Geospatial analytics and forecasting of Industry-related Cancer Risks

1. Introduction-Motivation

In 2020, nearly 1,800,000 new cancer cases and more than 600,000 cancer deaths are projected to occur in the United States[1], while $173 billion will be spent by the government on cancer treatments[2]. Cancer risk factors include biological factors such as genetic predisposition, lifestyle factors such as tobacco or alcohol, and factors related to work and living environments[25]. Manufacturing processes have undoubtedly been linked with adverse health risks[3], influencing the number of cancer patients, as it was proven for coal mining and lung cancer[24].

2. Problem Definition

Although multiple studies have investigated particular industries' impact on cancer risks [4–6], there is still limited knowledge about the local impact of industries as well as a need to identify and visualize the most influential industries in the county level. We propose a higher resolution analysis of industrial environmental impacts related to various types of cancer incidents at a county level throughout the USA. Big data analytics, predictive models and interactive visualizations are combined to find patterns between industry-related environmental indicators, county level demographics and cancer data. The goal of this project is to analyze, predict and visualize the cancer incidence rate of a county due to a particular industrial expansion (e.g. building a factory). We aim to build an intuitive interface to support informed decisions and predict health implications.

3. Survey

One of the oldest and most effective tools for determining the cause of disease is the study of its transmission in space and time, known as Spatial Epidemiology. Early pioneers[7] demonstrated that by plotting the location and incidence of disease on a map, it was possible to discover its origin and possible means of transmission. This field was advanced using improved mapping techniques and a greater understanding of disease vectors[8]. Growing availability of compute power, data, and Geographic Information Systems (GIS) enable researchers to identify geographic coincidence of disease clusters and pollution sources with high fidelity[9].

On the grounds that machine learning and spatial data mining (SDM) advancements stimulate the field of epidemiology, Bellinger et al.[10] have explored alternative methods to make predictions, find patterns and extract information. These techniques enable knowledge extraction from spatial big data[11]. Common algorithms for location prediction are: Linear regression model, Bayesian classifiers, neural networks and decision trees. But as the nature of SDM is different, these algorithms are not fully able to solve the problems that require Spatial AutoRegressive (SAR) and Markov-hybrid models[12] based on Bayesian Classifier algorithms[13]. Cluster analysis has also been an accurate technique for studying disease incidence in small regions[14]. A relevant study correlated lung cancer mortality rates with median salary in Florida[15] confirming the analytical capabilities via regression models and geographic distribution measurements as well as user-friendly data visualization through disease mapping and clustering. Another study looked into the link between residence and different types of cancer in Massachusetts using the same clustering approach[16].

4. Proposed Method

4.1. Innovation

So far, there are many tools that allow people to visualize information either about cancer, local industries or environmental factors. However, these tools are for siloed applications[17], thus making

it hard to get a global, more insightful view that can only be obtained by a congregation of these existing tools. Furthermore, existing research that does try to congregate cancer data with geospatial features[18] has been more focused on post-outbreak situations instead of trying to identify and predict health risks. At the core of our approach, we want to understand the global picture by aggregating datasets from our three aforementioned areas of interest, identify patterns through Machine Learning based analysis, after which we want to present our results through user-friendly software. The latter will provide information at the county level granularity on relationships between industries and a wide range of cancer types. Our key innovations are summarized as follows:

- Higher resolution analysis of cancer-related industrial impacts at the county level to discover the most or least contributing industries to cancer risks and validating that analysis by our Machine Learning models;
- Use clustering to find combinations of high cancer-rate industries for particular cancer types;
- Provide citizens and lawmakers with an intuitive user interface that allows them to visualize how changes in local industries would affect local incidence rates. This would allow them to make more informed decisions. For example, visualizing how adding an automobile factory in a county would impact local cancer incidence rates.

4.2. Description of approaches

4.2.1 Data Collection

For this project we have collected data from three primary sources First, Cancer Incidence Data (CID) for various forms of cancer were collected from the NIH's State Cancer Profiles website[21] using a Python web scraping script. Second, County Business Patterns (CBP)[22], which include data about the size of each industry in each locality was collected using an API exposed by the US Census Bureau. Finally, Industry Impact Data (IID), a collection of 24 industry indicators, ranging from job creation to CO2 emissions, was collected from USEEIO model[20] through an API exposed by the EPA. To complement the CBP data detailed above, we have collected IID values for all 24 indicators for a wide range of industries.

4.2.2 Data Cleaning and Integration

Each of the three data sets described above were gathered from government sources, and there was very little cleaning required beyond basic deduplication and filling missing values. Integration of this data set, however, was a much more arduous process. We had to unify several unique identifiers, such as state names, NAICS codes, and county FIPS IDs, to ensure consistent nomenclature. Our end data product is a pair of tabular data sets for the industry indicators (input features) and the cancer incidence rates (output labels). Each table contains a single row for each county in the U.S. Each row in the IID contains a column for each industry indicator for each industry classification. Each row in the CID contains a column containing annual cancer incidence rate for each cancer type. After we assemble this data, we set it up in a uniform manner for training and prediction purposes. To bring county-level resolution in our dataset, a calibration factor - the annual payroll of each industry in counties divided by the national annual compensation of the industry - is applied to the 24 industry indicators. When training a model, the input data point is the environmental impacts from all industries in a given county, which is being mapped to the output data point of average annual cancer incidents in the same county.. The full process for integrating the data is given in Figure 1. Twenty-four environmental impacts are tracked per industry and 70 higher-level industries per county.

