A Statistical Approach To Correlating Environmental and Demographic Factors to Cancer Incidences Across U.S. Counties

Shivakumar
Student
The Harker School
San Jose, CA, USA
19kaushiks@students.harker.org

***Abstract*—This study seeks to identify environmental and demographic factors contributing to cancer incidences, approximately 80% of which are classified as sporadic, across the United States. All data utilized was public, containing a variety of statistics by U.S. county. Through the use of statistical techniques, an average of 20 demographic and chemical factors were found to be strongly associated with increased or decreased cancer rates, across the four different cancer types examined. Even though not all of these factors may be causes, they have strong correlations with the cancer incidence rate, and thus deeper examination of them can reveal the true risk factors. Thus, this study represents a novel way of bringing to light various previously unexamined factors that could be causing cancer incidences and therefore could represent a major step forward in preventative medicine.**

Keywords—Sporadic cancer, Environmental causes of cancer, Pearson linear correlation coefficient, Linear regression

I. Introduction

A. Sporadic Cancer Incidences

Cancer, the second leading cause of death in the U.S., claims over half a million lives every year. The disease is generally classified into two categories: hereditary and sporadic. Sporadic cancer, unlike hereditary cancer, is *primarily* caused by external factors and accounts for approximately 80% of all cancer incidences [1]. This amounts to approximately 1.4 million new cases annually in the U.S. alone [2]. While much research is being conducted on genetic ties to cancer, and direct testing of carcinogens, there have not been many large-scale studies that examine environmental causes of cancer. This study focuses on identifying those little-examined environmental causes for cancer.

B. Cancer Incidence Variation

Sporadic cancer incidence primarily depends upon the environment, so it is apparent that cancer incidence varies by geographic location. For instance, the leading type of cancer varies from county to county in the U.S, indicating that there is a strong geographic factor in cancer incidences. Additionally, using data from the Centers for Disease Control and Prevention (CDC) data, it is possible to show that lung cancer incidences vary significantly by geographic location as illustrated in Fig. 1. There are significantly more counties with greater than 0.150% lung cancer incidences on the east coast than the west. Only 2 out of 79 counties with lung cancer high incidence rates lie in the western U.S. Based on this preliminary finding, it is hypothesized for this study that cancer incidences are linearly correlated with factors related to the demography and environment of each county of the U.S.

Fig. 1. Distribution of counties with high lung cancer incidence across the U.S. Each color of the pin represents a different county. As can be seen, there is a high amount of variation based on locality (the east coast has a much higher incidence rate), which motivates this study.

C. Prior Work

No prior work closely matches the research presented in this paper. Some studies, such as one by Simeonov and Himmelstein [3], examine the effect of one particular factor on cancer incidence rates, specifically elevation on lung cancer. Another study, by O'Connor et al. [4], focuses on external factors that control the differences in cancer rates between income levels. Both of the aforementioned studies ask narrow and specific questions about cancer incidences and are ultimately different in scope from the work that is presented in this paper. No other previous research has involved broad-scale data analysis on hundreds of factors like has been done in this study, to determine several environmental risk factors for sporadic cancers.

II. Methods

A. Data

Two types of data were used to analyze cancer incidences by region. One of these datasets, obtained from *County Health Rankings & Roadmaps*, consisted of 69 demographic factors (race, gender, poverty, healthcare costs, median income) for one year, listed by U.S. county (there are 3,141 counties) [5]. The other, provided by the *Environmental Protection Agency* (EPA) consisted of the emission levels of 274 chemicals, again by U.S. county, over a year's time [6]. The data for the cancer incidences per county was obtained from the governmental website, *State Cancer Profiles*, which was created by the National Cancer Institute (NCI) and CDC [7]. This data was presented as an average over a set of five consecutive years.

This study was performed for 3 different types of cancers (lung, colorectal, and pancreatic — the ones with the highest mortality rates), and all cancers in aggregate.

B. Demographic Data Analysis

1) Data Cleaning

To generate a table which contained both the cancer incidence data and the demographic data, SQL was used perform a left outer join operation on the two datasets, using the state and county names as the key for the operation. This formatted all of the data as required, with each row representing a county and each column representing a feature. A new column, effectively an $n \times 1$ label matrix, was created to record the rate of cancer incidences per ten thousand people, as this is a much more meaningful metric than the raw count of cancer incidences. This is due to the fact that, if using just the raw counts of cancer incidences, a county with a population of 1,000 and an annual rate of cancer incidences of 100 would appear to be healthier than a county with 10,000 individuals with a cancer incidence rate of 200. Thus, the rate of cancer incidences, in units of incidences per ten thousand people (PTTP), is a much stronger indicator of cancer risk — comparing the two counties would then indicate that the second county is healthier, as expected.

