

**Spatial Distribution and Association of the Relative Risk of Lung Cancer Incidence by Census Tracts
Between 2013-2017 Across New York State (NYS) with Particulate Matter (PM) 2.5 microns and Adult
Smoking Rates at the County Level**

Stony Brook University Program in Public Health

Hannah Wang, Tuyet-Anh Nguyen, Lan Lei

HPH 534 Spatial Analysis: Health Applications

Professor Das

May 17, 2021

Spatial Analysis of Lung Cancer Incidence in New York State

Background

Lung cancer is the leading cause of cancer deaths in New York State. According to the NYS Department of Health (n.d.), about 14,100 New York State residents are diagnosed with lung cancer, and about 8,000 people die from this disease each year. Many risk factors are associated with lung cancer, such as lifestyle, environmental exposure, and occupational exposures. Smoking is widely considered the leading cause of lung cancer. Additional risk factors include but are not limited to radiation and chemicals, family history, viruses and bacteria, certain hormones (Li et al., 2018; Wei et al., 2017).

Knowing the risk factors for lung cancer is key to taking steps to help prevent the disease, this spatial analysis project focused on the spatial distribution and association of lung cancer incidence in New York State between 2 exposures for comparison, which were PM2.5 and adult smoking rate. PM2.5 is characterized by its small particle size and large surface absorption ability which makes it possible to invade the small airways (Wang et al., 2020). Scientific studies have linked increased long-term population exposure to PM2.5 with elevated respiratory disease incidence by affecting lung function and worsening the medical conditions (Wolfe et al., 2019). Another risk factor we focused on is smoking. Approximately 87 percent of lung cancers are caused by cigarette smoking as reported by the Surgeon General Report (CDC, 2014). The relative risk will be increased with a higher number of cigarettes smoked consumed per day. Heavy smokers and those who began smoking at a young age are at an elevated risk; non-smokers who are regularly exposed to secondhand smoke will significantly increase their risk of lung cancer (McDowell, 2020).

The burden of cancer in New York State is considerable because of its notable size and population diversity. Nevertheless, the temporal trends of lung cancer mortality in New York State have decreased over time for most counties since 1994 (Juster, 2017). State lung cancer mortality rates decreased by 11.9% from 1994 to 2013, which was due to the enhanced screening and clinical practice

(Henschke et al., 2010). Cancer screening is intended to detect cancers at their early stages when treatment is most effective. The New York State Department of Health offers free lung cancer screening through cancer services program partnerships in every county in New York State; thus, CT lung screening has significantly reduced lung-cancer mortality by 36% from 2001 to 2005 (Henschke et al., 2010). Furthermore, NYS started fine particulate matter monitoring with a systematic PM_{2.5} monitoring network in 1998. The PM_{2.5} monitoring network primarily works for determining average environmental exposures for most populations (NYS Department of Health, n.d.). Lastly, smoking cessation programs in NYS provide comprehensive and affordable treatment choice. The NYS Smoker Quitline also provides free starter kits of nicotine replacement therapy (NRT) for eligible New Yorkers (NYS Department of Health, n.d.).

Because no previous study has been done to specifically evaluate the correlation between PM_{2.5} and smoking exposure and their attributions to lung cancer incidence in New York State, there is a potential need for research to fill in the gap in the literature. To improve our scientific understanding of the relationship between risk factors and clusters of disease, this project was conducted to identify the significant clustering of lung cancer and its risk factors in New York State.

Objective

While smoking cigarettes is the biggest risk factor for lung cancer and causes about 80% of deaths from the disease, people who do not smoke can develop lung cancer too (McDowell, 2020). A new study co-led by the US Centers for Disease Control and Prevention (CDC) and the American Cancer Society (ACS) consisting of a large sample of over 129,000 cases of lung cancer based on data from cancer registries in 7 states found that 12% of people in the United States who were recently diagnosed with lung cancer had never smoked cigarettes (Siegel, Fedewa, Henley, Pollack, & Jemal, 2021). Other risk factors for lung cancer include secondhand smoke exposure, occupational exposures, radon, air pollution, and genetic factors. The objective of this spatial analysis project is to investigate the spatial

distribution of the relative risk of lung cancer incidence by census tracts between 2013-2017 across NYS as well as evaluate for any correlation with the covariates of particulate matter (PM) 2.5 microns and adult smoking rates in percentage at the county level. Although previous spatial clustering analysis of lung cancer have been done using cancer incidence data from 2011-2015 (Health, 2018), this study's analysis uses data from 2013-2017, which will allow for comparison of clustering across two partially overlapping time periods. The alternative hypothesis for this analysis is that there is significant clustering of census tracts for lung cancer relative risk in New York State between 2013-2017.

Methodology and Study Design

This is a cross-sectional analysis of the state-wide population-based dataset from the New York State Cancer Registry (NYSCR), which provides incident lung cancer observed and expected cases aggregated at the census-tract level for the diagnosis years from 2013-2017. The analysis aims to investigate whether clustering is statistically significant or whether it is likely to be a chance occurrence or is simply a reflection of the distribution of the population at risk. In other words, does the observed spatial pattern differ significantly from the null hypothesis of complete spatial randomness? The statistical significance of a cluster will be determined using Monte Carlo simulation and permutations of simulated datasets.

Datasets

- United States Census Bureau "2010 TIGER/Line Shapefiles" from <https://www.census.gov/cgi-bin/geo/shapefiles/index.php>.
- 2013-2017 NYSCR Cancer Incidence by Census Tract publicly-use dataset provided courtesy of NYS Dept. of Health (NYSDoH) as a SAS datafile.
- 2019 County Health Rankings and Roadmaps for NYS -- <https://www.countyhealthrankings.org/app/new-york/2019/downloads> -- contains aggregated

measures of PM 2.5 and adult smoking rates at the county-level collated from other primary sources of data:

- The original source of the air pollution data is the Environmental Public Health Tracking Network, in which air particulate matter is the average daily density of fine particulate matter in micrograms per cubic meter (PM2.5) from 2014
- The original source of % of adults who are current smokers is the individual-level Behavioral Risk Factor Surveillance System survey from 2016

Statistical Analyses

Given that we have cancer incidence data aggregated at the census-tract level for **n = 4775 census tracts in NYS**, three different local methods of cluster detection were utilized to analyze spatial clustering in diagnoses of lung cancer in NYS to allow for comparison and contrasts of the spatial autocorrelation results across techniques. A combination of statistics should be used when studying local clustering to ensure different aspects of spatial patterns are identified and to check whether the results from different analyses are consistent. These three methods -- **local Moran's I**, **Getis-Ord Gi***, and **Kulldorff's spatial scan statistics** -- can identify the location and extent of statistically significant clusters or spatial patterns of the relative risk of lung cancer incidence over the period of 2013-2017, defined as the number of cases of cancer observed divided by the population-adjusted and age-adjusted number of cases expected in each census tract. Autocorrelation statistics for aggregated data provide an estimate of the degree of spatial similarity observed among neighboring values of an attribute over a study area. Expected incidence or cases is the number of people in a given census tract that would be expected to develop cancer within a five-year period if the census tract had the same rate of cancer as the State as a whole (NYSCR, 2017). The cancer rate for the entire state along with the age structure and the number of people in a census tract are used to estimate the expected incidence. Please note that the descriptions on the different spatial scan statistics excerpted below are taken from the ArcGIS

Desktop help function at <https://desktop.arcgis.com> as well as from the Spatial Analysis in Epidemiology book by Dirk Pfeiffer et al (Dirk U. Pfeiffer, 2008). SaTScan was downloaded for free at SaTScan.org and its spatial scan statistic calculation methodology are detailed on the website as well (SaTScan, 2005).

Local Moran's I statistic, also termed Local Indicators of Spatial Association (LISA) statistic, was calculated in **ArcGIS Map** by assigning data for each census tract to its centroid and then using spatial correlograms to evaluate the distances where spatial effects are greatest. The local Moran test detects local spatial autocorrelation in aggregated data by decomposing Moran's I statistic into contributions for each area within a study region. These indicators detect clusters of either similar or dissimilar disease frequency values around a given observation. The sum of the LISAs for all observations is proportional to the global Moran's I statistic. Moran's I can indicate whether there is spatial autocorrelation of the relative risks of lung cancer in New York state between 2013-2017 at the census-tract level by considering similarity between neighboring regions. A positive value for the I coefficient indicates that a feature has neighboring features with similarly high or low attribute values and that this feature is part of a cluster while a negative value for the I coefficient indicates that a feature has neighboring values with dissimilar values and that this feature is an outlier; values of 0 indicates the null hypothesis of no clustering. The cluster/outlier type (COType) field distinguishes between a statistically significant cluster of high values (HH), cluster of low values (LL), outlier in which a high value is surrounded primarily by low values (HL), and outlier in which a low value is surrounded primarily by high values (LH). Statistical significance is set at 95 percent confidence level. Analysis was done with and without false discovery rate (FDR) correction. FDR correction reduces the p-value threshold from 0.05 to a value that better reflects the 95% confidence level given multiple testing. The conceptualization of spatial relationships field was set as "inverse distance," the distance method field was set as "Euclidean distance," and the number of permutations set at "999."

Another method of cluster detection used is the hot spot analysis tool in **ArcGIS Map** that calculates the **Getis-Ord Gi* statistic** for each feature in a dataset. The resultant z-scores and p-values identifies where features with either high or low values cluster spatially. This tool works by looking at each feature within the context of neighboring features. To be a statistically significant hotspot, a feature will have a high value and be surrounded by other features with high values as well. The Gi* statistic returned for each feature in the dataset is a z-score. The local sum for a feature and its neighbors is compared proportionally to the sum of all features; when the local sum is very different from the expected local sum, and when that difference is too large to be the result of random chance, a statistically significant z-score results. Analysis was done both with and without FDR correction. If FDR correction was applied, then the statistical significance was adjusted to account for multiple testing and spatial dependency. The conceptualization of spatial relationships field was set as “inverse distance” and the distance method field was set as “Euclidean distance.” For statistically significant positive z-scores, the larger the z-score is, the more intense the clustering of high values (hot spot). For statistically significant negative z-scores, the smaller the z-score is, the more intense the clustering of low values (cold spot).

The cluster detection software called **SaTScan** was used to search for clusters in datasets using the Poisson probabilistic population-based model, where the number of cases is compared to the background population data and the expected number of cases in each unit is proportional to the size of the population at risk. SaTScan employs Kulldorff’s method of cluster detection to identify areas of unusual disease patterns and calculates the **Kulldorff’s spatial scan statistic**, which is the test statistic using observed data and then re-calculates the test statistic using a specified number (999) of simulated data sets (or permutations) to generate the expected distribution of the test statistic under the null hypothesis. The likelihood of obtaining the value for the test statistic derived from the observed data is then calculated and expressed as a p-value. In this spatial scan method, for each specified location, a

series of circles of varying radii is constructed, wherein each circle absorbs the nearest neighboring locations that falls inside it, and the radius of each circle is set to increase continuously from zero until some fixed percentage of the total population is included. Both parameters of 25% and 50% of the total population were tried with only minor differences in the number of clusters detected. For each circle, the alternative hypothesis is that there is an elevated risk of disease within the circle compared to that outside.