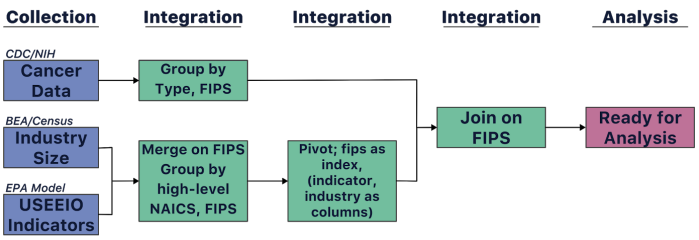

Figure 1: Data cleaning and integration

4.2.3 Analysis and algorithms

Two machine learning techniques are applied on the dataset: regression and clustering. We use regression to complete two goals: the first one is that we want to build a reliable regression model that is able to predict the cancer incidence rate based on different factors. We use this model for two main goals: the first one is to validate mathematically the basic analysis that we have done for each county to determine which industry affects most cancer incidence rate thanks to the external information that we have. We do this once we have an accurate prediction model, and we extract the features that have contributed the most to producing the prediction. It would make sense that these features would be the industries that contribute most to the cancer incidence rate. The second part where we use our regression model is when we want to know how the incidence rate in the county changes if, for example, we add a new factory in that county. Given that our data is set in a feature space where the number of samples to the number of features is close to a 100:1 ratio, we think that a Support Vector Regression will yield good results[23]. Outside the realm of SVMs and in the realm of classical regression, we test Ridge Regression which is similar to an Ordinary Least Squares approach, but uses L2 regularization which we have interest in because this can penalize certain weights of the input features from changing drastically during the training process. This is of interest because we could want robust weights for counties with values for certain features orders of magnitude larger than the same features in other counties. Another would be Lasso-type regression where its structure is very similar to Ridge Regression but the affinity function is normalized by the number of samples. Elastic Net is also of interest due to its mixture of optimizing on disparate data as Lasso does and weight-penalization from the Ridge Regression.

To supplement our findings, we used clustering techniques for classification on our dataset to try to find a general prediction of low, mid and high number of incidents rates per county, depending on the amount of each impact factor in that county. The knowledge thus gained would allow legislators to make better informed business decisions, when it comes to their effect on the health of their communities. To do this we found new features, using Linear Discriminant Analysis (LDA), that reduced the dimensions of our 24 environmental factors. These new features were then used to predict if the number of incident rates of a specific cancer type is within, above or lower than 1 standard deviation of the mean. We will also use PCA to find the most important factors contributing to each cancer type.

4.2.4 Visualizations

To explore the geospatial trends in the data, and the relationships between industry indicators and cancers rates, we have provided an interactive dashboard built on HTML, Javascript and D3 (available at: http://142.93.73.45:8000/). A national-scale, county-resolution choropleth is a natural way to begin looking for visual explanations of the geospatial distribution of various cancers. Users can toggle between cancer data and industry data using a set of controls presented above the map (Appendix A, figure 1). Furthermore within each data set, users can select subsets such as a specific type of cancer or a subset of industry. For example, a user could start by plotting the incidence rate of prostate cancer, identifying 'hot spot' regions, and then begin plotting industry

subsets such as agriculture or manufacturing to search for overlap. In addition, users will be able to toggle between the absolute color scale and one which uses the difference from the national average. This interactive choropleth will enable rapid exploration of national level trends with resolution to the county level. By toggling between data types, users can begin to see correlations between them.

In addition to the map, the UI will display an ordered, horizontal bar graph that depicts cancer incidence rates in the selected county by incidence rate. This chart will display both the actual and predicted cancer rates grouped next to each other for each of the top five cancers to demonstrate the performance of our regression models. Next to the bar chart, a scatterplot has been drawn to demonstrate the trend of distribution of data between specific environmental impacts from a given county (X-axis) and its cancer incidents (Y-axis) visualised across all counties. The X-axis is a drop-down for various environmental indicators, while the Y-axis changes depending on a drop-down selection of all cancer incidents or specific cancers (e.g. breast cancer). There is a checkbox that allows the user to choose between the default option of the number of cancer incidents in the county, and the cancer incidence rates per 100 thousand in the given county. There is also a logarithmic scale for both the axes to make the scatter plot compact in nature. We notice that for some parameters, there is a point of inflection in the distribution of data, such that after a critical point in this parameter, the cancer incidence value starts increasing dramatically. This scatter plot serves as a preliminary analysis and displays a general trend of cancer incidences with respect to environmental impacts tracked by USEEIO.

This interactive visualization enables users to rapidly explore and identify features in a data set with more than 1.5M unique data points. However, the interactivity really shines when user's begin to ask 'what if?' questions about the relationship between individual industries and cancer rates in a given locality (Appendix A, figure 2). Specifically, the dashboard includes a set of five sliders next to the choropleth which can be used to manipulate the amount of each of the top five industries present in the currently selected county. When a user adjusts the sliders and clicks the predict button, the dashboard dynamically adjusts the size of each industry and its indicators and then predicts new annual counts for each of the top five cancers (Appendix A, figure 3). Specifics regarding the models and data flow are given in another section. Together, this tool set enables users to determine which cancer types are prevalent in a region of industry, begin to identify potential industrial correlations, and then evaluate what the effect of increasing or decreasing the amount of each industry would have on public health.

5. Experiments and Evaluation

5.1. Testbed description

There is one main question that we try to answer quantitatively in our experiments: what is the correlation (if any) between the industrial size present in a county (number of established industries, number of their employees, industrial byproduct output, etc.) and that county's affinity towards reported cancer cases? Overlapping that question, we want to know if certain combinations of these features correlate to the same. Once a model is fitted, we score its prediction capability based on the equation:

$$score = (1 - \frac{u}{v}) \quad \text{where} \quad u = \sum_{i=1}^{n} (y_{true_i} - y_{pred_i})^2 \qquad v = \sum_{i=1}^{n} (y_{true_i} - \frac{1}{n} \sum_{i=1}^{n} y_{true_i})^2$$

which allows models to have a range of score from 1 to -∞ where 1 is 100% accuracy and lower values are arbitrarily worse. This is known as 'R^2 scoring'. Then we choose the best and examine its weights.