After this data was generated, significant data cleaning was performed. All counties with population below 15,000 individuals were filtered out, because a change in one incidence could affect the rate of incidences significantly. Thus, these counties added significant noise to the data. Furthermore, the counties which had fewer than 3 reported incidences did not report the exact number of incidences, so these counties were completely eliminated in the filtered data, again to avoid noise. After filtering, approximately 1,000 out of the 3,141 counties remained, and this number varied slightly for different types of cancer. Table I shows the exact numbers of counties remaining for the different types of cancers. An illustrative example of the usefulness of the filtering is shown in Fig. 2. As can be seen, the filtering causes the trends in the data to appear much more clearly, resulting in a much cleaner dataset where trends can be observed much more easily.

2) Correlation Analysis

The following steps were conducted for each type of cancer. Out of the 69 demographic features, only those with $|p| > 0.3$, where p represents the Pearson linear correlation coefficient are chosen, to select features that appear to have strong correlations with cancer incidence rates. Afterwards, the values for each of the features are normalized to their average, as described in Equation 1, in order to facilitate comparison. The number of counties is represented by n and the number of features is represented by m in all equations. *Val (County, Feature)* represents the original data matrix, and *NormVal (County, Feature)* represents the normalized feature matrix.

TABLE I. Filtered County Counts

Type	*Lung*	*Colorectal*	*Pancreatic*	*All*
County Count	*996*	*996*	*1025*	1029

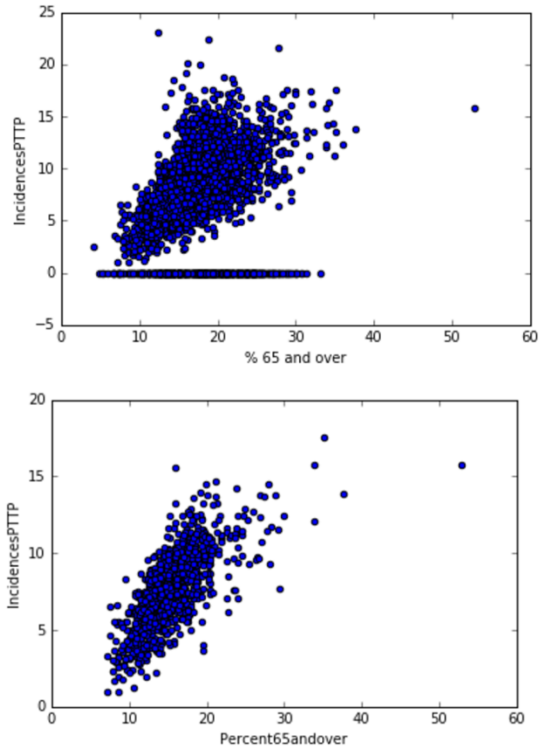

Fig. 2. Illustration of cancer incidence rate as a function of the percent of people in a county that are 65+ years old, before (above) and after (below) cleaning the study.

$$NormVal\ (County,\ Feature) = \frac{Val\ (County,\ Feature) \times n}{\sum_{i=1}^{n} Val\ (i,\ Feature)} \quad (1)$$

With *NormVal (County, Feature)* as the feature matrix, and *CancerIncidences (County)*, representing the incidence count per ten thousand individuals, as the label matrix, the data was fed into a linear regression model. This was done not to build a predictive model, but to rank the various factors in terms of relative importance, when all taken into account. After the linear regression model was built, the features with p values greater than 0.05 were removed, as they were not strong indicators of increasing or decreasing cancer rates. The remaining features were then ranked according to their coefficients.

3) Interpretation

The results were interpreted as such: features with the highest coefficients in the trained model would be the strongest candidates for increasing cancer rates, and those with negative coefficients, when increased, would cause a reduction in the amount of cancer incidences. To perform a sanity check on the model, it was compared with an "average-predicting model," one that would always predict the average number of cancer incidences in a county. Additionally, a model trained on three-quarters of the training data and tested on the remaining portion was also compared to the average-predicting model.

