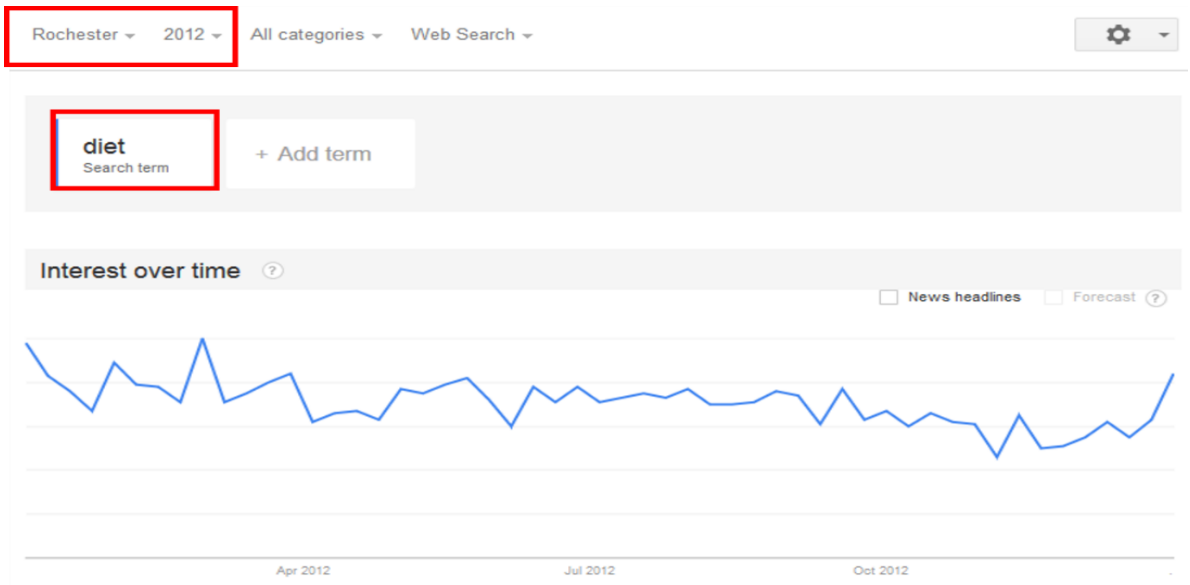
APPENDIX-1

Google Trends graph for keyword of "diet", default settings.