5.2. Results – Details of the experiments performed

5.2.1. Initial exploration

To begin with, several of the regression techniques listed section 4.2.3 were tested. The input features for the results mentioned below were the aggregated (summation) environmental impacts for all industry types in a given county (24 environmental impacts) and the count of every lowest-level industries (335 industries) in a given county. Our results (Appendix B, figure 2) shows Ridge Regression achieved the best score of ~97.4%, therefore we choose to inspect its final weight distribution for clues. Looking closer at the index of the feature corresponding to the feature label in our data frame (Appendix B, figure 2a), we find that the largest weight corresponds to the metal output of an industry. This surprised our initial assumption that smog output of an industry would be the largest contributor to cancer.

5.2.2. Regression

For this part, we aggregated the industry indicators per county and trained various ML models to predict the annual average count of each cancer type per that county. The ML models used for this purpose are random forest, logistic regression, SVR, multi-task lasso, and multi-task elastic net with (Appendix B, figure 4). In one experiment, we tried training these models only on the PCA components. In another experiment we tried to train on the most important factors from the last row of the covariance matrix (Appendix B, figure 6). Both these experiments resulted in worse metrics than ML models trained on all factors. Out of all the models, we decided to proceed with the Multi-Task Lasso regression model trained on all factors that achieved ~90% accuracy during testing. In addition, by analyzing the RF feature importance (Appendix B, figure 4a, right) we realized that Food is the most prominent factor contributing to the annual average count of all cancer types, followed by Metal then Municipal Solid Waste.

5.2.3. Classification

In a supervised learning experiment, we used LDA to create new features that reduced the dimensions of our 24 environmental indicators into single metrics. For classification purposes, we created a new variable. This variable had a value of 0 if a specific cancer's incidents_rates in a county were 1 standard deviation below the mean, 1 if incidents rates were within a standard deviation of the mean, and 2 if the incidents rates were 1 standard deviation above the mean. Moreover, we used LDA to predict this classification from the LDA features that we created earlier. The average R^2 evaluation of this classification from the random forest regressor for all cancer types was ~88% (Appendix B, figure 5). Our model had an accuracy of 100 percent when it came to predicting this number for cancers such as esophagus, ovary and brain. In this method, we managed to give a general prediction of low, mid and high number of incidents rates per county, instead of predicting the actual number. The results from PCA (Appendix B, figure 6) were also very similar to the results of the feature importance for each cancer from the random forest regressor. Both methods gave a high importance to the food factor almost for all cancer types. The next factors following were Metal and Municipal Solid Waste.

5.2.4. Experimenting with a two model pipeline

In our previously discussed prediction models, the input vector contains the 24 industry indicators. Now when we want to predict how the cancer incidence rate or the cancer count increases or decreases if we add an automobile factory in that county for example (i.e. increasing the number of establishments, employees and annual pay of that factory), we don't know how this change will

affect the 24 environmental factors corresponding to that industry in that particular county. In order to get past this issue, the assumption that we made above was that these environmental factors would change linearly as a function of the annual pay or the number of employees. This assumption might hold for factors such as 'Minerals and Metals', however we could not say the same thing for factors such as 'Human Health - Respiratory Effects'. As a result, our initial plan was to use a two-model pipeline, where the first model would predict how the 24 environmental factors would change, which we would then feed to the model that predicts the cancer data. However, we tried all the same models as those discussed in the Regression section, most of these models performed very badly. As a result, for the sake of coherence, we chose not to use the two-step model, thus sticking with the linearity assumption.

6. Conclusions and Discussion

Our predictive model achieved a satisfying function for mapping industrial indicators to cancer rates, but we were unable to find a similarly satisfying mapping from the localized industry size to their indicators. After the analysis, it is evident that there are certain industrial features indicated as a high risk of cancer to those living in its vicinity. In our findings we confidently see that the amount of metal an industry outputs is heavily weighted toward overall cancer risk. Curiously, we found that the food indicator of an industry is weighted the heaviest in our model. From a ground truth baseline of industry size, our model is able to predict near actual cancer values which gives confidence that adjusted industry quantities give satisfying approximations of adjusted cancer rates per county. On the subject of limitations; a large assumption was made in the linear scaling of those industry quantities which could contribute to the error deviating from baseline. In the future, we would pursue our 2-model pipeline further and to map the industry size more accurately to their indicators.

We succeeded in offering an intuitive framework for the visualization of cancer cases in U.S. counties that allows the user to ask questions about the role sizes of certain industries play in the cancer rates of their local populace. We shed light on this so that we may arm lawmakers and citizens with relevant information, thus allowing them to take appropriate measures to decrease industry-related preventable health hazards, and to make informed decisions on business opportunities by prioritizing the health of their communities. Another practical implication of this work stems from the open-access interactive interface that can inform citizens or potential home buyers about the least exposed-to-cancer counties in an intuitive way. The burden of cancer can be caused by numerous factors, including genetic factors, smoking, obesity, among other things and these factors are still not well understood. However, we believe light was satisfyingly shed in at least one corner of this darkened room.

"All team members have contributed similar amount of effort"

References

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2020. *CA. Cancer J. Clin.* (2020). doi:10.3322/caac.21590

2. Park, J. & Look, K. A. Health Care Expenditure Burden of Cancer Care in the United States. *Inq. J. Heal. CareOrganization, Provision, Financ.* **56:1–9**, 1–9 (2019).

