Correlation in Cause of Death: Investigating Unconventional Relationships

M, CONDON, University of Colorado, Boulder
D. GORTHY, University of Colorado, Boulder
B. MELVIN, University of Colorado, Boulder

Conventional fatality studies focus mainly on demographic trends, disease prevalence and temporal trends. The thoroughness of the CDC's data collection allows other, unconventional factors to be considered. This project is comprehensive and will demonstrate every step of a data mining project.

General Terms: Data Mining, Preprocessing, Data Gathering, Trend Prediction, Data Cleaning

Additional Key Words and Phrases: Python, R, Hadoop, Hive

1 INTRODUCTION

The Center for Disease Control and Prevention (CDC) gathers mortality statistics for the United States on an annual basis. We will begin our project with the data gathered by the CDC in 2014. The CDC mines the data for common causes of death among large demographics and provides a basis for comparison to other years. We will begin by replicating the results given by the CDC to ensure that we establish an accurate baseline to gauge the validity of our results. We will answer questions such as: *Is there a correlation between race and cause of death?* We will then expand our analysis to estimate the income of an individual by factoring in race, age, marital status and place of death to answer questions like: *Is there a correlation between income and lifespan?* We will then deviate from traditional analysis and search for relationships between factors like day of death, place of death and manner of death. Finally, we will estimate the date of each death and pull in external financial, solar activity and accident datasets. Hopefully, we can answer questions such as: *Does the number of deaths spike in the event of a solar flare?* The focus of our project will be on questions that CDC and traditional secondary sources do not look to answer.

2 PREVIOUS WORK

2.1 Center for Disease Control

The work that has been done in previous years from CDC is trends of all deaths from previous years until 2016. There is an in depth document written by CDC that explains trends such as causes of death associated with a certain demographic. Another focus of the CDC is investigating the mortality rate of and change in mortality of infant and newborn deaths.

Primarily, the CDC focuses on tracking mortality causes and rates temporally for many demographics; for example, CDC first identified an increase in the number of deaths attributed to Alzheimer's in recent years.

2.2 Secondary Analysis

Organizations such as LiveScience take the data provided by the CDC and investigate questions not researched by the primary source. Examples of the work that LiveScience mined from the information were "Top 2 Deadly Diseases", "Top Killers", and "Respiratory Diseases & Attacks". LiveScience, among many other secondary researchers, focus on generating graphics to spread awareness of diseases and other preventable causes of death common in the United States.

3 PROPOSED WORK

So far, we have collected all the death data for 2014. The dataset provided by the CDC contains several redundant attributes that can be removed for faster processing time. For example, they provide three different attributes relating to cause of death -- because we are looking for the most specific cause of death, the other two less-specific attributes are redundant.

The main focus of our project is investigating correlation in unconventional relationships; consequently, we will gather daily stock data, major sporting occurrences and solar/space activity, among other supplementary datasets, for every day of 2014. Additional preprocessing must be performed on the CDC dataset to make this supplementary data applicable.

To pull in other datasets, it is imperative that we establish some method of estimating the date of the death. The CDC cannot disclose the exact date for privacy reasons; however, they do provide the month and day of the week of death. We have decided that equally distributing the deaths across all instances of a day of the week will be a good estimation. For example, if a month contains four Mondays and there are 1000 deaths on Mondays that month, we estimate that 250 deaths occurred on each Monday that month. This will have a dampening and potentially biased effect on our results, but it is a necessary approximation to make the CDC set compatible with other data.

It is not clear what significance threshold is appropriate for the results we obtain from our set. And due to the variation in each metric (i.e. S&P 500 movement vs. solar data), we will need to establish thresholds for each metric that we pull into the death dataset.

As explained in the literature survey, the bulk of studies on this set pertain to demographics and specific diseases or causes of death. Our goal is to deviate from this trend and look for correlations in factors not directly related to the CDC dataset. In order to validate our results, we will start with replicating the work explained in the literature survey and then moving on to analysis relationships between the CDC set and other daily metrics.

4 DATA SET

As described in the *Proposed Work* section, the CDC dataset contains the bulk of the factors surrounding the actual event of each death. The key attributes that will be our initial focus are years of education, age, sex, marital status, month and day of the week of death and manner of death. This base set will be used to verify our initial results and will be used to construct a baseline for later results. After initial analysis, we will estimate the date of each death in order to add in other attributes.

Using the CDC data set as a foundation, we will incorporate the Census Bureau's economic data, daily economic performance data such as the S&P 500 and the government bond rates, solar activity data, geomagnetic field data and daily accident data. This list is not comprehensive due to the availability of daily statistics across a variety of subjects.

5 EVALUATION METHODS

As explained in the *Proposed Work* section, establishing a baseline for results comparison is essential to justify any other results we generate. Luckily, the CDC's dataset has already been widely studied and provides many easily-understandable metrics for comparison. A trivial first result we will replicate is the leading causes of death for each combination of age, race and sex demographics.