Linear regression analyses were also conducted for the dependent variable of relative risk for lung cancer incidence over the period of 2013-2017 using predictor variable of PM_{2.5} while controlling for the average smoking rates at the county level using SAS 9.4 and SPSS.

Data Management Using Microsoft Excel and ArcGIS Map

The NYSCR cancer incidence dataset was provided as a SAS datafile, which had to be converted into an Excel format. Although there were a total of 4919 census tracts in NYS based on the 2010 US Census Tract shapefile (Bureau, 2010), merging this information with the NYSCR cancer incidence data led to matching of 4775 census tracts. This is most likely because some census tracts had too few cases to be shown for confidentiality reasons, and thus these census tracts are combined with neighboring census tracts and are listed with the census tract with which they were combined with numerical values given for the combined group of census tracts (Bureau, 2010). To merge the NYSCR cancer incidence dataset to the US Census Tract shapefile in ArcGIS Map, a “GEOID” had to be created in Excel using the concatenation function to combine the fields containing state code, county code, and census tract number in the NYSCR dataset to match up with the GEOID unique identifier for each census tract in NYS that was already available in the US Census Tract shapefile’s attribute table. Thus, these two datasets were joined in ArcGIS Map using this unique identifier.

The exposure dataset containing information on smoking rates and PM 2.5 at the county-level were taken from the County Health Rankings database and merged with the 2010 US Census County

shapefile using ArcGIS Map join by table function, leading to 100% matching of all 62 counties in NYS. This newly merged exposure dataset at the county-level is then joined with the cancer dataset at the census-tract level using the unique identifier of county code, wherein each county-level observation is replicated for all the census-tracts within that county. Finally, once all the data have been joined together, the data had to be extracted and saved as a permanent shapefile on which geoprocessing and geostatistical analysis can be conducted. This newly created shapefile was also converted into an Excel file into for analysis in SaTScan.

Study Methodology Strengths and Limitations

Assumptions of the local Moran's I method of cluster detection include the following: population at risk is evenly distributed within the study area, the correlation or covariance is the same in all directions (isotropic), accounting for the population at risk within each test area/region (by normalizing by population), and the attribute of interest is normally distributed. The Kulldorff's spatial scan statistic reports only the circles with the highest likelihood ratio that are non-overlapping and that are statistically significant. However, the same problem remains that large areas with large populations (and thus large absolute excesses) and small elevations in risk are more often identified due to their inherently greater statistical power compared to smaller sub-clusters within these areas that have higher elevations in risk but lower likelihood ratios (Boscoe, McLaughlin, Schymura, & Kielb, 2003). Advantages of Kulldorff's spatial scan statistic is that it can identify circular clusters of any size, located anywhere within a study area, while controlling for multiple hypothesis testing (Boscoe et al., 2003).

Limitations with respect to the datasets used include the following -- given that the unit of analysis was at the census-tract level, the study lacked individual-level exposure information, individual-level information regarding duration of exposure, and prior residence history. Migration of people into and out of census tracts and the length of residency of the cohort in each census tract are not accounted for in this analysis. The covariates of smoking percentage originated from the 2016 BRFSS self-reported

survey while the air pollution (PM 2.5) exposure data originated from the Environmental Public Health Tracking Network in 2014, in which both variables were aggregated and averaged at the county-level in the County Health Rankings & Roadmaps database. Having different scales of data makes any correlation to be suspect and may not give enough granular information to find other correlations. Given that there is a long period of latency between exposure, whether from smoking or air pollution, and the development of cancer, using exposure data during a period that overlaps with the period of cancer diagnosis requires the assumption that the air particulate matter and the adult smoking rates remains unchanged over time. Finally, and most importantly, causal inference for lung cancer clustering or even exposure association with lung cancer at the individual level is extremely limited and cannot be made given that this a population-based cross-sectional study design with risks of ecological fallacy.

Results

Table 1: SPSS output of the linear regression analysis of relative risk for **lung cancer** incidence at the census tract level using predictors of percent adult smoking rates and levels of air particulate matter 2.5 microns at the county level for the entire NYS. The analysis found that percent adult smoking is positively (coefficient = 0.039) and significantly ($p = 0.000$) correlated with the relative risk for incident lung cancer, but there is no statistically significant association with PM 2.5 and lung cancer relative risk ($p = 0.928$)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.217 ^a	.047	.047	.485392133206667

a. Predictors: (Constant), Avg_Daily, Pct_Adult

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	55.541	2	27.770	117.868	.000 ^b
	Residual	1124.310	4772	.236		
	Total	1179.850	4774			

a. Dependent Variable: Lung_RR

b. Predictors: (Constant), Avg_Daily, Pct_Adult

Coefficients^a

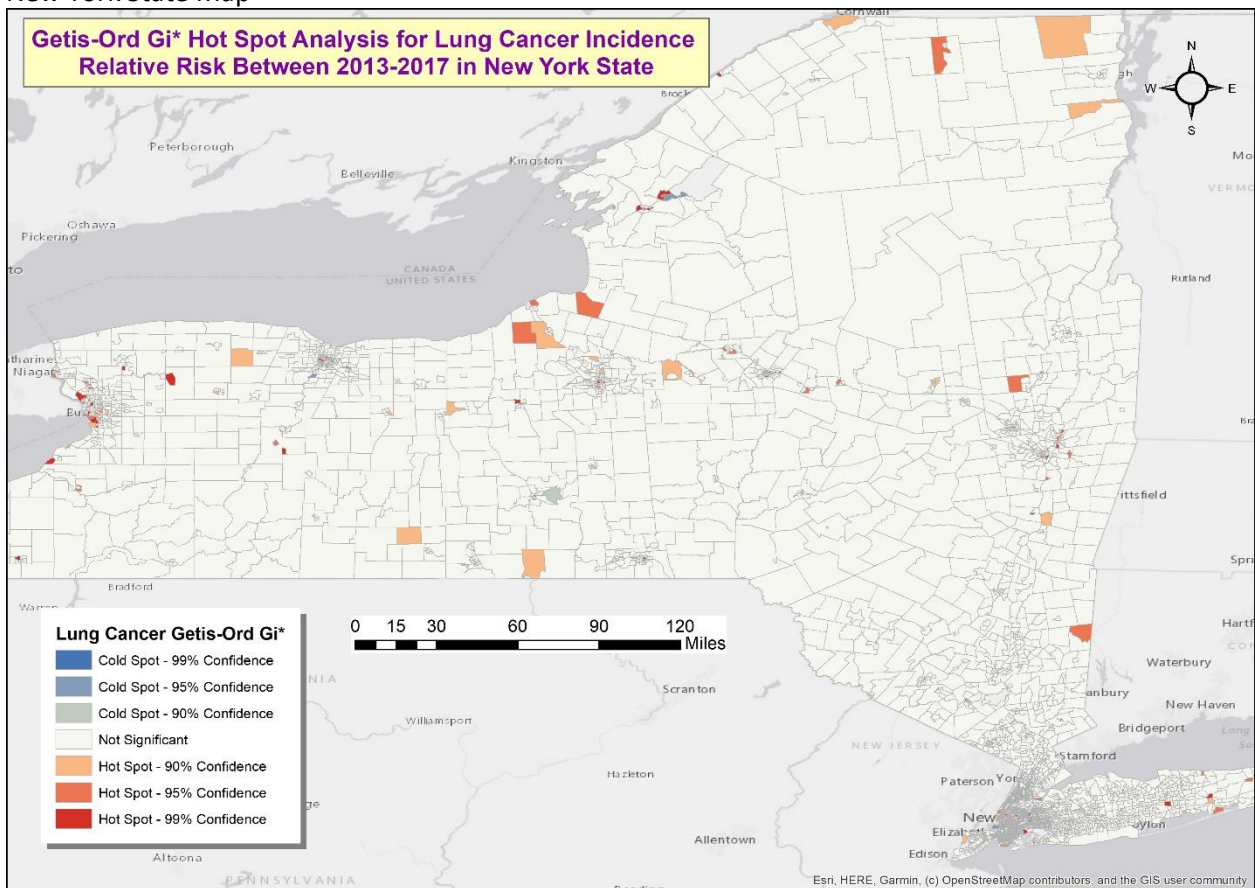
Model	Unstandardized Coefficients	Standardized Coefficients	t	Sig.
-------	-----------------------------	---------------------------	---	------

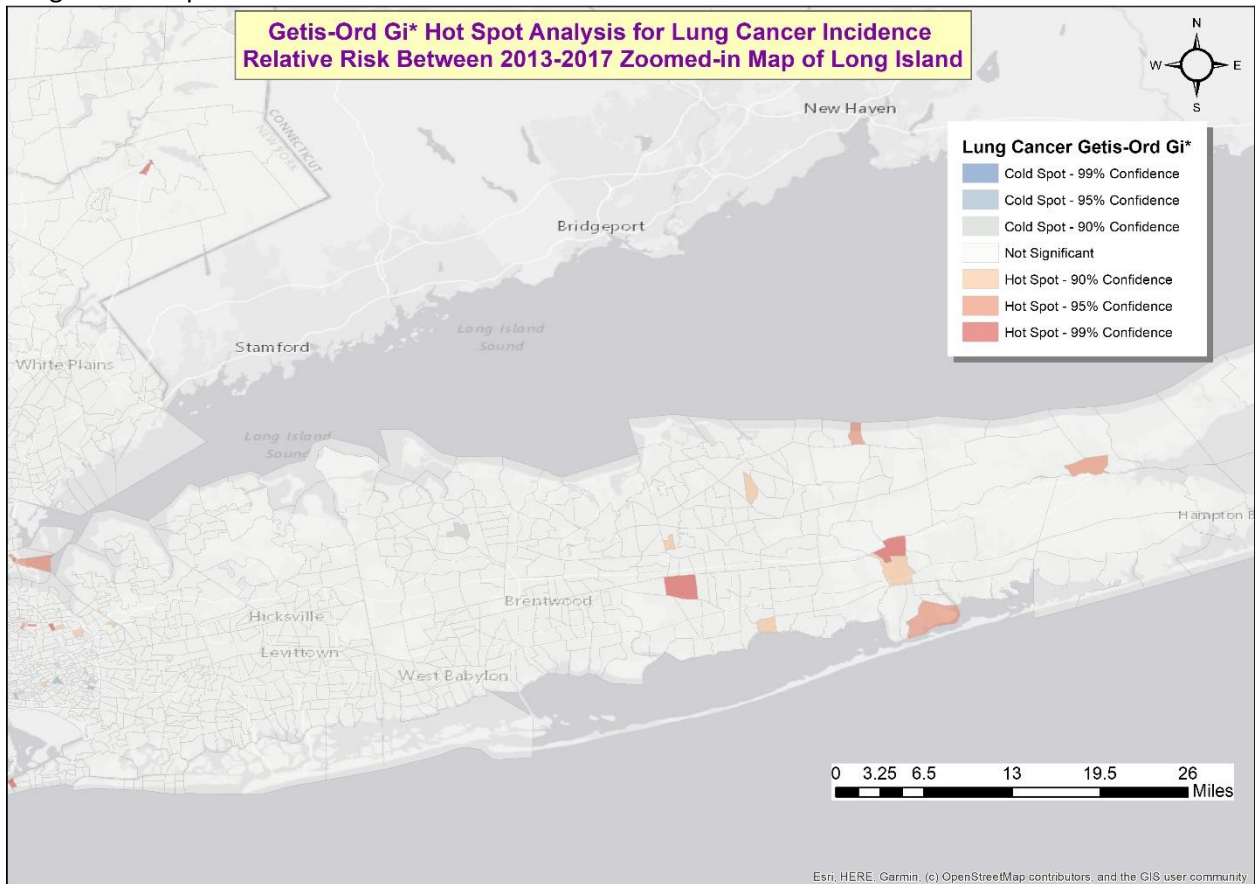
		B	Std. Error	Beta		
1	(Constant)	.451	.083		5.441	.000
	Pct_Adult	.039	.003	.217	15.143	.000
	Avg_Daily	-.001	.007	-.001	-.090	.928

a. Dependent Variable: Lung_RR

Figures 1A-D: Maps of Getis-Ord Gi* Hotspot Analysis for Incident Lung Cancer Relative Risk **without FDR correction**. All hotspots on Long Island become insignificant with a smaller number of significant hotspots in NYC when FDR correction was applied.