3. Crimmins, A. *et al.* The Impacts of Climate Change on Human Health in the United States: A Scientific Assessment. *U.S. Glob. Chang. Res. Progr.* (2016).

4. Kogevinas, M. *et al.* Cancer risk in the rubber industry: A review of the recent epidemiological evidence. *Occupational and Environmental Medicine* (1998). doi:10.1136/oem.55.1.1

5. Fernández-Navarro, P., García-Pérez, J., Ramis, R., Boldo, E. & López-Abente, G. Proximity to mining industry and cancer mortality. *Sci. Total Environ.* (2012). doi:10.1016/j.scitotenv.2012.07.019

6. Lacourt, A., Pintos, J., Lavoué, J., Richardson, L. & Siemiatycki, J. Lung cancer risk among workers in the construction industry: Results from two case-control studies in Montreal. *BMC Public Health* (2015). doi:10.1186/s12889-015-2237-9

7. McLeod, K. S. Our sense of Snow: The myth of John Snow in medical geography. in *Social Science and Medicine* (2000). doi:10.1016/S0277-9536(99)00345-7

8. PAVLOVSKY, E. N. Natural Focal Localization of Human Transmissible Diseases and the Concept of Landscape as an Epidemiological Factor. *Meditsinskaya Parazitologiya i Parazitarnye Bolezni* (1944).

9. Nuckols, J. R., Ward, M. H. & Jarup, L. Using geographic information systems for exposure assessment in environmental epidemiology studies. *Environmental Health Perspectives* (2004). doi:10.1289/ehp.6738

10. Bellinger, C., Mohomed Jabbar, M. S., Zaïane, O. & Osornio-Vargas, A. A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health* (2017). doi:10.1186/s12889-017-4914-3

11. Vopham, T., Hart, J. E., Laden, F. & Chiang, Y. Y. Emerging trends in geospatial artificial intelligence (geoAI): Potential applications for environmental epidemiology. *Environmental Health: A Global Access Science Source* (2018). doi:10.1186/s12940-018-0386-x

12. Mishra, V. N., Rai, P. K., Prasad, R., Punia, M. & Nistor, M. M. Prediction of spatio-temporal land use/land cover dynamics in rapidly developing Varanasi district of Uttar Pradesh, India, using geospatial approach: a comparison of hybrid models. *Appl. Geomatics* (2018). doi:10.1007/s12518-018-0223-5

13. Kumar, A., Kakkar, A., Majumdar, R. & Baghel, A. S. Spatial data mining: Recent trends and techniques. in *2015 International Conference on Computer and Computational Sciences, ICCCS 2015* (2015). doi:10.1109/ICCACS.2015.7361319

14. Elliot, P., Wakefield, J. C., Best, N. G., & Briggs, D. J. (2000). *Spatial epidemiology: methods and applications. Oxford University Press. Oxford Uni Press* (2000). doi:10.1093/acprof

15. Wang, H. *et al.* Epidemiological data analysis in TerraFly Geo-spatial Cloud. in *Proceedings - 2013 12th International Conference on Machine Learning and Applications, ICMLA 2013* (2013). doi:10.1109/ICMLA.2013.166

16. Vieira, V., Webster, T., Weinberg, J., Aschengrau, A. & Ozonoff, D. Spatial analysis of lung, colorectal, and breast cancer on Cape Cod: An application of generalized additive models to case-control data. *Environ. Heal. A Glob. Access Sci. Source* (2005). doi:10.1186/1476-069X-4-11

17. Carroll, L. N. *et al.* Visualization and analytics tools for infectious disease epidemiology: A systematic review. *Journal of Biomedical Informatics* (2014). doi:10.1016/j.jbi.2014.04.006

18. Roquette, R., Painho, M. & Nunes, B. Spatial epidemiology of cancer: A review of data sources, methods and risk factors. *Geospat. Health* (2017). doi:10.4081/gh.2017.504

19. Beale, L., Abellan, J. J., Hodgson, S. & Jarup, L. Methodologic issues and approaches to spatial epidemiology. *Environ. Health Perspect.* (2008). doi:10.1289/ehp.10816

20. Yang, Y., Ingwersen, W. W., Hawkins, T. R., Srocka, M. & Meyer, D. E. USEEIO: A new and transparent United States environmentally-extended input-output model. *J. Clean. Prod.* (2017). doi:10.1016/j.jclepro.2017.04.150

21. National Program of Cancer Registries and Surveillance, Epidemiology, and End Results SEER*Stat Database: NPCR and SEER Incidence – U.S. Cancer Statistics 2001–2016 Public Use Research Database, November 2018 submission (2001–2016), United States Department.

22. U.S. Census Bureau (2018). 2012-2017 County Business Patterns. Retrieved from API at api.census.gov/data/2017/cbp.

23. Jakkula, V. Tutorial on Support Vector Machine (SVM). *Sch. EECS, Washingt. State Univ.* (2011).

24. Jenkins, W. D., Christian, W. J., Mueller, G., & Robbins, K. T. (2013). Population cancer risks associated with coal mining: a systematic review. PloS one, 8(8), e71312. https://doi.org/10.1371/journal.pone.0071312

25. Blackadar C. B. (2016). Historical review of the causes of cancer. World journal of clinical oncology, 7(1), 54–86. https://doi.org/10.5306/wjco.v7.i1.54