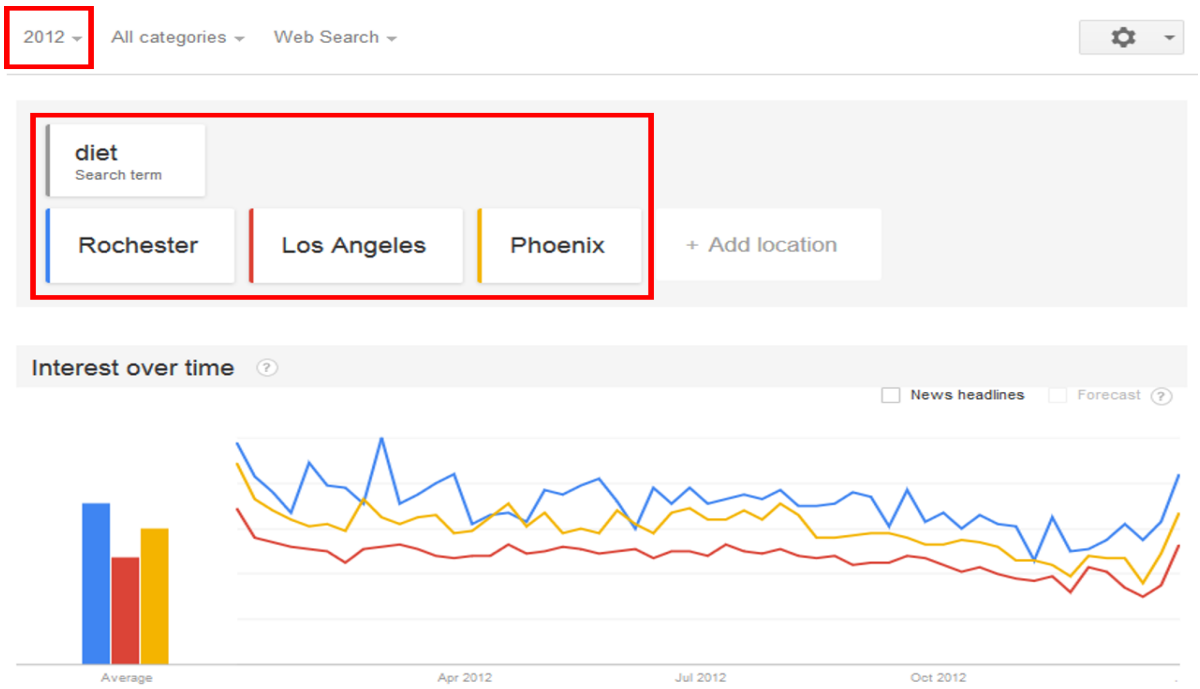
APPENDIX-2

Google Trends graph for keyword of "diet", Rochester Metro Area, 2012.



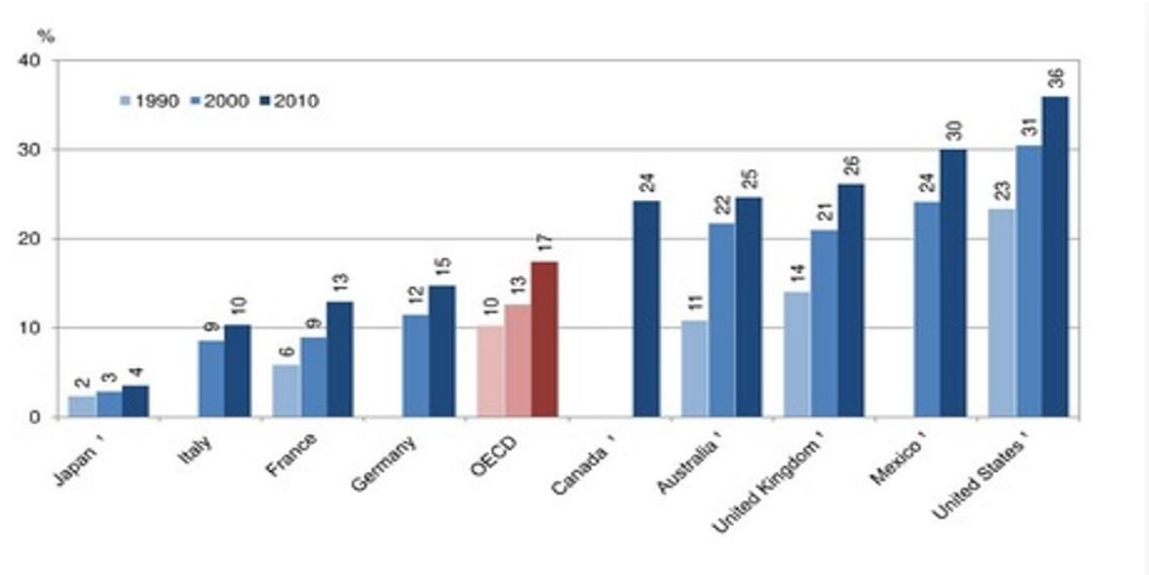
APPENDIX-3

Google Trends graph for keyword of "diet", Rochester, Los Angeles and Phoenix Metro Areas, 2012.



APPENDIX-4

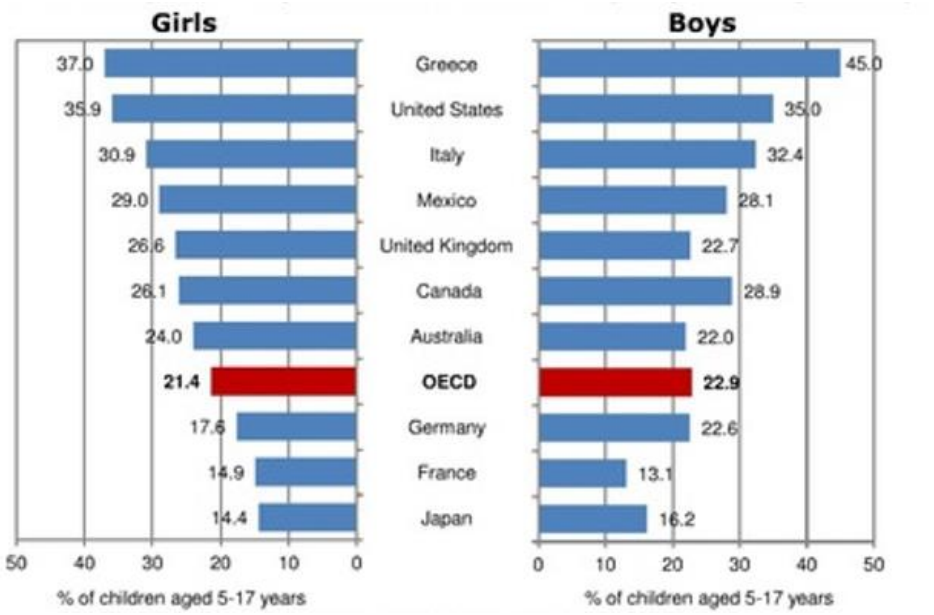
Measured obesity rates from 1990 through 2010.