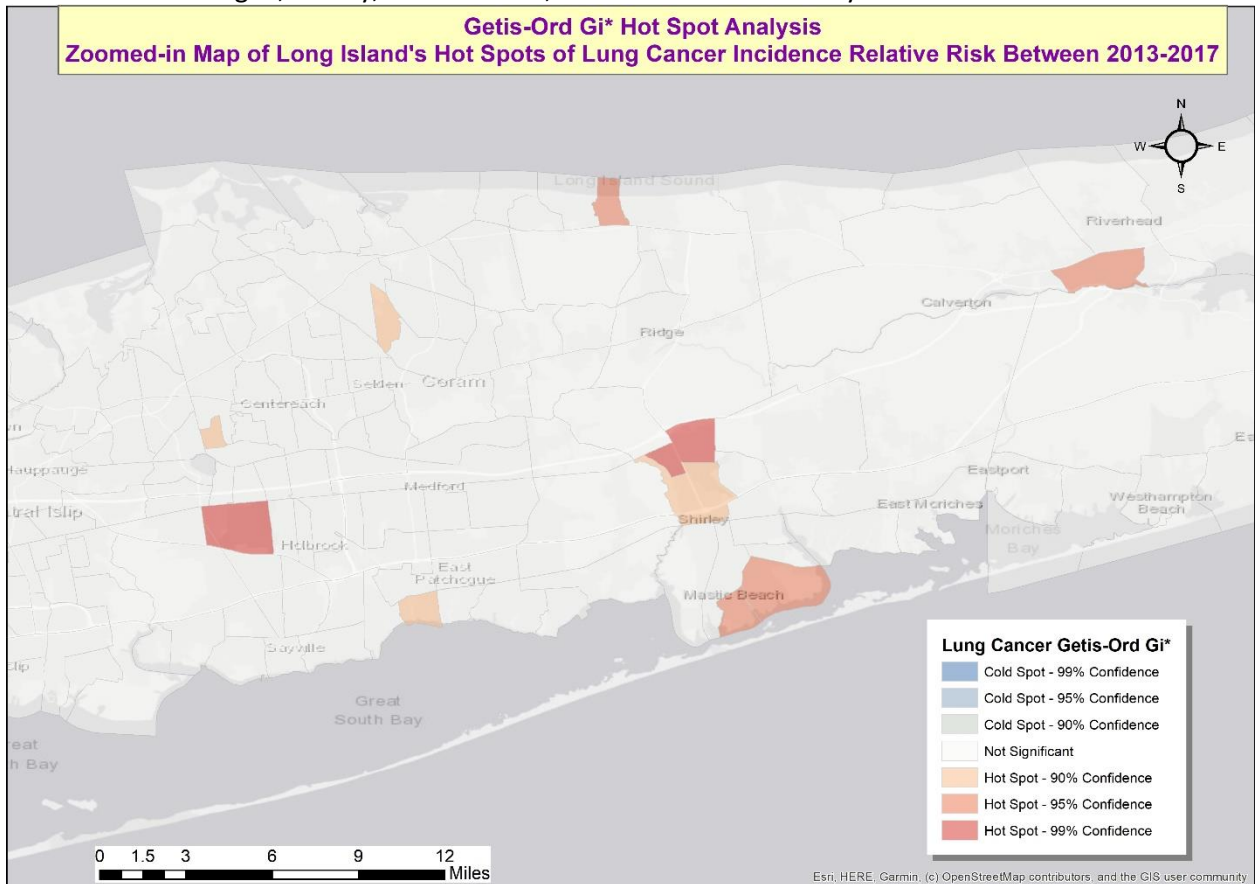
A. New York State Map



B. Long Island Map

C. Long Island Hot Spot Zoomed-in: Hotspots of lung cancer incidence are scattered spanning from Long Island MacArthur Airport, Lake Grove in the western parts to as far east as near Riverhead, as