Appendix

A. Visualization Demo

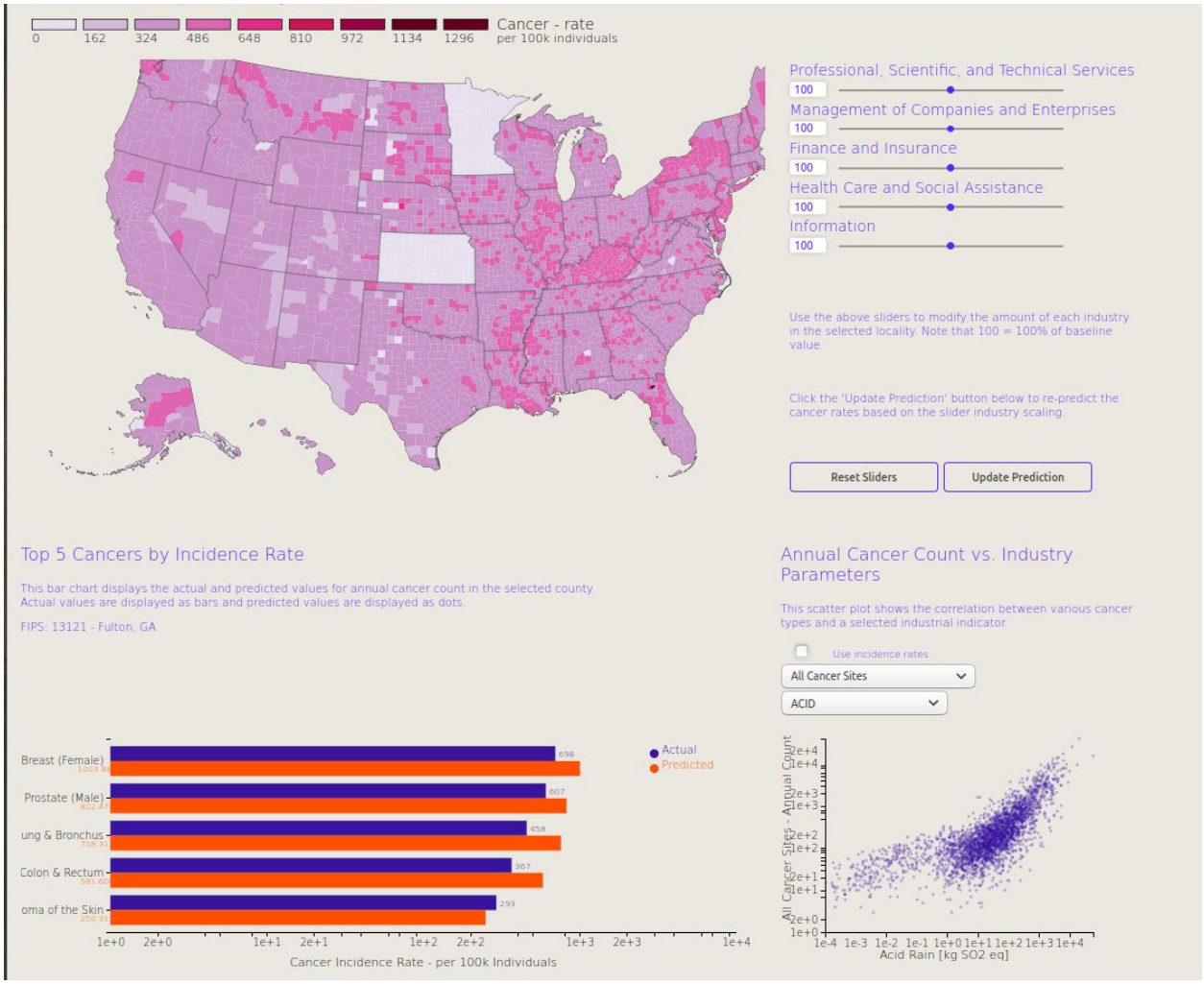

Figure 1: Visualization Dashboard at a National View

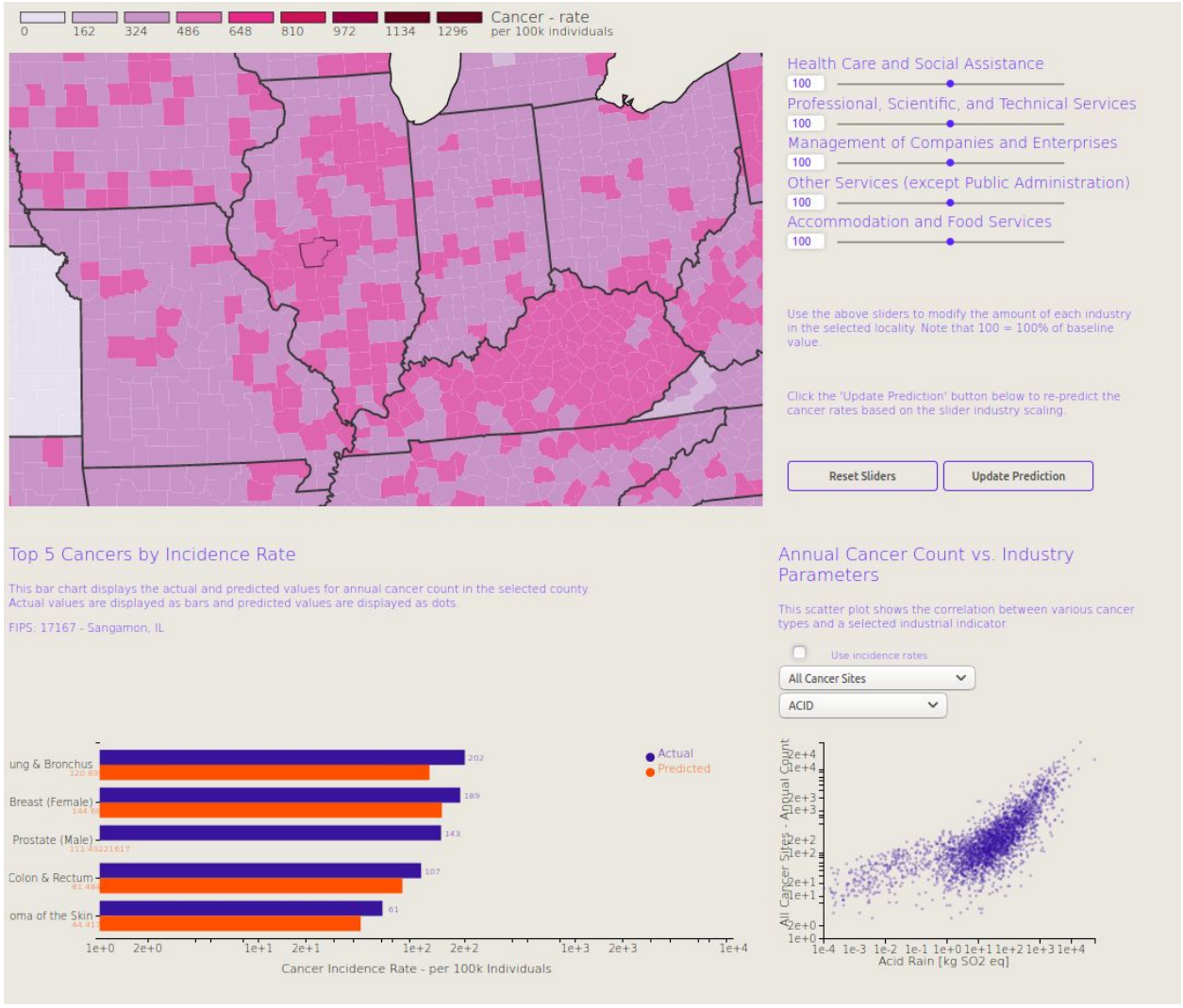