Using the statistics that CDC and LiveScience published, we will expand our results to encapsulate irregularities in causes of death across demographics. Sources like the two mentioned have also studied the correlation between income level and average age of death – this will also be used to validate our initial results.

As we deviate from traditional studies, establishing our validity metrics will be essential to assert that any correlations we find are not only significant, but are not influenced by external factors. At this point in the project, we will need to research other unconventional studies to validate our results. However, it is not likely that we will find a study that investigates the relationships between solar flare activity and death rates; to address the lack of a concrete validation metric, we will use each other's results to verify our findings. By drawing conclusions independently, we will be more likely to catch any errors in our calculations by comparing results.

We must also address the issue of inconsistent validity thresholds for the external daily metrics. Daily S&P 500 performance will have a different scale and implication than daily solar activity. These two examples also draw attention to preconceived assumptions that we have of potential correlations – poor stock market performance should correlate with higher suicide rates, if any correlation exists. However, we have no basis for assuming a relationship between mortality statistics and solar activity. In some cases, common sense metrics must suffice when establishing results; in other cases, these metrics must be researched and based off of previous work.

6 TOOLS

6.1 Python. The simplicity and performance of Python 3 will allow us to perform fast preprocessing on the CDC dataset. The language supports several statistical, analytical and visualization plugins, making it an obvious starting point for the project.

6.2 R. Although R is normally used for larger projects, we will use it when seeking correlation between the CDC's demographic data and the supplementary external data pulled in from other sources. It may not be necessary to use R, but our goal is to gain experience in a breath of data mining and data visualization tools.

6.3 Git/Github. An industry standard in collaborative repository tools, Github will serve as our method of demonstrating the development process in a way that supports safe group development.

6.4 Slack. Considering the use of messaging tools used by industry, slack will be our main platform for communication.

6.5 SQL. Maintaining our goal of utilizing tools used in industry, we will use an SQL database to store our preprocessed data.

6.6 Hadoop/Amazon E3/Hive. If time allows, we will work with these three tools in order to gain experience in other data mining tools. Our dataset does not require the use of these tools, but using them on a smaller dataset will ensure that processing time does not inhibit our learning of the tools.

7 MILESTONES

7.1 Milestones Achieved So Far

7.1.1 Data cleaning. Redundant CDC death codes removed – only two sections were deemed repetitive. This did reduce the size of the dataset by 52 MB and made the preprocessing significantly faster.

7.1.2 Date estimation. A key issue with the dataset arose upon further investigation; the dataset contained the month and *day of the week* of every death. To pull in external data, date estimation was an absolute requirement to continue with this project. After research into variable estimation like dates, we chose to randomly assign a day attribute to every instance based on the number of days in that month. For example, there are five Friday's in January of 2015; every time a Friday in January is encountered, one of the five possible dates are randomly assigned to the instance.

7.1.3 Income estimation. Income is estimated by the U.S. Census Bureau's 2014 data per race for every death in the dataset.

7.1.4 Daily/Binary data addition. Unconventional data is the cornerstone of our analysis. The current list of added data is: federal holiday list, Friday the 13th list, full moon list, mass shooting data, S&5 500 data, and solar activity data. This list is not comprehensive and Python scripts in the repository allows for efficient addition of daily and binary data.

7.1.5 File size management. Because we are using Dropbox to share project scripts and dataset, Github's 100 MB file size limit required a reduction in the size of our dataset. We created one folder containing the entire split dataset and another holds smaller files for running trials of our computations on.

7.1.6 Establishing conventional significance metrics. An initial set of significance values for traditional areas of study is listed in the *Results* section. The significance values pertain to the relationships between traditionally studied values like race, level of education, cause of death, and month of death. These metrics will likely change, but are based off of previous studies on mortality.

7.1.7 Initial results validation. Our first step in confirming our results is the reproduce the results that the CDC's analysis shows. Our initial results are listed in the *Results* section. These computations were largely straightforward but confirmed that any preprocessing of the data did not get unintentionally modified.

7.2 Milestones Left to Achieve

7.2.1 Additional external datasets. While we have added several external datasets to our CDC sets, we believe that this list in not comprehensive. We are planning to add in weather data, tornado data, number of NFL games, average amount of daylight, and average U.S. temperature.

7.2.1 Establishing significance metrics. Significance metrics for unconventional data – any data that we added to the CDC set – will be more difficult to establish. Without purely estimating the significance values, we will need to consider other research to find grounds for our reasoning.

7.2.2 Results Generation. Once metrics have been established, we will perform an in-depth analysis to find correlations between a wide range of external factors measured daily. This will be the most thought-intensive portion of the project because the significance metrics are not studied and will need to be justified. We will begin with looking for significant correlations between one unconventional variable and one conventional variable. After this, we will expand our search to include multiple-variate correlations and unconventional-unconventional variable comparison.

7.2.3 Results summary. All analysis is complete; results will be depicted graphically and statistically. All results will be justified using previous results, established metrics, estimated metrics and/or statistical significance metrics.