far south as Patchogue, Shirley, Mastic Beach, and as far north as Rocky Point



D. NYC Map

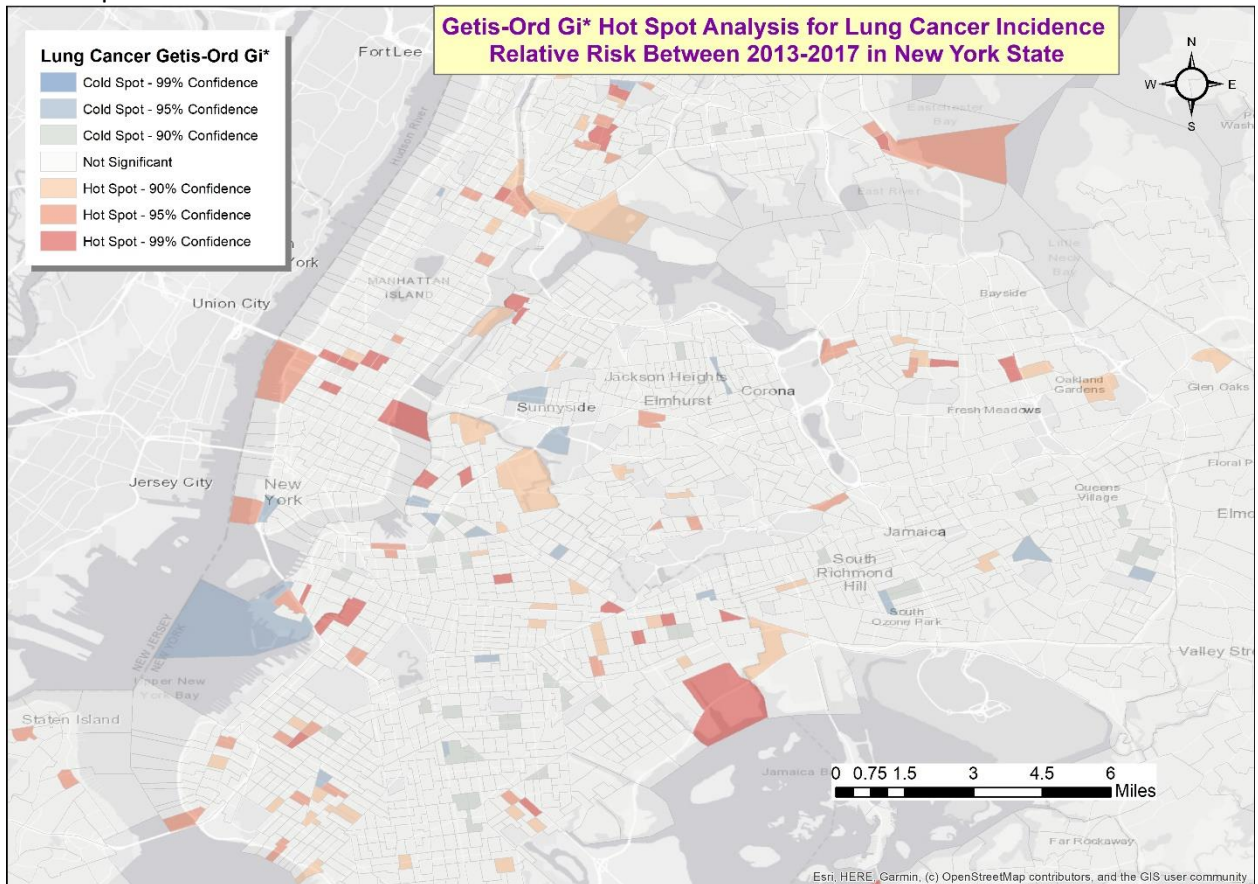

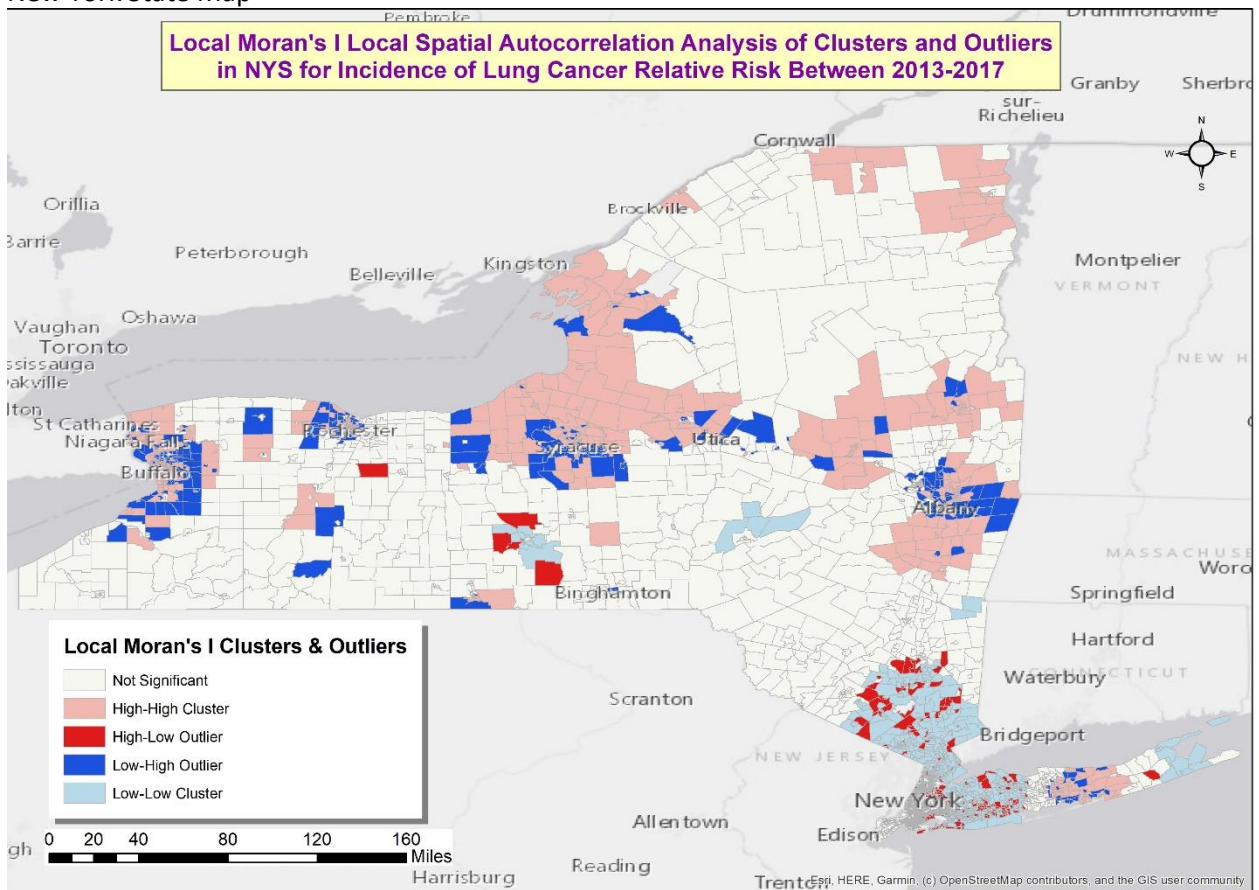


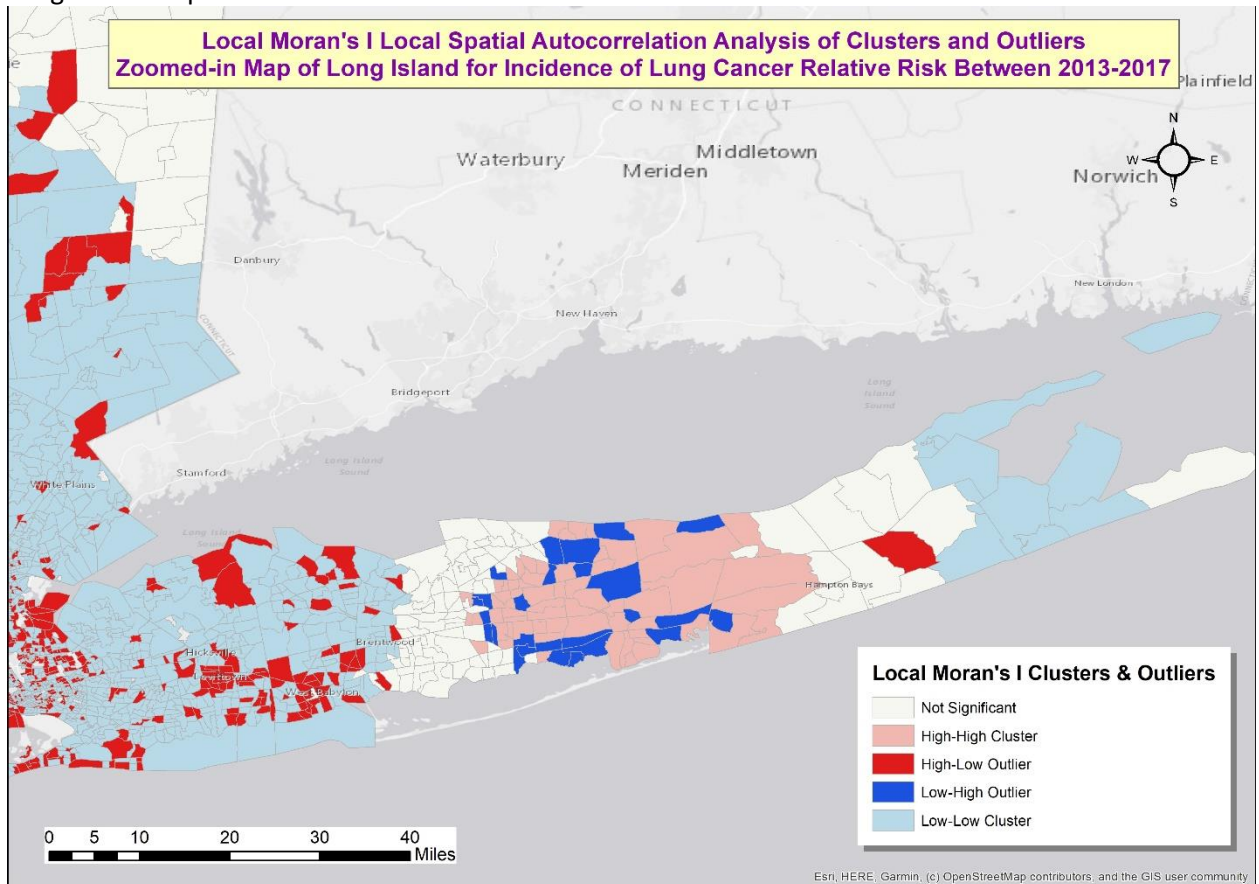
Table 2: Getis-Ord Gi*Attribute Sample Table from ArcGIS Map

Table								
								
Lung Cancer Getis-Ord Gi*								
	FID	Shape *	SOURCE_ID	Lung_RR	GiZScore	GiPValue	NNeighbors	Gi_Bin
	140	Polygon	140	0.811599	-0.441272	0.659016	58	0
	141	Polygon	141	1.13288	0.205432	0.837234	59	0
	142	Polygon	142	0.765092	-0.535227	0.592493	57	0
	143	Polygon	143	0.769023	-0.527252	0.598018	57	0
	144	Polygon	144	0.968767	-0.124899	0.900603	55	0
	145	Polygon	145	1.6963	1.342172	0.17954	182	0
	146	Polygon	146	1.61126	1.168706	0.242522	67	0
	147	Polygon	147	1.35821	0.666555	0.505057	216	0
	148	Polygon	148	1.76792	1.490189	0.136175	221	0
	149	Polygon	149	2.46062	2.886652	0.003894	173	3
	150	Polygon	150	0.937489	-0.183381	0.854499	187	0
	151	Polygon	151	0.908461	-0.242744	0.808204	149	0
	152	Polygon	152	0.909649	-0.238507	0.811488	238	0
	153	Polygon	153	0.984942	-0.083031	0.933827	247	0
	154	Polygon	154	1.07981	0.102482	0.918374	174	0
	155	Polygon	155	1.29888	0.54231	0.587605	173	0
	156	Polygon	156	1.43283	0.813146	0.416134	179	0
	157	Polygon	157	0.846657	-0.366942	0.713662	174	0
	158	Polygon	158	1.15987	0.265633	0.790522	179	0
	159	Polygon	159	1.82922	1.61234	0.106888	176	0
	160	Polygon	160	1.13185	0.209145	0.834335	179	0
	161	Polygon	161	1.27778	0.501899	0.615738	178	0
	162	Polygon	162	2.08729	2.132111	0.032998	178	2
	163	Polygon	163	1.34398	0.638486	0.523158	177	0

Figures 2A-D: Maps of Local Moran's I Spatial Autocorrelation Analysis for Lung Cancer Incidence Relative Risk **with FDR correction**.

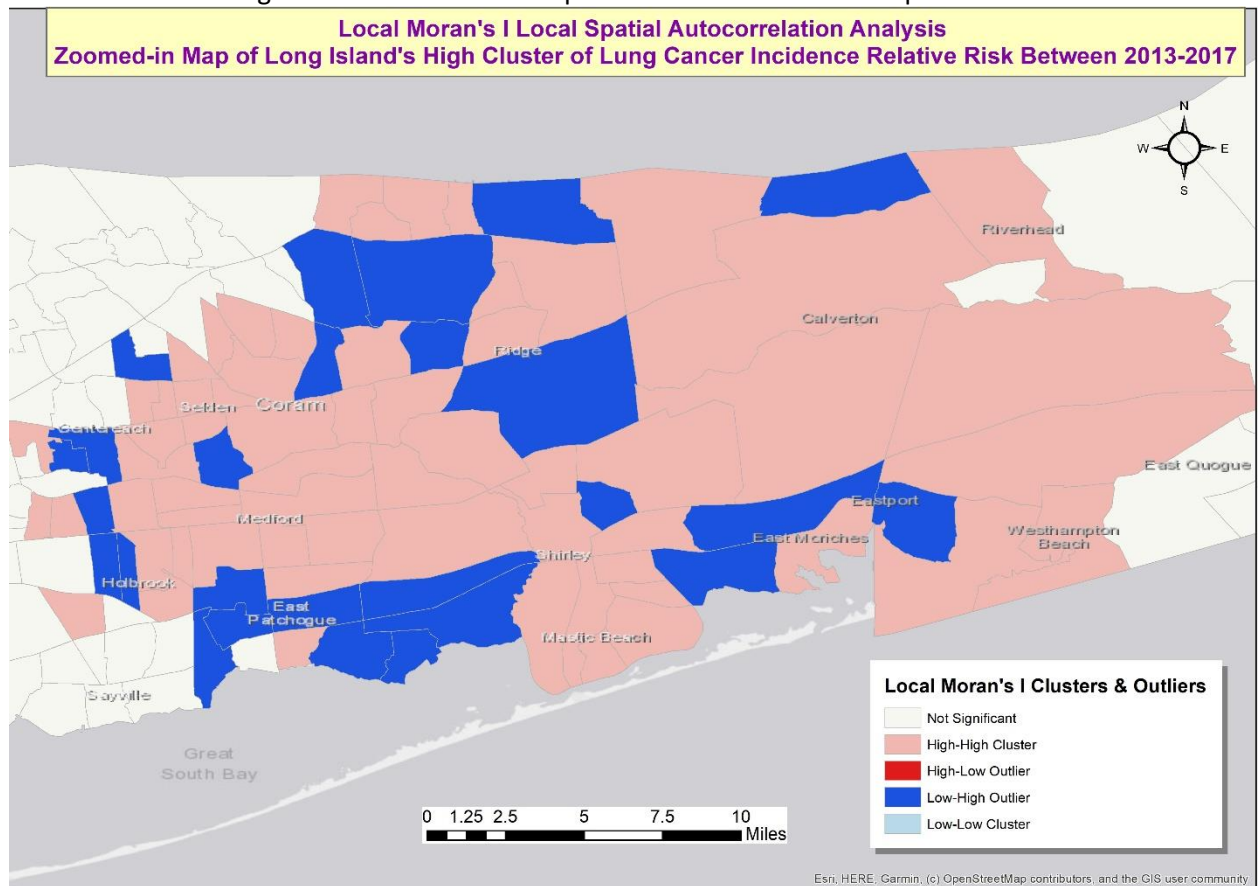
A. New York State Map



B. Long Island Map

- C. Long Island's High Cluster Map:** The high-high clusters in this map overlap with some of the hotspots in the Local Moran's I map of Long Island, indicating similar region with elevated risk for lung cancer, although Local Moran's I map's high-high cluster is more expansive and covers a lot more census

tracts in this same region than the scatter hotspots in the Getis-Ord Gi* map.