Figure 2: Visualization Dashboard at a County View (Sangamon, IL) with Baseline Predictions

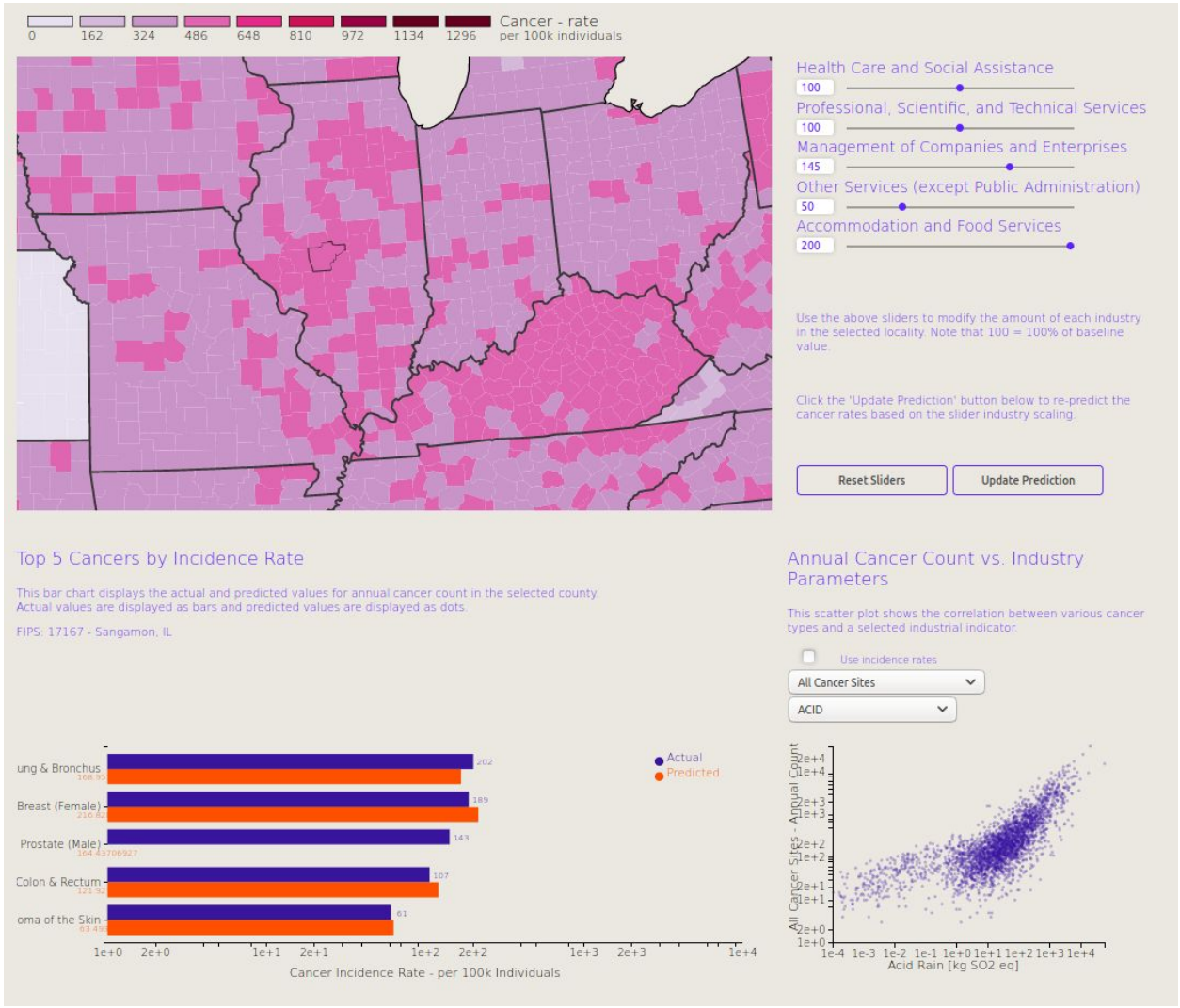

Figure 3: Visualization Dashboard at a County View (Sangamon, IL) with Altered Predictions