An additional point worth noting is that this approach combines analysis of correlation and causation. The initial filtering by Pearson correlation coefficient extracts features that show strong promise — that is, those that have strong relationships with cancer incidence rates. However, when run through the multiple linear regression model, the signs of expected coefficients of these features were sometimes reversed, because the OLS (Ordinary Least Squares) regression accounts for the interdependence of the features themselves, known as collinearity. This phase of the analysis focused more on the contribution of each feature to cancer incidences, rather than each feature's correlation to other features that caused cancer incidences.

In essence, the first half of the demographic data analysis emphasized correlations, and the second half focused more on causations, by taking into account feature collinearity. Thus, this study presents a hybrid approach, first selecting the "interesting" demographic factors, and then ranking them on their causal significance.

C. Chemical Data Analysis

1) Data Cleaning

Chemical data, as compared to the demographic data, was much sparser. While a diverse array of industries producing and emitting various chemicals exists nationally, each county does not contain many industries. Therefore, each county may only have data for a few of the 274 chemicals registered in the EPA dataset. Because of this sparsity, the same type of analysis as done for the demographic data (Pearson correlation coefficient filtering combined with normalization and linear regression) could not be performed; instead, a different methodology was adopted.

2) Correlation Analysis

The chemical emissions data was batched into two categories, 0 — not contaminated/no data (signifying no emission and thus contamination) and 1 — contaminated (having a level three times the average). Then, an algorithm was used to determine which of the chemicals were statistically significant with respect to increasing cancer incidences. *CancerIncidence(0, i)* returns the cancer incidence count per ten thousand people in the i^{th} county without contamination of a chemical, and *CancerIncidence(1, i)* returns the same for the i^{th} county with contamination.

$$avgIncidencesOnes = \frac{\sum_{n=1}^{numOnes} CancerIncidence(1, n)}{numOnes} \quad (2)$$

$$avgIncidencesZeros = \frac{\sum_{n=1}^{numZeros} CancerIncidence(0, n)}{numZeros} \quad (3)$$

$$\partial = avgIncidencesOnes - avgIncidencesZeros \quad (4)$$

$$stDevZeros = \sqrt{\frac{\sum_{n=1}^{numZeros} (CancerIncidence(0, n) - avgIncidencesZeros)^2}{numZeros}} \quad (5)$$

$$numStdDeviations = \frac{\partial}{stDevZeros} \quad (6)$$

Equations (2) through (6) were used to find the number of standard deviations by which cancer incidences increased in the presence of a chemical. All of the chemicals with greater than two standard deviations were selected as statistically significant, and in practice, most of the identified chemicals had greater than three standard deviations of statistical significance, thus eliminating the possibility that just random chance led to chemicals' identification. adopted.

3) Interpretation

After determining the statistically significant chemicals, the percentage increase in cancer incidence rates that contamination was accompanied by was plotted for each chemical, as a metric for determining the magnitude of impact of each chemical on cancer incidence rates.

III. Results And Discussion

A. Demographic Data Analysis Results

Fig. 3 illustrates the results for the demographic data analysis, showing the relative importances of the factors as determined by the study. These factors range from the left hand side, where they have strong positive associations with cancer incidences, and to the right hand side, where they have strong negative associations with cancer incidences (they contribute to a lessening of cancer incidences). An average of 10 factors were identified as statistically significant across the cancers.

As expected, age consistently ranks as one of the most significant factors of increased cancer incidence rates. Counties with greater amounts of Non-Hispanic Caucasian individuals also have greater cancer incidence rates, possibly implicating race as a significant contributing factor to increased cancer risk. Furthermore, for pancreatic cancer, diabetes ranks as one of the primary factors, which is expected, as there are studies that have shown that diabetes can cause pancreatic cancer [8].

Across the four cancers, a lack of education presents itself as being associated with a high cancer incidence rate. While a lack of education itself does not cause cancer, it could be that unhealthy habits or lack of awareness, caused by not being educated well, could lead to a higher risk. Additionally, the