- D. NYC Map: High-high clusters are seen in Manhattan and Brooklyn but not in other boroughs (as was seen in Getis-Ord Gi* map)

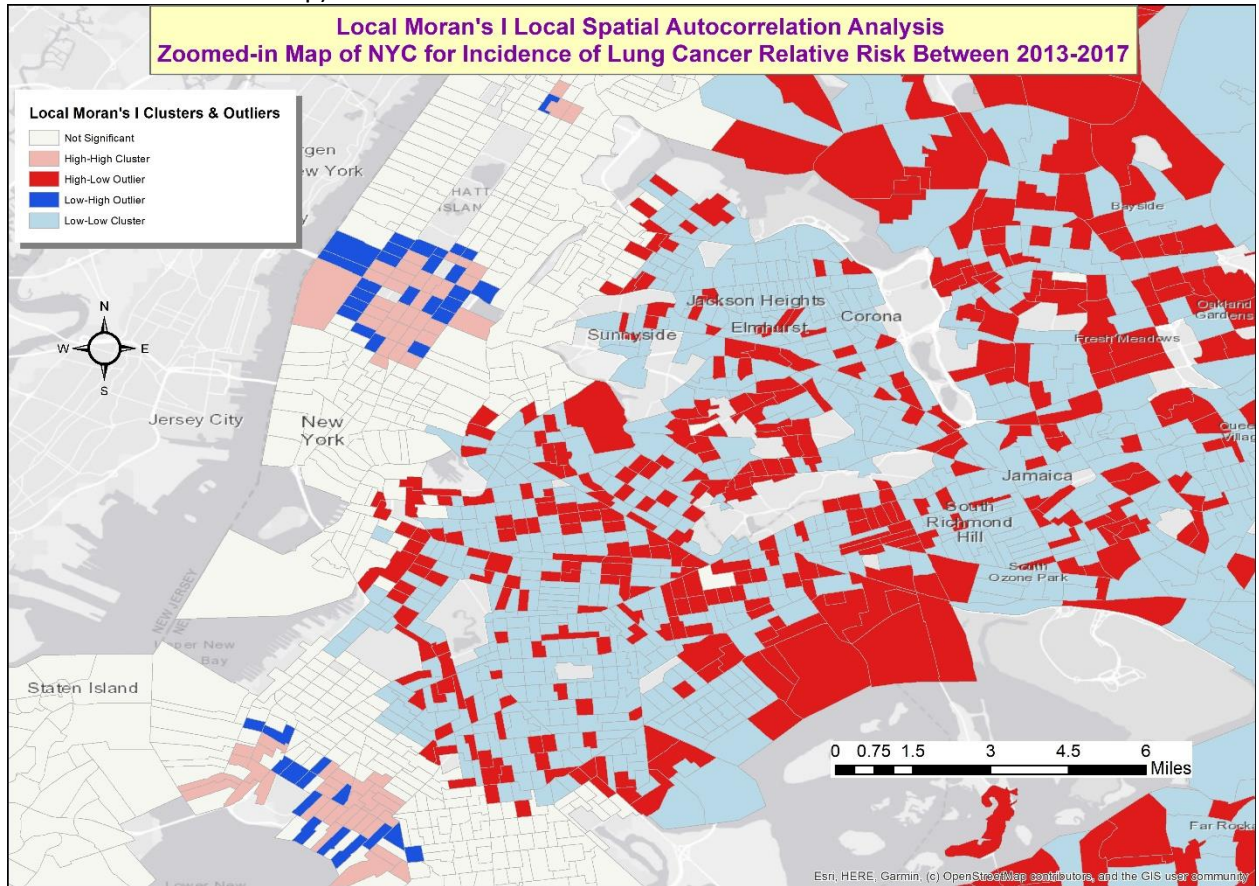
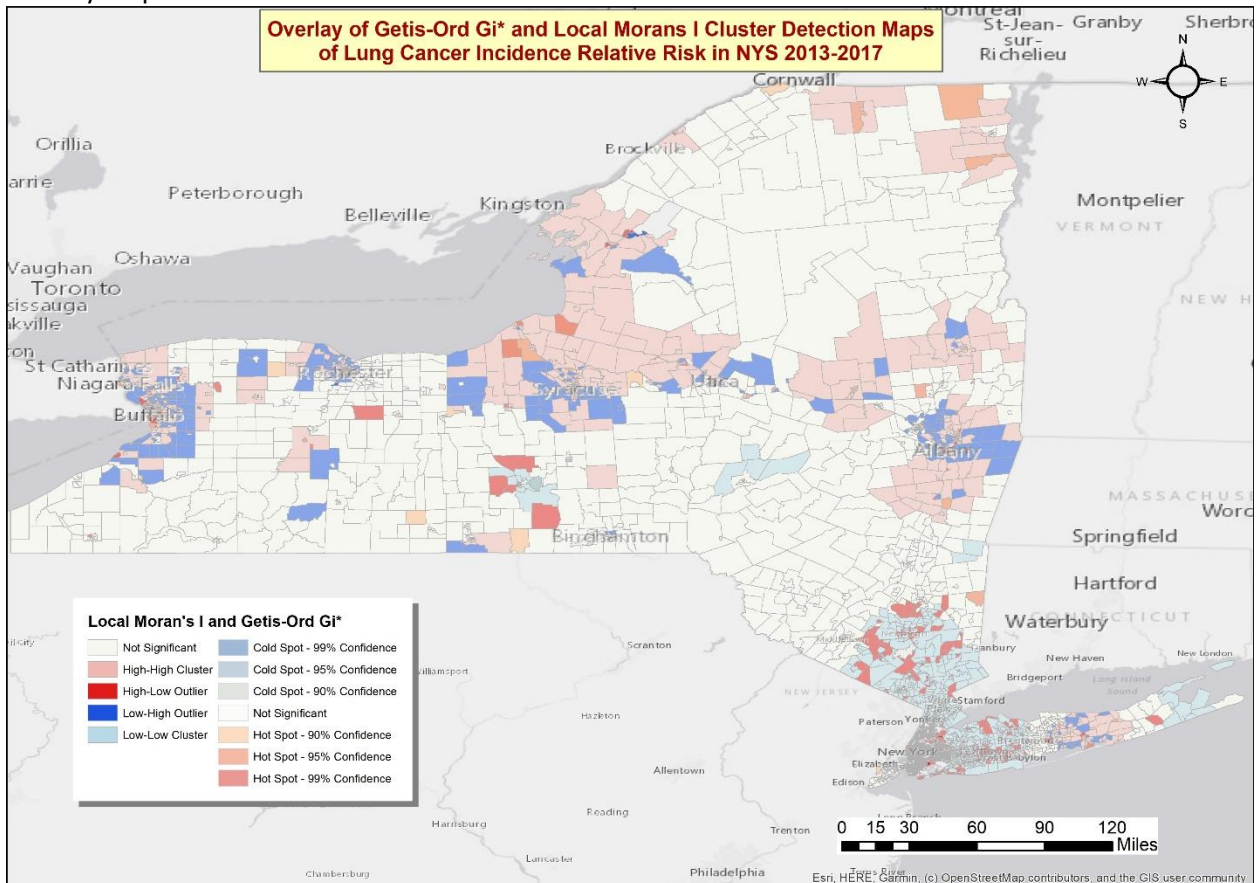


Table 3: Local Morans I Attribute Sample Table from ArcGIS Map

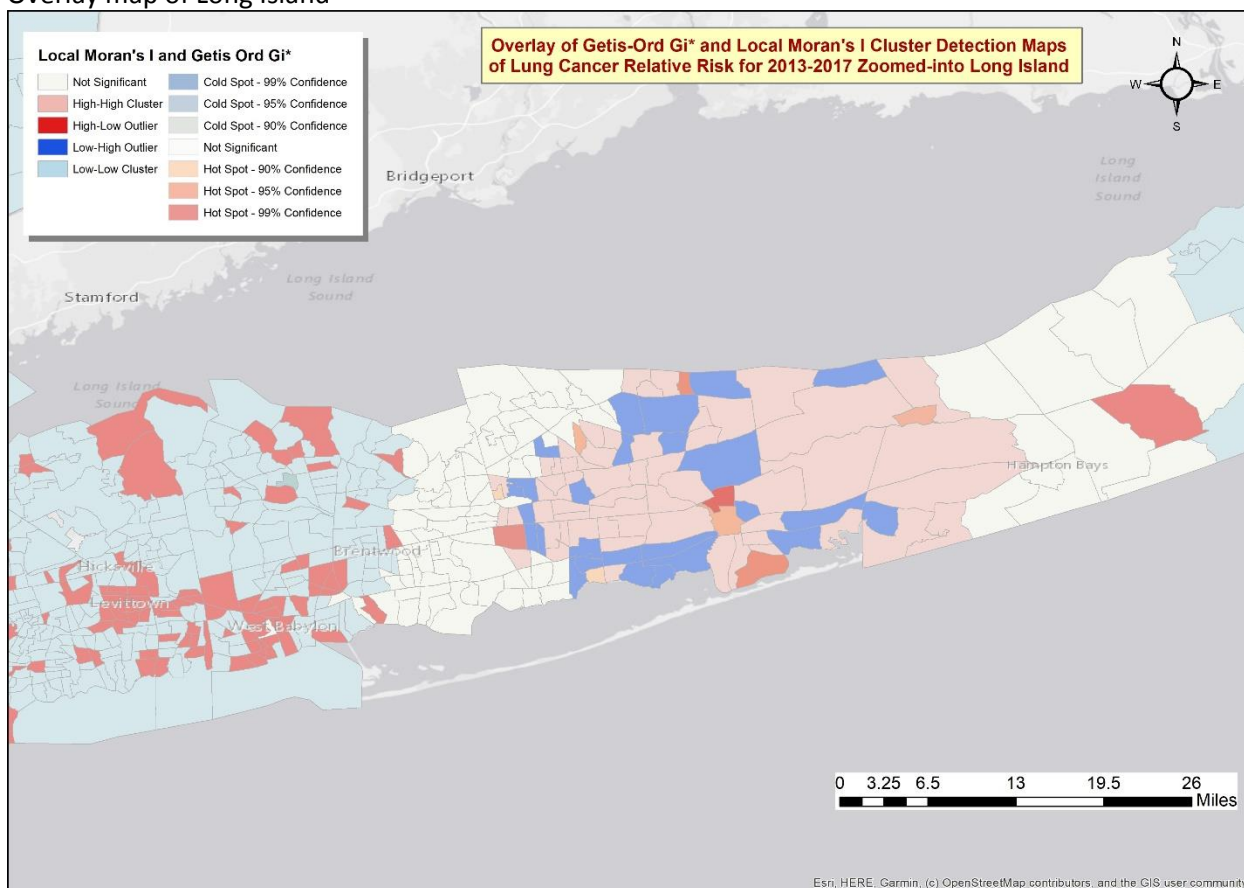
Table											
LungCancer_MoransI_ClustersOutliers											
OBJECTID *	Shape *	SOURCE	Lung_RR	Shape_Length	Shape_Area	LMiIndex IDW	LMiZScore IDW 3	LMiPValue IDW 30	COType IDW 3	NNeighbors IDW 30	
1	Polygon	0	0.690558	0.482458	0.010674	0.000287	1.621818	0.037			26
2	Polygon	1	1.37609	0.703	0.019407	0.000502	1.727057	0.049			36
3	Polygon	2	1.34933	0.244291	0.002579	0.000845	2.746504	0.01	HH		37
4	Polygon	3	0.89622	0.638527	0.017141	-0.000311	-2.855694	0.011	LH		40
5	Polygon	4	0.984569	0.727397	0.022314	0.000019	0.958132	0.149			19
6	Polygon	5	0.966511	0.773379	0.027637	-0.00004	-1.323534	0.096			27
7	Polygon	6	0.801951	0.22073	0.001812	-0.000508	-2.429309	0.021	LH		36
8	Polygon	7	0.724261	0.463428	0.011037	-0.000207	-1.271536	0.086			24
9	Polygon	8	1.34862	0.567023	0.01449	0.000938	3.210792	0.006	HH		80
10	Polygon	9	0.894144	0.04085	0.00008	-0.003882	-5.203747	0.001	LH		173
11	Polygon	10	1.20945	0.435203	0.011386	0.001016	4.230749	0.002	HH		177
12	Polygon	11	1.25313	0.029703	0.000054	0.00516	4.03678	0.003	HH		180
13	Polygon	12	1.26551	0.035003	0.000067	0.006235	4.401066	0.002	HH		177
14	Polygon	13	0.979356	0.064011	0.000094	-0.000778	-2.356313	0.017	LH		176
15	Polygon	14	0.743628	0.051122	0.000139	-0.003646	-2.481712	0.011	LH		176
16	Polygon	15	0.581342	0.063658	0.000141	-0.005509	-2.407278	0.013	LH		177
17	Polygon	16	1.58489	0.113369	0.000596	0.003468	1.77549	0.042			174
18	Polygon	17	0.468027	0.195835	0.0008	-0.003606	-2.407402	0.023	LH		169
19	Polygon	18	1.49823	0.113718	0.000285	0.005595	2.523809	0.009	HH		174
20	Polygon	19	1.4388	0.08533	0.000169	0.004172	1.923394	0.035			173
21	Polygon	20	0.988197	0.0681	0.000168	-0.000432	-2.138425	0.029	LH		174
22	Polygon	21	1.40363	0.05582	0.000137	0.005275	2.763214	0.013	HH		175
23	Polygon	22	0.957233	0.70435	0.026202	-0.000055	-1.303103	0.088			39
24	Polygon	23	0.675794	0.260965	0.003292	-0.000713	-2.101878	0.024	LH		36