B. Model Experimental Results

Algorithm	Score
Support Vector Regression (linear)	-2.8138950637684386
Support Vector Regression (radial basis function)	-0.08660670689272254
Ridge Regression	0.973769356194365
Lasso	0.9548959635651908
Elastic Net	0.9735836527099999

Figure 2: Initial Omni-State-Space to Cancer Rates Model Tests and Scores

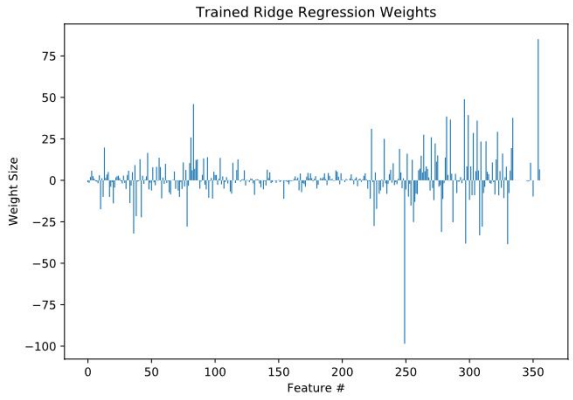

Figure 2a: Initial Ridge Regression Weights for Omni-State-Space Input

Algorithm	Score
Linear Regression (standard scaled input)	0.0388868300359007

Linear Regression (non-scaled input)	0.11150001835481553
Random Forest (non-scaled input, estimators = 100)	0.011901444372792494
Random Forest (standard scaled input, estimators = 100)	0.9261049905173594
Random Forest (standard scaled input, estimators = 800)	0.9311754629911315
Elastic Net	0.046084232236825874
Ridge Regression	0.1115001227804399
Lasso	0.0990408675208958
Lasso (LARS optimizer)	0.027145855239856496
Multi-Task Elastic Net	0.10879486613823279
Multi-Task Lasso	0.10399256911880145

Figure 3a: Industry Scaling to Indicator Model Tests and Scores

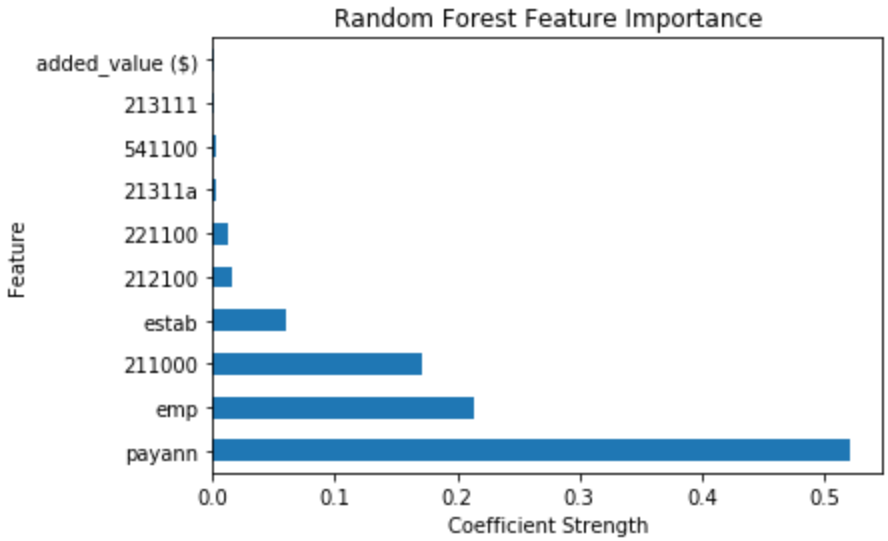

Figure 3b: Random Forest Feature Analysis for Industry-Size Input

Algorithm	Score
Multi-Task Elastic Net	0.7833450560976941
Multi-Task Lasso	0.8959449771883728
Lasso (single output average, alpha = 0.01)	0.8604212356280521
Lasso (single output average, alpha = 0.001)	0.864774596122108
Elastic Net (single output average)	0.032863967062976084
Ridge Regression (single output average, alpha = 0.5)	0.8703244136041673
SVR (single output average, epsilon = 0, max_iterations = 64, tol = 0.1)	-0.15505755597593907
Random Forest Regressor (single output average, estimators = 100, max_depth = 8)	0.8126109753659849

Figure 4: Indicator to Cancer Rates Model Tests and Scores

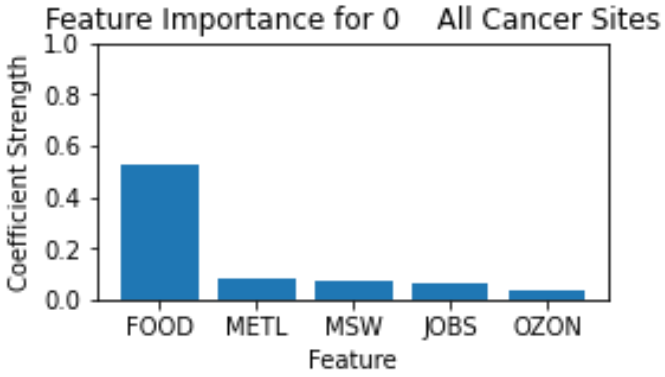

Figure 4a: Feature Analysis for Indicator Input Model

Algorithm	R2 Score (all environmental factors)	R2 Score (HTOX and HC)
Random Forest Regressor	0.8700949707169692	0.41378470045227933