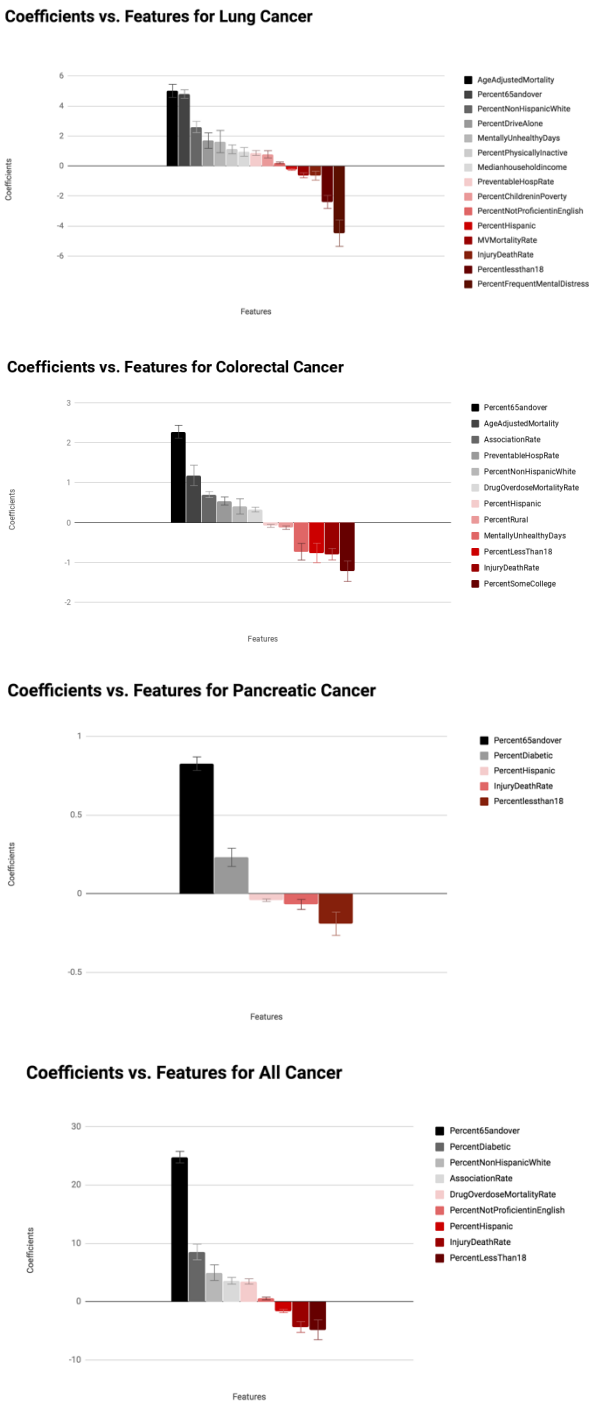

Fig. 3. Illustration of the relative importances of the demographic factors as per the demographic data analysis. There are four graphs, three for the different types of cancers, and the fourth for the aggregate study. The factors with the most positive coefficients are indicated to have strong causatory relationships with cancer incidences, and negative coefficients indicate a inhibitory relationship with cancer incidences.

percentage of children in poverty ranks as a factor for lung cancer. Again, while poverty does not cause lung cancer, poor living conditions and exposure to toxic substances certainly could. Revealing these problems, as this study is doing, can lead to legislative action on the part of each county to address these problems to minimize people's cancer incidence risk. There are also certainly some factors that are also unlikely to lead to true causes, such as the percentage of people who drive alone, but the majority of factors are either causal or can easily lead to discovery of actual risk factors.

Furthermore, Fig. 4 illustrates the performances of the trained linear regression model compared to a model that always predicts the average cancer incidences across counties. This was used as validation to ensure that the model did pick up on the features and obtain the right coefficients. As is evident, the linear regression model significantly outperformed the average predictor, achieving approximately half the percentage error across the different types of cancers. Additionally, an experiment was run where 5 random 75%/25% train/test splits were generated for each type of cancer, and the percentage errors were averaged. These average percentage errors and their standard error of the mean are reported in Table II, along with the results of the two experiments compared in Fig. 4. The performance of the model with only three-quarters of the training data is very comparable to the model trained on all the data. This illustrates that the model is easily able to generalize to counties that it has not seen before, and it serves as further validation for the approach taken in this study, as it shows the robustness of the multiple linear regressor.