Source: OECD Health Data 2012 Dissemination and Results (2012)

APPENDIX-5

Children aged 5-17 years who are overweight (including obese).



Source: Health at a Glance 2011: OECD Indicators, Non-medical Determinants of Health, Overweight and Obesity Among Children.

APPENDIX-6

Number of respondents in BRFSS from 2004 through 2012.

Year	Number of respondents
2004	303,822
2005	356,112
2006	355,710
2007	430,912
2008	414,509
2009	432,607
2010	451,075
2011	504,408
2012	475,687

APPENDIX-7

A page from Codebook 2012.

Computed body mass index

CalculatedVariables: 7.17 Calculated Variables **Type:** Num

Column: 1644-1647 **SAS Variable Name:** _BMI5

Prologue:

Description: Body Mass Index (BMI)

Value	Value Label	Frequency	Percentage	Weighted Percentage
1 - 9999	1 or greater Notes: WTKG3/(HTM4*HTM4) (Has 2 implied decimal places)	450,212	100.00	100.00
BLANK	Don't know/Refused/Missing Notes: WTKG3 = 777 or 999 or HTM4 = 777 or 999	25,475		

Computed body mass index categories

CalculatedVariables: 7.18 Calculated Variables **Type:** Num

Column: 1648 **SAS Variable Name:** _BMI5CAT

Prologue:

Description: Four-categories of Body Mass Index (BMI)

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Underweight Notes: _BMI5 < 1850 (_BMI5 has 2 implied decimal places)	7,803	1.73	1.88
2	Normal Weight Notes: 1850 <= _BMI5 < 2500	151,985	33.76	34.66
3	Overweight Notes: 2500 <= _BMI5 < 3000	162,768	36.15	35.77
4	Obese Notes: 30000 <= _BMI5 = 9999	127,656	28.35	27.69
BLANK	Don't know/Refused/Missing Notes: _BMI5 = 9999	25,475		

APPENDIX-8

BMI averages of some metro areas in 2012.

Metro Area	BMI Average	Metro Area	BMI Average
Albuquerque	27.225	Las Vegas	27.434
Atlanta	27.493	Miami	27.122
Baltimore	27.897	New York	27.057
Boston	27.208	Phoenix	27.223
Chicago	27.874	Rochester	27.235
Denver	26.440	San Francisco	26.510
Houston	27.710	Tampa	27.318

APPENDIX-9

Example URLs for keyword "weather".

http://www.google.com/trends/trendsReport?q=weather&content=1&export=1&geo=US%2CUS-NY-538&cmpt=geo (note that 538 is metro area code of Rochester)

http://www.google.com/trends/trendsReport?q=weather&content=1&export=1&geo=US%2CUS-WA-819&cmpt=geo (note that 819 is metro area code of Seattle-Tacoma)

http://www.google.com/trends/trendsReport?q=weather&content=1&export=1&geo=US%2CUS-TX-623&cmpt=geo (note that 623 is metro area code of Dallas-Fort Worth)

APPENDIX-10

Rescaled search volume indices for some metro areas. (keyword "diet", year 2012)

Metro Area	Rescaled Search Volume Index	Metro Area	Rescaled Search Volume Index
Alpena	0	Orlando	35.088
Baltimore	34.823	Philadelphia	31.349
Boston	30.511	Portland	37.616
Cleveland	35.575	Raleigh-Durham	36.798
Columbia	45.809	Salt Lake City	37.735
Denver	32.627	Seattle-Tacoma	34.137
Lafayette	56.486	Syracuse	42.025
Meridian	0	Washington	29.622
New Orleans	41.841	Wichita	43.724

APPENDIX-11

Datasets comparable across keywords and time:

In order to successfully implement benchmarking idea here, it has been chosen to download data for each keyword along with some benchmark keyword that has being searched frequently for every region and time combination. Appropriate benchmark keyword selection should be made according to the pool of keywords in question.

