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

ACM Reference format:

Ben Trovato, G.K.M. Tobin, Lars Thørväld, Lawrence P. Leipuner, Sean Fogarty, Charles Palmer, John Smith, and Julius P. Kumquat. 1997. SIG Paper in word Format. *ACM J. Comput. Cult. Herit*. 9, 4, Article 39 (March 2010),4 pages.
DOI: 10.1145/1234

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, interest rate data and solar/space activity, among other supplementary datasets, for every day of 2014. Additional preprocessing must be performed on the CDC dataset in order to make this supplementary data applicable.

In order 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 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 similar to 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; in order 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 Established results validation. Replicating the results of the CDC's studies is imperative to establish a baseline for future results.

7.2 Data cleaning. Remove redundant CDC attributes such as less-specific death codes. Cleaning redundant data will yield faster results generation in more rigorous analysis.

7.3 Attribute estimation. The two key attributes we must estimate are income level – based on age, race and sex – and date of death – using method described in *Proposed Work*.

7.4 Data addition. Once the date of each death has been established, we will add in supplementary unconventional data. Although a comprehensive list has not been established, we have found data for 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.

7.5 Conventional supplementary attribute analysis.

7.5.1 Establishing significance metrics. Previous studies on economic conditions related to death will be a useful metric for validating our results. Results may vary for study to study, metrics will involve estimation.

7.5.2 Results Generation. Once metrics have been established, we will perform an in-depth analysis in an attempt to find correlations between economic status and mortality attributes.

7.6 Unconventional supplementary attribute analysis.

7.6.1 Establishing significance metrics. A lack of previous studies will force our significance metrics to be based entirely on estimation.

7.6.2 Results Generation. Once metrics have been established, we will perform an in-depth analysis in an attempt 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.

7.7 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.

8 SUMMARY OF PEER REVIEW SESSION

We initially established that the size and breadth of the CDC dataset ensured that we have a sample large enough to justify basic correlations. The breadth of the attributes and the availability of initial comparison results ensures that our methods will be, initially, easily validated.

The largest concern of the session was the estimation of the date of each death. We are unable to obtain the date of death from the CDC, so estimation is our only option if we want to add in other daily datasets. No method of date estimation is perfect, so we decided to evenly distribute the deaths across each death (described in detail in *Proposed Work*). This will influence our results and must be disclosed when presenting the results, but we concluded that it will dampen any spike in deaths on a particular date. To account for this, we decided to adjust our validation sensitivity.

Finally, we discussed the tools we will be using in this project. While gaining experience in Hadoop, Amazon E3 and Hive would be beneficial, the bulk of our work will be done in Python and R in order to obtain results. If any time is left, we will experiment with the tools and perform further investigation.

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>.