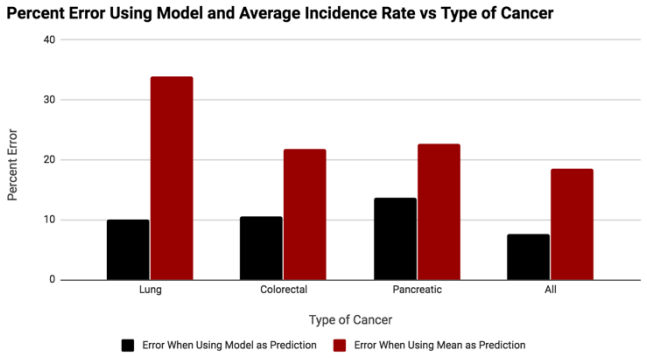

Fig. 4. Performance of linear regression model as compared to a fake model that always predicts the average cancer incidence rate. As can be seen, the trained linear regression model obtained much lower percentage errors compared to the average-predicting model, illustrating that it picked up on the correct features and that the coefficients provided by it are indeed accurate.

TABLE II. PERCENTAGE ERRORS FOR DIFFERENT EXPERIMENTS

Type	Lung	Colorectal	Pancreatic	All
Error With All Training Data Used	*10.1%*	*10.6%*	*13.6%*	*7.7%*
Error With ¾ of Training Data Used	*12.8 ± 0.5%*	*11.0 ± 0.2%*	*13.6 ± 0.3%*	8.0 ± *0.2%*
Error When Using Mean As Prediction	*34.0%*	*21.9%*	*22.6%*	*18.5%*

B. *Chemical Data Analysis Results*

Fig. 5 shows the chemical data analysis results, which illustrate the ranking of statistically significant chemicals by the percentage increase in cancer incidence that they are accompanied by. As can be seen, several chemicals were identified for each type of cancer, and the most number of chemicals identified was for all cancers, highlighting the benefits that can arise from using a combination of all cancers in identifying carcinogens. On average, 10 chemicals were marked as statistically significant, across the four types of cancers studied.

2-Chloroacetophenone is represented by the yellow color in Fig. 6 and is the statistically significant chemical with the most contaminated counties. It is an extremely toxic compound used in tear gas, and the fact that it, and many other harmful chemicals are produced and released in high amounts into the atmosphere should be of concern [9]. Methyl Isocyanate is also a chemical that emerges as statistically significant in two of the different cancers, and it is used in the manufacture of pesticides [10]. It is represented by the orange color in Fig. 6, and is present in high amounts in two counties, including San Diego, California. This opens a whole new range of possibilities of future topics to study using a similar type of analysis — the effects of different types of and levels of pesticides on cancer incidence rates in U.S. counties. Lastly, *p*-Phenylenediamine, represented by the green color in Fig. 6, is a chemical used in dyes and, which is known to cause irritation in humans but is not classified in terms of carcinogenicity [11]. The chemical *p*-Phenylenediamine is also used in the manufacture of plastics like kevlar [12]. Orange, Texas, one of the counties marked as having high levels of this pollutant, is home to polymer factories, which could very well be increasing cancer incidences in the county [13]. These three chemicals (2-Chloroacetophenone, Methyl Isocyanate, p-Phenylenediamine) are illustrative examples of the goals of this research — to bring to light these emissions that are highly likely to contribute to increased cancer rates in counties.

To validate this analysis, a map showing counties which were contaminated with statistically significant chemicals for lung cancer was generated (Fig. 6). It shows a striking similarity with Fig. 1, which also shows a large east coast bias, but for lung cancer incidence. Given that chemicals in the dataset are produced all across the U.S., this map serves as validation for the chemical data analysis, which has selected chemicals that are predominantly on the east coast, and which align with high lung cancer incidence rates.

IV. Conclusions

A. *Primary Conclusions*

Overall, using by-county data has a variety of distinct advantages. This method makes it possible to uncover not only factors that are direct, but also indirect causes of cancer, which cannot be done through laboratory testing. Moreover, this study has enabled the identification of characteristics of counties that have high cancer incidence rates.