Figures 3A-D: Overlay of Spatial Analysis Maps of Local Moran's I with FDR correction and Getis-Ord Gi* without FDR correction, which allows for comparisons between findings using these two methods of cluster detection. Overall, the output from both cluster detection methods lined up relatively well in a several places including in the region on Long Island spanning from Centereach to Riverhead, in which high-high cluster from the Moran's I map overlaid on the hotspots from Getis-Ord Gi*. However, there were areas with statistically significant Moran's I values but statistically non-significant Getis-Ord Gi* values, and vice-versa.

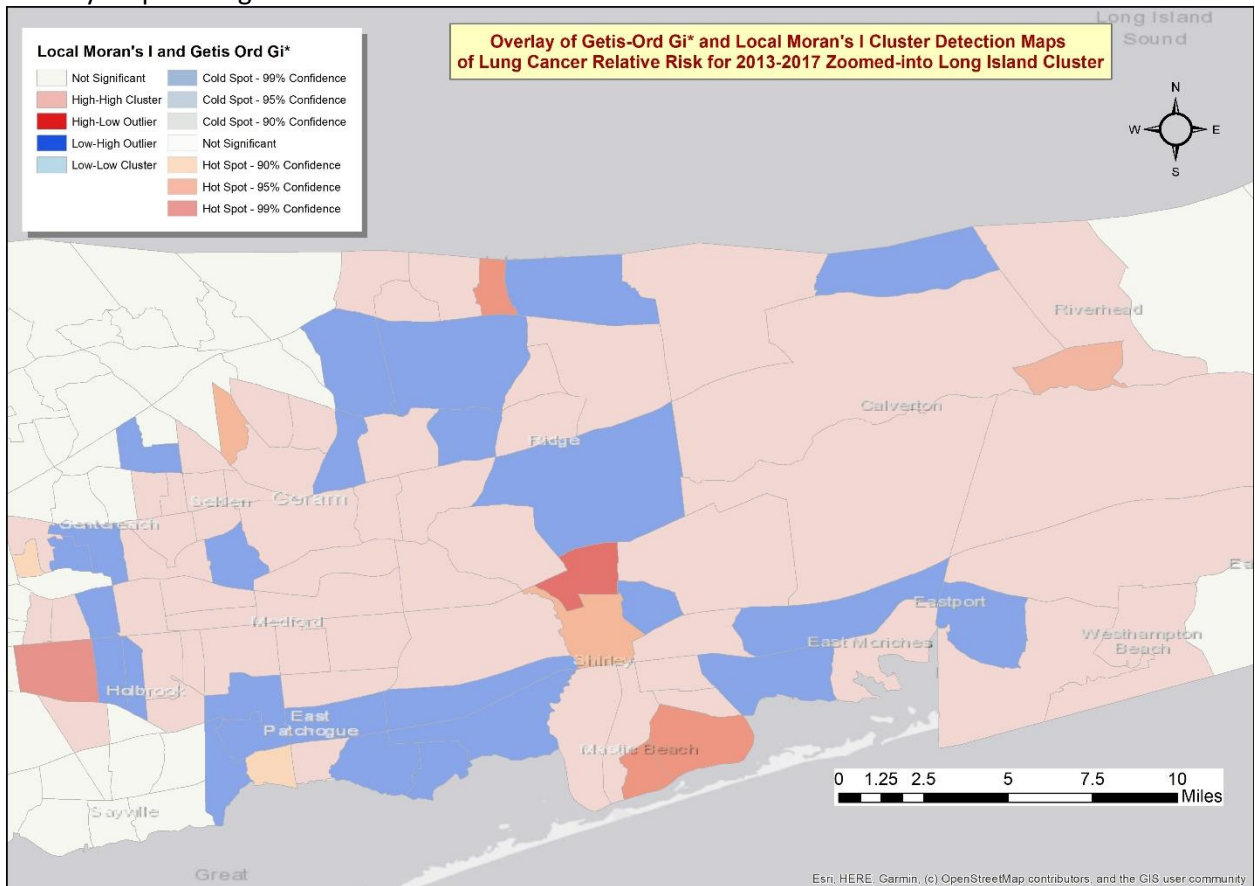
A. Overlay map of NYS



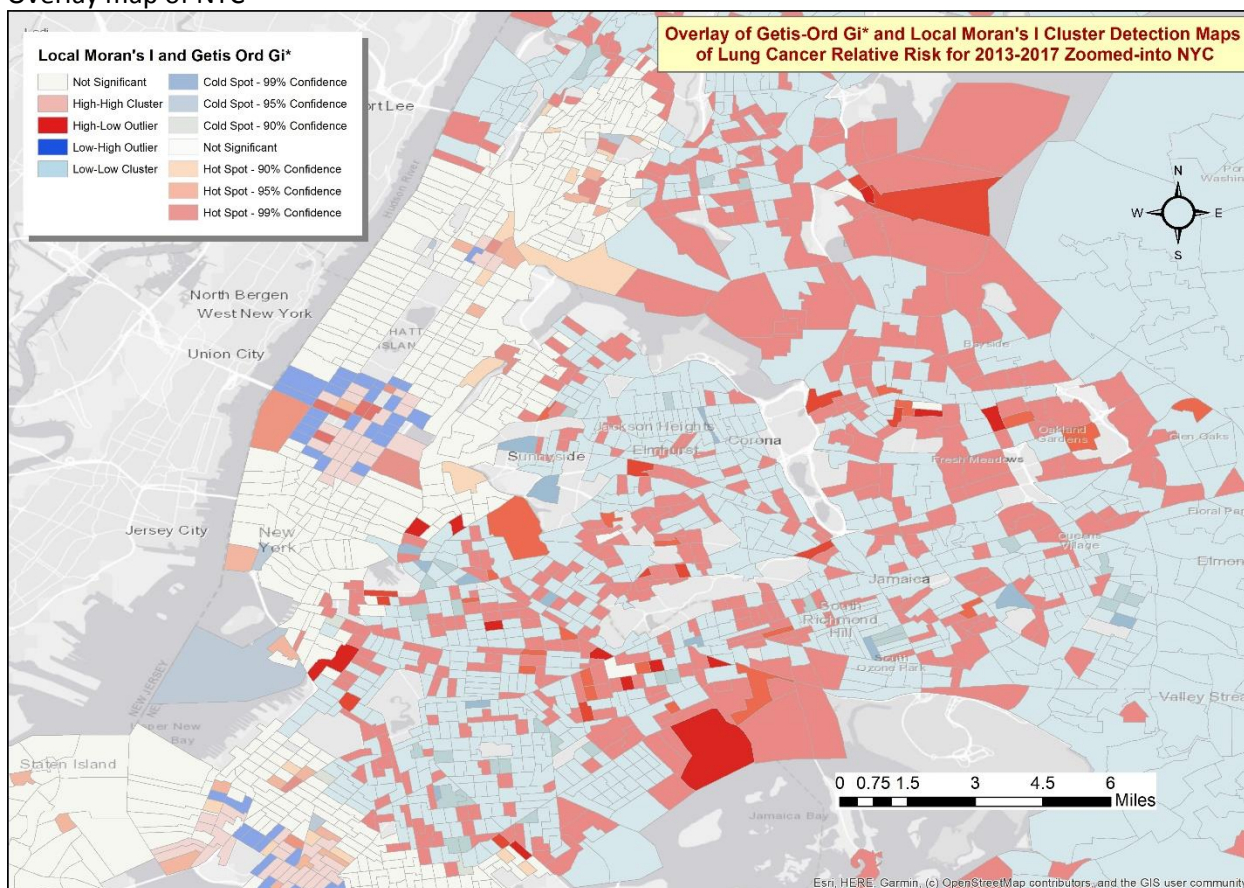
B. Overlay map of Long Island



C. Overlay map of Long Island cluster



D. Overlay map of NYC

**Table 4:** SaTScan input parameters and output

Maximum Spatial Cluster Size: 50 percent & 25 percent of population at risk

Spatial Window Shape: Circular

Isotonic Scan: No

Number of Replications: 999

Purely Spatial analysis scanning for clusters with high or low rates using the Discrete Poisson model.