Logistic Regression	0.06170212765957447	0.11461408330990763
SVR	0.1925047582004683	-0.08589664158110089

Figure 5: Feature Analysis for Indicator Input Mode

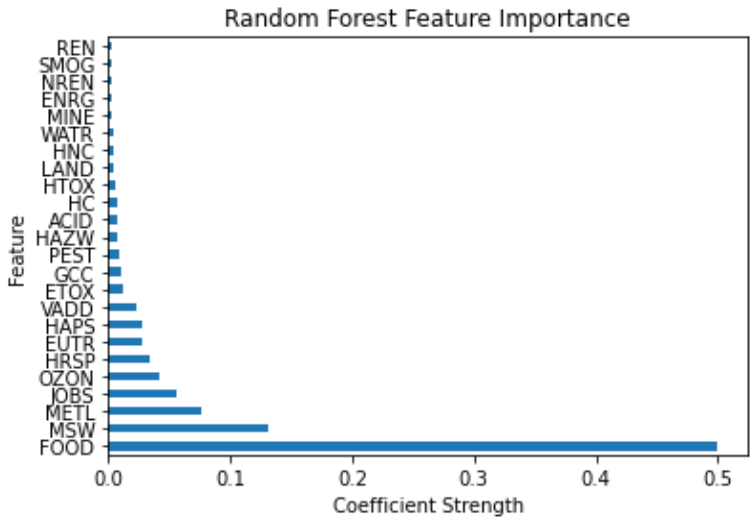

Figure 5a: Feature Analysis for Indicator Input Model

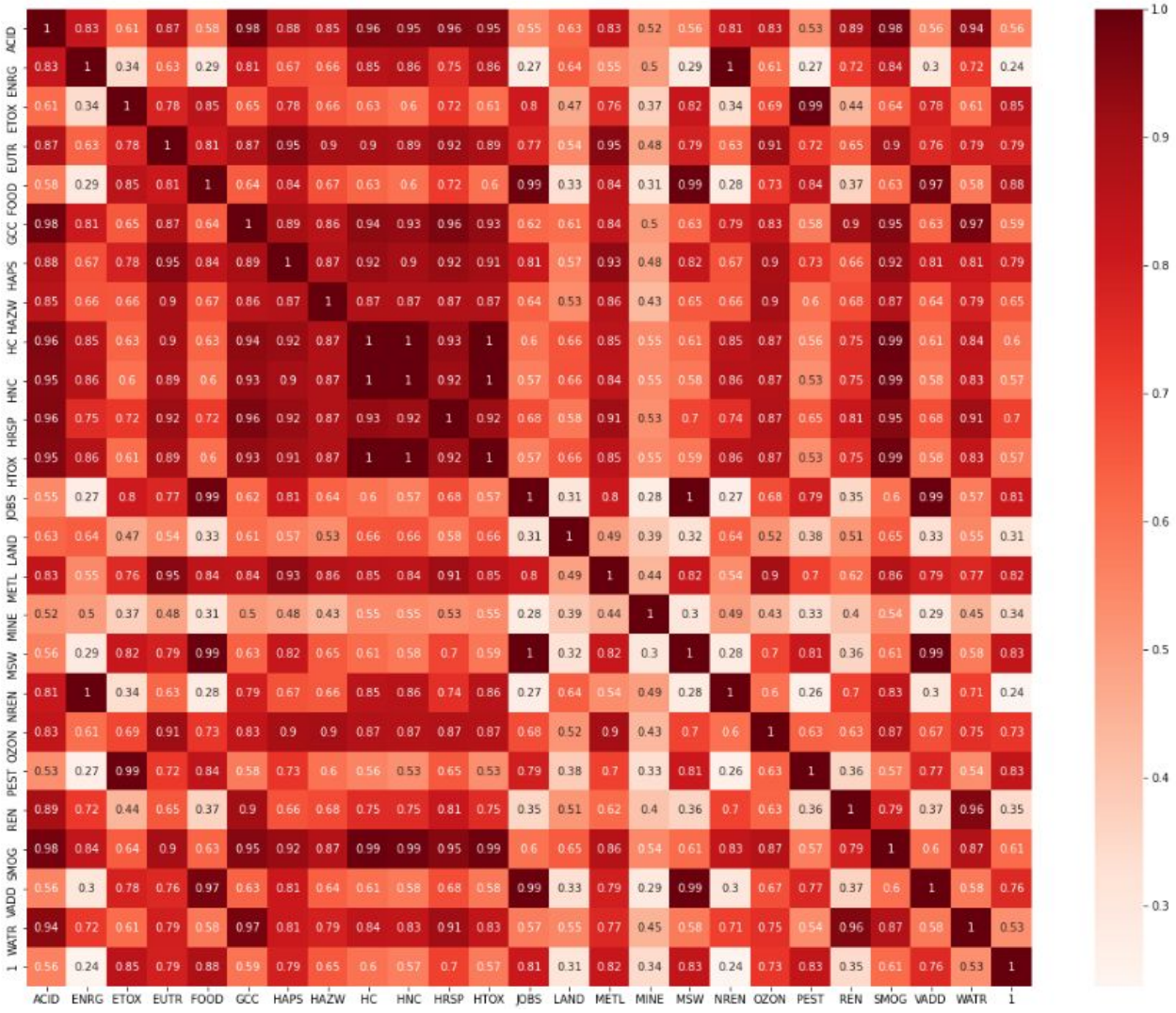

Figure 6: Covariance matrix for all cancer types.