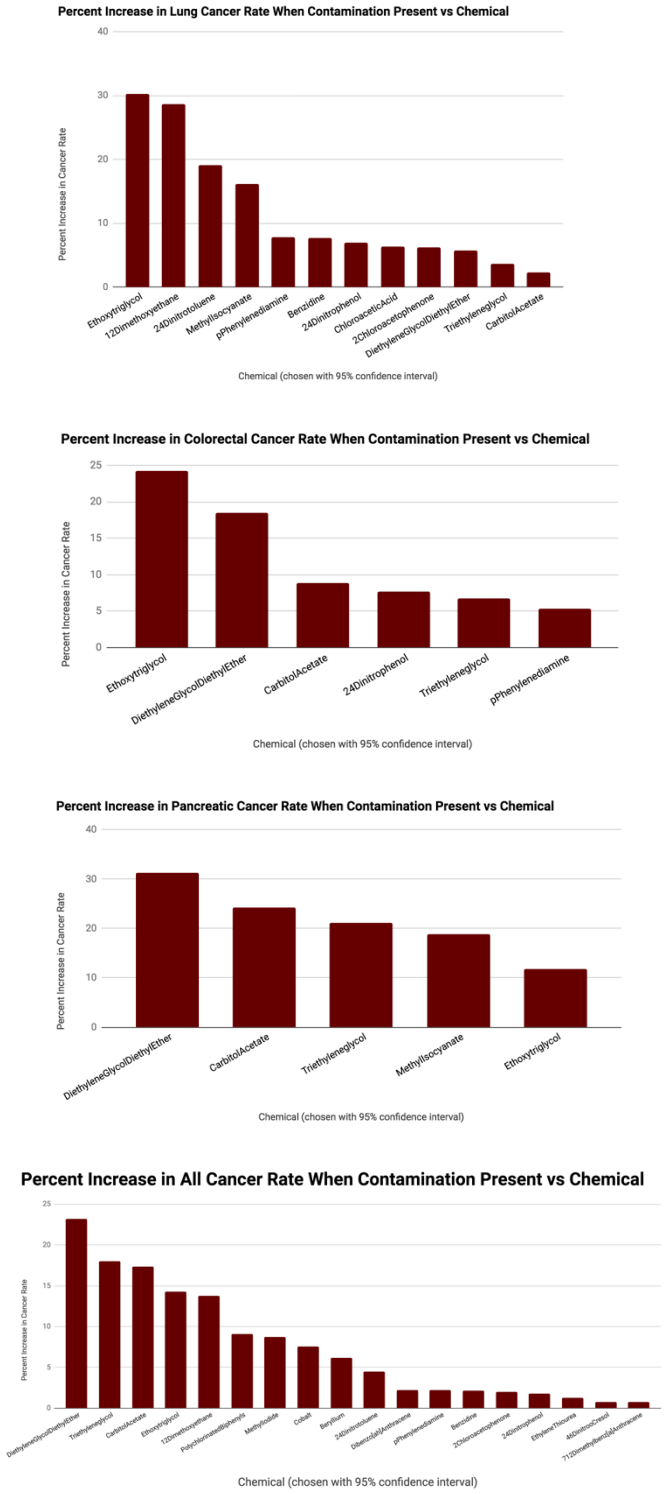

Fig. 5. The chemicals identified as statistically significant through the chemical data analysis are ranked in order of the percentage increase in cancer incidences that they are accompanied by, for each of the four cancer types.

This type of high-level analysis is extremely powerful since it enables counties to identify their unique problems and take action. As physical inactivity emerges as a risk factor for lung cancer through this study, a county that has especially high levels of inactivity, such as Roane County, Tennessee, can invest in building public exercise centers, in order to reduce the average physical inactivity of the population. Additionally, other measures, such as improving physical education in schools, and improving sports programs for both children and adults, can help combat physical inactivity.

On the other hand, counties can also work on finding techniques to reduce emissions of toxic chemicals that have been identified as statistically significant. For example, Colbert County, Alabama produces high levels of 2-Chloroacetophenone, which is a statistically significant chemical for lung cancer. Thus, it may invest in finding alternate methods of production that do not release this toxic chemical as a by-product.

In essence, this study brings to light not only direct risk factors (eg: diabetes for pancreatic cancer), but also previously unconsidered risk factors, by unveiling other factors (eg: poverty) that are highly associated with them. Being able to identify, through this large-scale data analysis, what actually causes increased cancer incidence rates, is extremely valuable — it can lead to officials intervening to improve the health of the public.

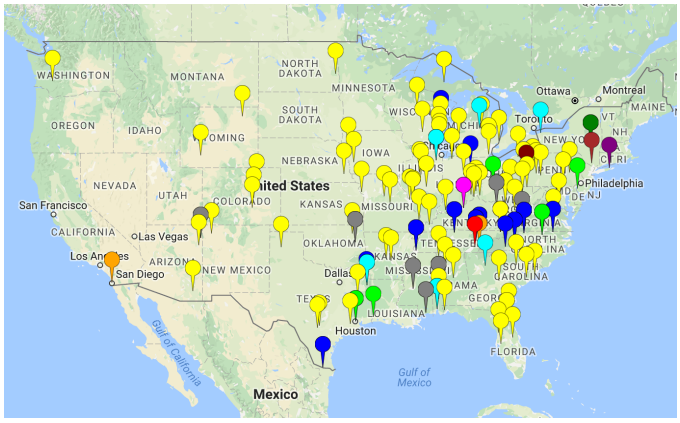

Fig. 6. A map of counties that have contaminated levels of chemicals that were identified as statistically significant for lung cancer through chemical data analysis. Each color in this map represents a different statistically significant chemical.