Study period: 2013/1/1 to 2017/12/31

Number of locations: 4775

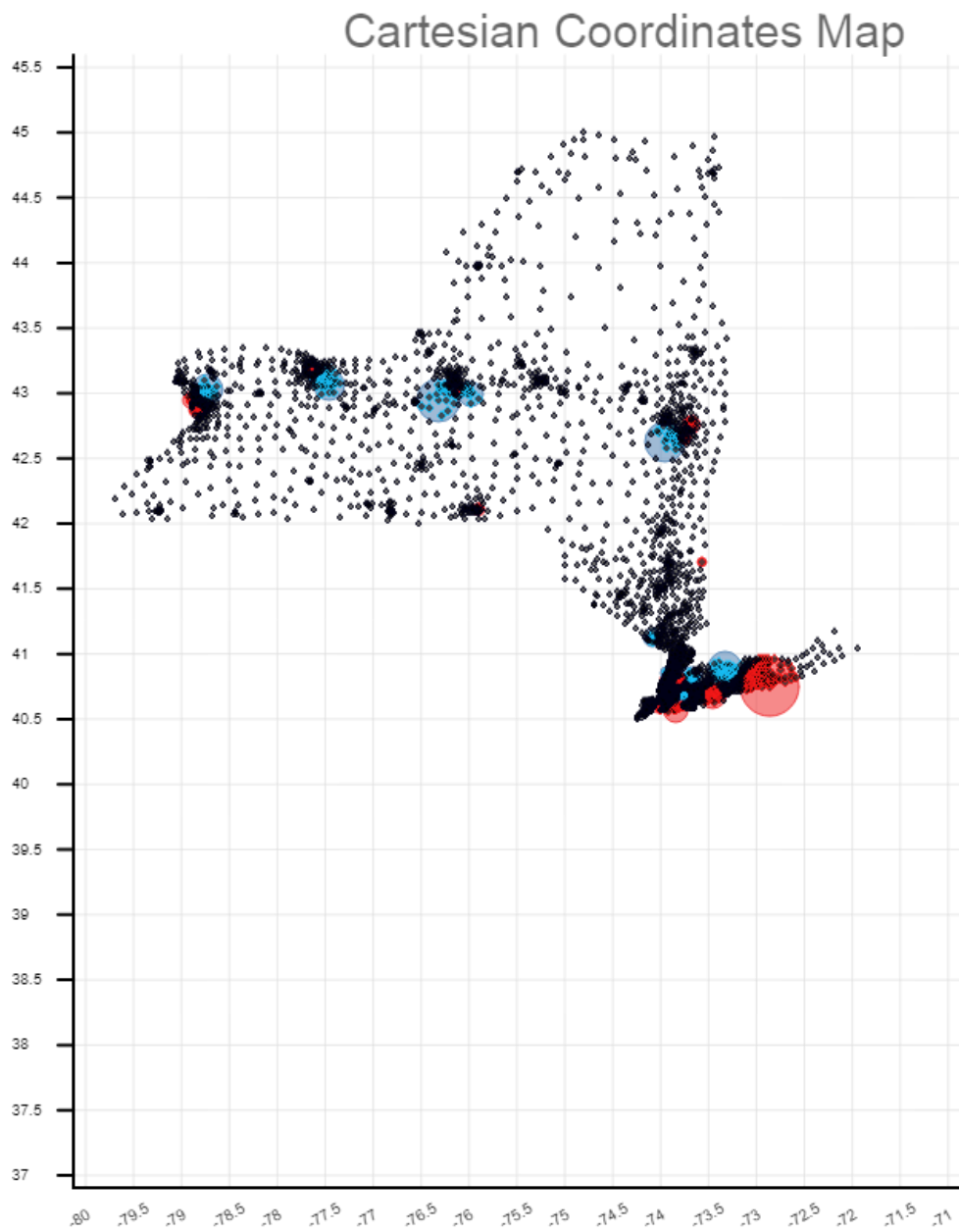
Population, averaged over time: 69789

Total number of cases: 69789

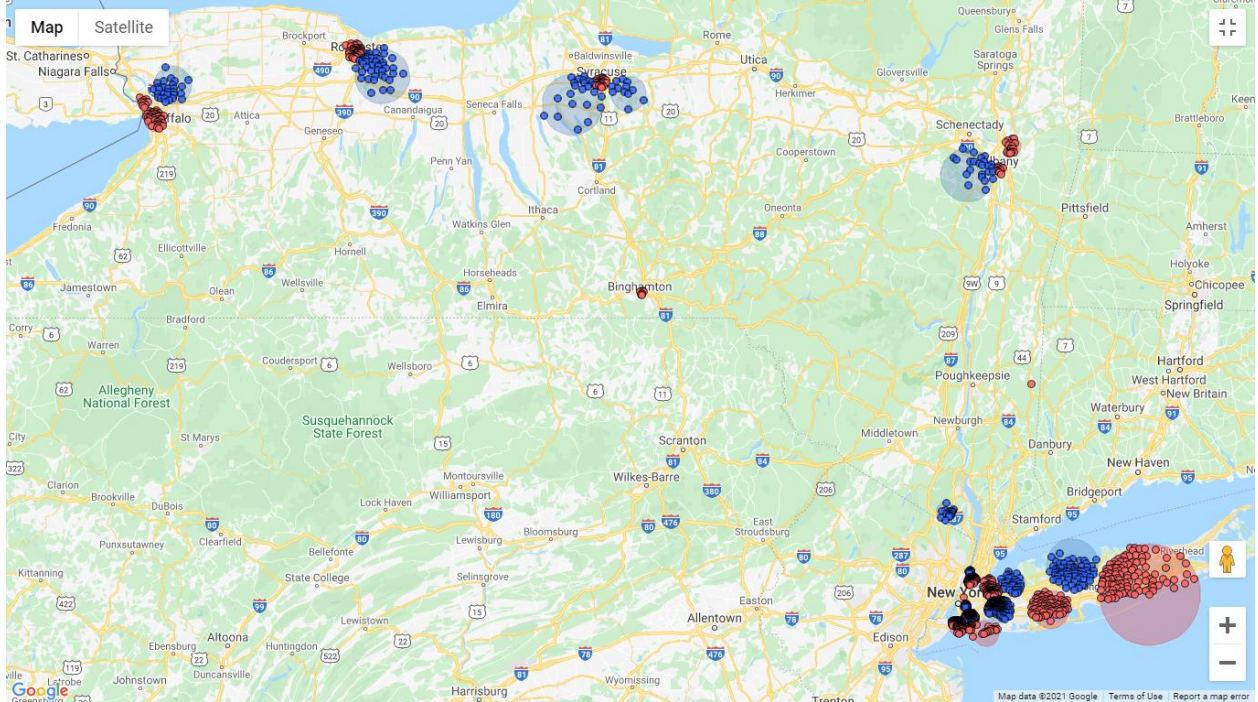
Annual cases / 100000: 20002.3

Maximum Spatial Cluster Size	25% of population at risk	50% of population at risk
Total # of Clusters (high & low)	60	54
# of Significant Clusters (high & low)	30	28
Total # of Significant Cluster Locations	1388	1305

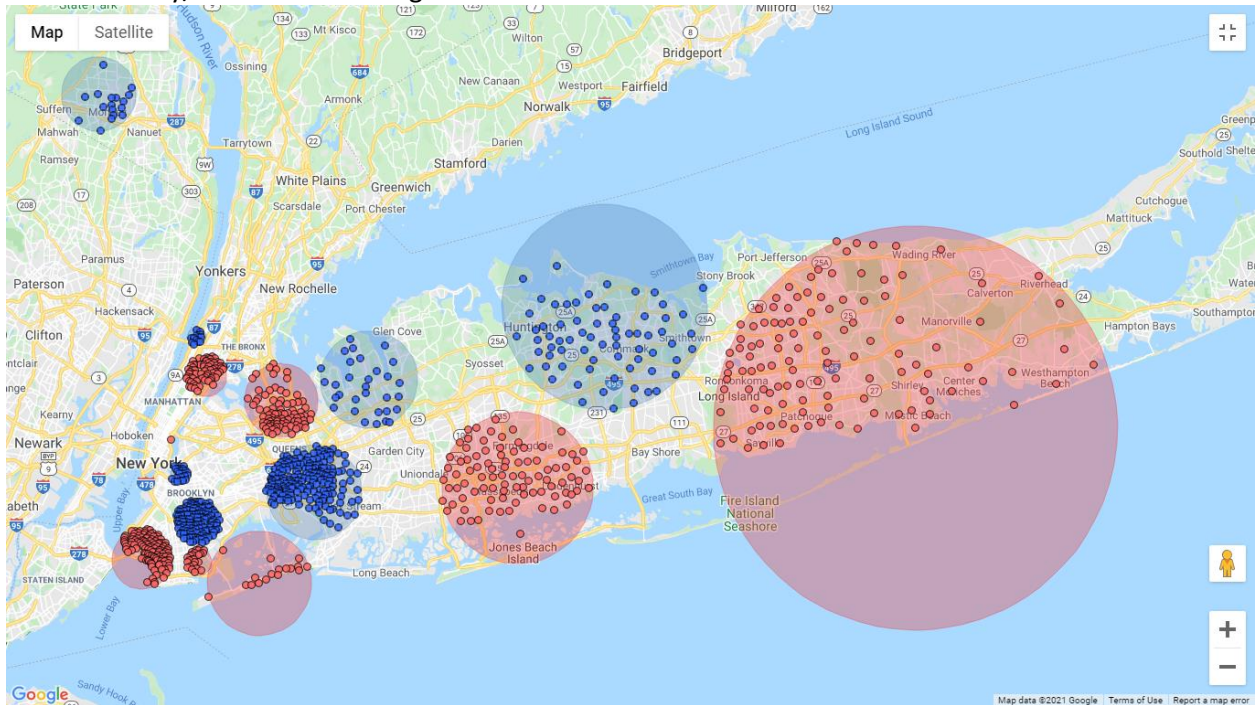
Figure 4A-C. SaTScan Spatial Analysis Maps where blue indicates statistically significant low cluster areas and red indicates statistically significant high cluster areas using circular parameter of 50% of population.

A. Cartesian Coordinate Map of NYS

- B. Google Map of NYS: The low and high clustering in SaTScan appear to congregate around population centers**

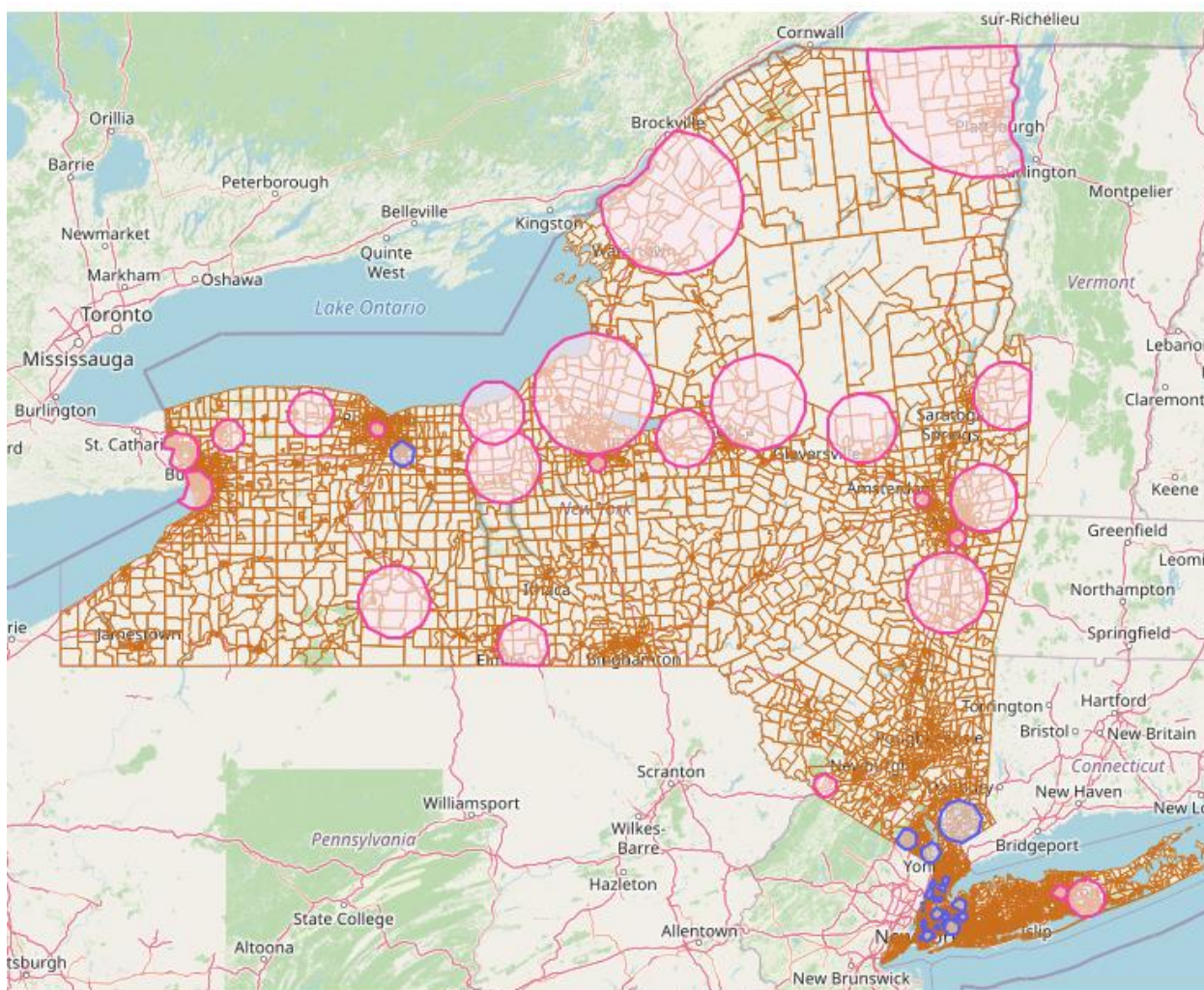


- C. Google Map of NYC and Long Island: The same spatial hot spot or high cluster for lung cancer spanning from Ronkonkoma to Riverhead is reproduced in SaTScan, indicating consistency between all three methods of cluster detection with respect to this Long Island cancer cluster. However, SaTScan also appears to detect a smaller high cluster in the Farmingdale and Lindenhurst region in Nassau County/southwestern Long Island.**