Once we download data for one keyword's search volume indices relative to the benchmark keyword, we are getting a table of indices where vertical axis stands for time, weekly intervals, t=1,2,….,T and horizontal axis stands for keywords, in this case we have two of them as the keyword in question, k=1,2,….,K and benchmark keyword. Note that for each region, r=1,2,….,R that we are interested in separate set of downloads should be made.

Let J_{rtk} represents search volume index of time period t for the download made for keyword k and region r and J^*_{rtk} represents search volume index of time period t for the benchmark keyword relative to the keyword in question and again for region r. For one single region and for K number of keywords, K separate download should be made and each download will contain tables with T rows and 2 columns filled with J_{rtk} and J^*_{rtk} values.

In order to normalize search volume indices so that they will be comparable across keywords and time, one of the keywords in question should be picked as base and all others' indices should be adjusted accordingly. An educated selection is made based on performance measure of $\frac{\sum_{t=1}^{T} J_{rtk}}{\sum_{t=1}^{T} J^*_{rtk}}$ and keyword $k_r^{\#}$ is selected as base region where $k_r^{\#} = argmax_k \frac{\sum_{t=1}^{T} J_{rtk}}{\sum_{t=1}^{T} J^*_{rtk}}$ for each r=1,2,…R

Let J'_{rtk} denotes adjusted index values. Base keyword's index values will remain unchanged $J'_{k_r^{\#}\,tk} = J_{k_r^{\#}\,tk}$ and others will be updated as $J'_{rtk} = \frac{J_{rtk} * J^*_{k_r^{\#}\,tk}}{J^*_{rtk}}$ for all r=1,2,…R

Resulting J'_{rtk} values are comparable across keywords and time. We have shown estimation of set of I'_{rtk} within the text and estimation of set of J'_{rtk} above. Next, final reconciliation step should be conducted to merge these two sets to get index values which are comparable across all three dimensions of regions, keywords and time.

<u>Reconciling two data sets:</u>

Let V_{rtk} stands for actual search volume of keyword k, in region r for a time interval of t.

For a given time t, we can get two sets of ratios: one contains ratios of regions and one for ratios of keywords.

For the first set of ratios, let's assume that we are interested in keywords of k=1,2,…,K for some number of regions R, where r=1,2,…,R then for each keyword k we have;

$\frac{V_{1tk}}{V_{2tk}}$, $\frac{V_{2tk}}{V_{3tk}}$,…., $\frac{V_{R-1tk}}{V_{Rtk}}$ and they are equal to $\frac{I'_{1tk}}{I'_{2tk}}$, $\frac{I'_{2tk}}{I'_{3tk}}$,…., $\frac{I'_{R-1tk}}{I'_{Rtk}}$, respectively.

For the second set of ratios, let's assume that we are again interested in same regions of r=1,2,…,R for same keywords of k=1,2,…,K then for each region r we have;

$\frac{V_{rt1}}{V_{rt2}}$, $\frac{V_{rt2}}{V_{rt3}}$, ……, $\frac{V_{rtK-1}}{V_{rtK}}$ and they are equal to $\frac{J'_{rt1}}{J'_{rt2}}$, $\frac{J'_{rt2}}{J'_{rt3}}$,….., $\frac{J'_{rtK-1}}{J'_{rtK}}$, respectively.

For each time t for a total of R*K unknown volume values we have 2*R*K – R – K equations. However, as all are in terms of proportions, it is impossible to derive numeric values of search volumes. On the other hand, for each time interval t, once V_{1t1} set to some number z_t then all other V_{rtk} can be expressed in multiples of z_t.

Then starting from t=1, for each pair of time t and t+1 we can estimate the ratios of $\frac{V_{rtk}}{V_{rt+1k}}$ for all regions and keywords. Both datasets discussed above are already comparable across time so actually $\frac{V_{rtk}}{V_{rt+1k}}$ is equal to $\frac{I'_{rtk}}{I'_{rt+1k}} = \frac{J'_{rtk}}{J'_{rt+1k}}$

Given that we have the numeric values for ratios $\frac{I'_{rtk}}{I'_{rt+1k}}$ and $\frac{J'_{rtk}}{J'_{rt+1k}}$ we can estimate ratios between z_t values. Then if one set z_1 = v, some arbitrary constant, all other z_2, z_3,…. z_{t-1}, z_t can be expressed in multiples of v. This actually means that all V_{rtk} can be expressed in multiples of v, where V_{111} is actually equal to v and V_{rtk} is now comparable across all of the three possible dimensions of regions, keywords and time.