B. *Future Research*

Future research that can be conducted involves locating industries where risk factors identified by this study are high and assessing the geographic relationship with high cancer incidence areas in a quantitative manner. Additionally, combining granular cancer data (eg: neighborhood) with local environmental data to build especially accurate models would yield an interesting study as well. It could also utilize other types of datasets, such as air quality, pesticide use, and healthcare quality, which may provide additional insights into causes of increased cancer rates in certain counties.

Additionally, tuning methods for determining strengths of correlations and making the analysis process more automated can lead to this study's utilization in solving many related problems. The thresholds used can be tuned to be more restrictive or sensitive; furthermore, only Pearson correlation coefficient thresholding and only multiple linear regression can all be explored, to emphasize the correlation or causation aspects of the demographic data analysis more than the other — as of now, the study weights both aspects equally.

Overall, the results of this study are extremely promising, and they lay the groundwork for many potential future investigations into environmental causes of cancer.

V. Acknowledgements

I would like to thank Dr. Ramakrishnan Srikant for helpful discussions pertaining to statistical data analysis techniques.

References

[1] Northwell.edu. (2018). *Hereditary Cancer Genetics Program at SIUH - sporadic-vs-hereditary-cancer | Northwell Health*. [online] Available at: https://www.northwell.edu/find-care/locations/hereditary-cancer-genetics-program-staten-island-university-hospital/sporadic-vs-hereditary-cancer [Accessed 1 Sep. 2018].

[2] National Cancer Institute. (2018). *Cancer Statistics*. [online] Available at: https://www.cancer.gov/about-cancer/understanding/statistics [Accessed 2 Sep. 2018].

[3] Simeonov, K. and Himmelstein, D. (2015). Lung cancer incidence decreases with elevation: evidence for oxygen as an inhaled carcinogen. *PeerJ*, 3, p.e705.

[4] O'Connor, J., Sedghi, T., Dhodapkar, M., Kane, M. and Gross, C. (2018). Factors Associated With Cancer Disparities Among Low-, Medium-, and High-Income US Counties. *JAMA Network Open*, 1(6), p.e183146.

[5] County Health Rankings & Roadmaps. (2018). *Explore Health Rankings | Rankings Data & Documentation*. [online] Available at: http://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation [Accessed 2 Sep. 2018].

[6] Aqs.epa.gov. (2018). *Download Files | AirData | US EPA*. [online] Available at: https://aqs.epa.gov/aqsweb/airdata/download_files.html [Accessed 2 Sep. 2018].

[7] Statecancerprofiles.cancer.gov. (2018). *State Cancer Profiles*. [online] Available at: https://statecancerprofiles.cancer.gov/incidencerates/index.html [Accessed 2 Sep. 2018].

[8] Li, D. (2011). Diabetes and pancreatic cancer. *Molecular Carcinogenesis*, 51(1), pp.64-74.

[9] Epa.gov. (2018). *2-Chloroacetophenone*. [online] Available at: https://www.epa.gov/sites/production/files/2016-09/documents/2-chloroacetophenone.pdf [Accessed 2 Sep. 2018].

[10] Pubchem.ncbi.nlm.nih.gov. (2018). *Methyl isocyanate*. [online] Available at: https://pubchem.ncbi.nlm.nih.gov/compound/methyl_isocyanate [Accessed 2 Sep. 2018].

[11] Pubchem.ncbi.nlm.nih.gov. (2018). *P-Phenylenediamine*. [online] Available at: https://pubchem.ncbi.nlm.nih.gov/compound/p-Phenylenediamine [Accessed 2 Sep. 2018].

[12] Explain that Stuff. (2018). *How1does Kevlar work? | Why is Kevlar so strong?*. [online] Available at: https://www.explainthatstuff.com/kevlar.html [Accessed 2 Sep. 2018].

[13] Alloy Polymers. (2018). *Home*. [online] Available at: https://www.alloypolymers.com/ [Accessed 2 Sep. 2018].