7.2.4 *Results Heart*: Further our analysis by finding results relating to what location, race, etc... is more prominent to die from a certain heart-related disease/event.

8 RESULTS

Income Preprocessing and external data research took up more time than previously anticipated. We have been able to reproduce the several of the statistics stated by the CDC in 2014. This was a simple analysis, but it shows that our dataset behaves the same way as it did originally. After extensive preprocessing, making sure that the underlying data had not changed. Also, below is an estimated list of significance level from the limited amount of time that we have had to estimate significance levels.

Conventional Comparison:

Comparison Subjects	Estimated Low Significance	Estimated High Significance
Sex/Ethnicity – Cause of Death	5% from total average	15% from total average
Month of Death – Cause of Death	5% from overall average	10% from overall average

The following significance levels are largely speculative. Part of the remaining work involves researching previous analysis papers to find some basis for significance levels. These may be adjusted depending on the calculations of the dataset.

Unconventional Comparison:

Comparison Subjects	Estimated Low Significance	Estimated High Significance
Date Related – Cause of Death	10% from total average	15% from total average
Date Related – Place of Death	5% from overall average	15% from overall average
True/False Data – Cause of Death	5% from overall average	10% from overall average

CDC main results table:

Rank[1]	Cause of death (based on ICD–10)	Number	Percent of total deaths	2014 crude death rate	2014	Percent change 2013 to 2014	Ratio Male to female	Ratio Black[2] to white
...	All causes	2,626,418	100.0	823.7	724.6	–1.0	1.4	1.2
1	Diseases of heart (I00–I09,I11,I13,I20–I51)	614,348	23.4	192.7	167.0	–1.6	1.6	1.2
2	Malignant neoplasms (C00–C97)	591,700	22.5	185.6	161.2	–1.2	1.4	1.1
3	Chronic lower respiratory diseases (J40–J47)	147,101	5.6	46.1	40.5	–3.8	1.2	0.7
4	Accidents (unintentional injuries) (V01–X59,Y85–Y86)	135,928	5.2	42.6	40.5	2.8	2.0	0.8
5	Cerebrovascular diseases (I60–I69)	133,103	5.1	41.7	36.5	0.8	1.0	1.4
6	Alzheimer's disease (G30)	93,541	3.6	29.3	25.4	8.1	0.7	0.8
7	Diabetes mellitus (E10–E14)	76,488	2.9	24.0	20.9	–1.4	1.5	1.9
8	Influenza and pneumonia (J09–J18)	55,227	2.1	17.3	15.1	–5.0	1.3	1.1
9	Nephritis, nephrotic syndrome and nephrosis (N00–N07, N17–N19,N25–N27)	48,146	1.8	15.1	13.2	0.0	1.5	2.0
10	Intentional self-harm (suicide) (*U03,X60–X84,Y87.0)	42,826	1.6	13.4	13.0	3.2	3.6	0.4
11	Septicemia (A40–A41)	38,940	1.5	12.2	10.7	0.0	1.2	1.8
12	Chronic liver disease and cirrhosis (K70,K73–K74)	38,170	1.5	12.0	10.4	2.0	2.0	0.6
13	Essential hypertension and hypertensive renal disease (I10,I12,I15)	30,221	1.2	9.5	8.2	–3.5	1.1	2.1
14	Parkinson's disease (G20–G21)	26,150	1.0	8.2	7.4	1.4	2.3	0.5
15	Pneumonitis due to solids and liquids (J69)	18,792	0.7	5.9	5.1	–1.9	1.9	1.0
...	All other causes (residual)	535,737	20.4	168.0	...	...	...	...

Our results (mimicked results table):

Rank	Cause of Death	Number	Percent of Total Deaths
1	Diseases of heart	614,346	23.4
2	Alzheimer's Disease	93,537	3.6
3	Suicide (Self-Harm)	42,820	1.8
4	Parkinson's Disease	26,147	1.0

REFERENCES

[1] "Death in the United States." Kaggle. Center for Disease Control and Prevention, 2016. Web. 2 Mar. 2017.
[2] DeNavas-Walt, Carmen, and Bernadette D. Proctor. "Income and Poverty in the United States: 2013." Census.gov. United States Census Bureau, Sept. 2014. Web. 2 Mar. 2017.
[3] Geggel, Laura . "The Odds of Dying." LiveScience. N.p., 9 Feb. 2016. Web. 2 Mar. 2017.
[4] "Historical Treasury Rates." U.S. Department of Treasury, n.d. Web. 2 Mar. 2017.
[5] "National Solar Radiation Data Base." National Renewable Energy Laboratory, n.d. Web. 2 Mar. 2017. <https://maps.nrel.gov/nsrdb-viewer/>.
[6] "S&P 500 (^GSPC)." Yahoo Finance, n.d. Web. 2 Mar. 2017. <https://finance.yahoo.com/quote/%5EGSPC/history?p=%5EGSPC>.