Figures 5A-B. Maps taken from the NYS Department of Health's (NYSDoH) cancer mapping website using lung cancer incidence data from 2011-2015 (Health, 2018) where the cancer highlighted areas in pink indicate higher than expected and those in blue indicate lower than expected rates. Comparing our results using data from 2013-2017 to the NYSDoH's results indicated that some similarities such as the higher-than-expected cluster in middle-eastern part of Suffolk County remained consistent over the periods from 2011 – 2015 to 2013-2017. However, there were some hotspots that were detected in NYC in the current 2013-2017 analysis that were not evident in the 2011-2015 NYSDoH's map.

A. NYSDoH's Lung Cancer Map of NYS



B. NYSDoH Cancer Map Zoomed-in to Long Island

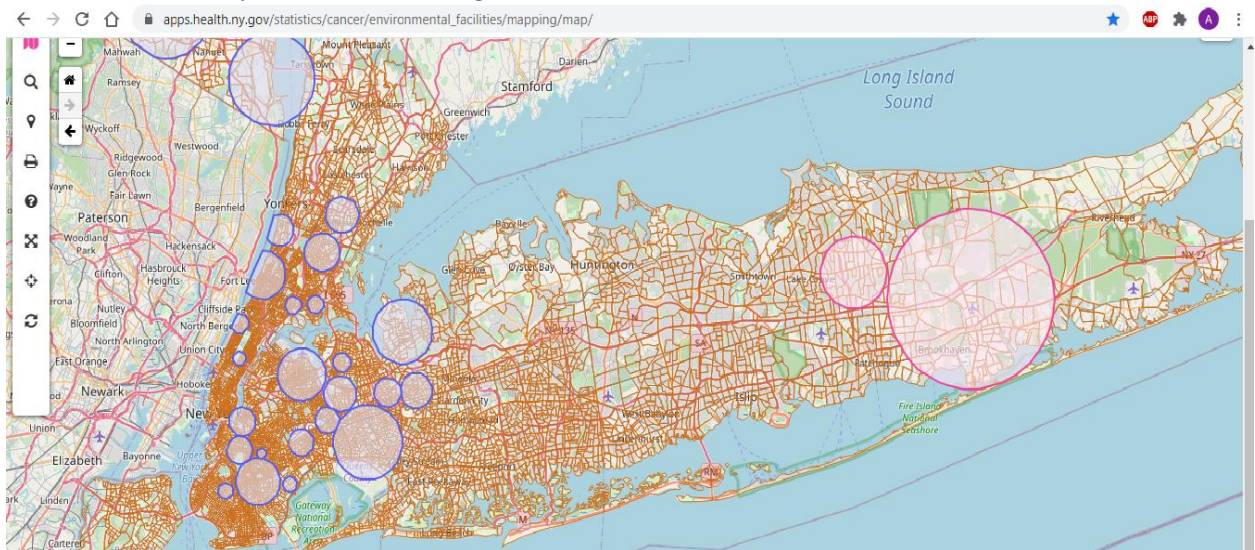
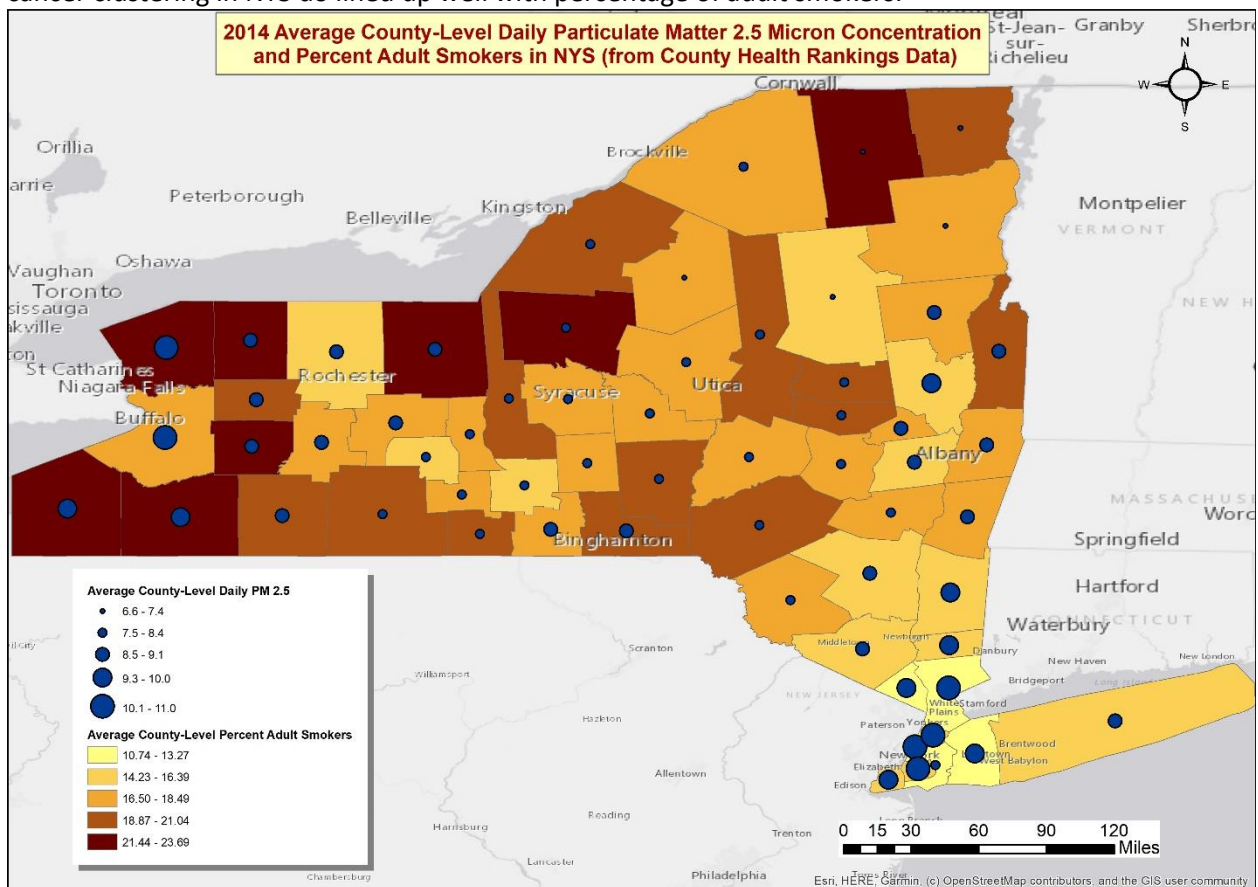


Figure 6: NYS Map of PM 2.5 Levels and Adult Smoking Rate at the County-Level. Some of the high lung cancer clustering in NYS do lined up well with percentage of adult smokers.



Results Summary

There appeared to be consistency in the findings of all three methods of cluster detection with respect to the Long Island lung cancer cluster or hotspot spanning from Ronkonkoma to Riverhead in Suffolk County. This Long Island lung cancer cluster using data from 2013-2017 also aligned well with the lung cancer cluster identified using the 2011-2015 data in the NYSDoH map. The etiology or causation behind this cluster in Suffolk County Long Island appeared unclear since the adult smoking rate here (15%) was close to the NYS average rate (14%) and was actually much lower than certain parts of upstate NY at smoking rates of 20+ percent (Roadmaps, 2019). With respect to New York City (NYC), all three methods (Getis-Ord G_i^* , Local Moran's I , SaTScan's Kulldorff statistic) identified hotspots of lung cancer in Mid-Lower Manhattan and South Brooklyn from the 2013-2017 dataset. In contrast, the NYSDoH map did not identify these hotspot areas using data from 2011-2015; in fact, there appeared to be no lung cancer hotspot in NYC from the NYSDoH map. Looking at the entire NYS map overall using all three cluster detection methods, the areas of higher-than-expected lung cancer incidence appeared to match up well, including in the northeast corner of upstate NY, Albany, Syracuse, Rochester, and Buffalo regions; these hotspot areas were also reflected in the NYSDoH map using data from 2011-2015. Some but not all the areas with the highest smoking rates and PM 2.5 levels corresponded with higher incidence of lung cancer relative risk.

Discussion

This study is the first work attempting to use the most recent Census Tract level evidence to determine lung cancer clustering within New York State (NYS) and to analyze the correlation between certain risk factors (e.g., smoking and air pollution) and lung cancer incidence.

Our work has demonstrated significant lung cancer hot spots in Getis-Ord G_i^* analyses around populational aggregation centers in upstate New York, including Albany, Syracuse, and Buffalo. Similar patterns were detected in local Moran's I analyses, with extra high-high clusters identified around Rochester, Plattsburgh, Utica-Saratoga Springs, and Suffolk county. Furthermore, local Moran's I

exhibited noticeable low-low clusters of lung cancer cases around New York City and in Nassau county. Although the results from local Moran's I and Getis-Ord Gi* lined up with each other in many locations, there were still discrepancies. One possible explanation for such difference could be that for the former, only neighboring features (lung cancer incidence in our case) around the study area were analyzed, while the latter included the features within the study area to generate the model. For instance, local Moran's I detected many low-low clusters in Brooklyn and Queens, while Getis-Ord Gi* showed significant hot spots in the same areas. And according to the key difference between the two models, we can tell that the hot spots detected by Getis-Ord Gi* could be the "true" center since it targeted all data within and around, and a feature with a very high value would show up as a hot spot regardless of the surrounding features. On the contrary, local Moran's I excluded features within the study area and therefore, when it showed low-low clusters, it could be because the features in the study area are relatively high (a hot spot) and the analysis will always detect low value in the neighborhood. In such cases, we could expect discrepancies to be observed; therefore, an important lesson learned is that we should cautiously examine both the global and local distribution before we make any conclusions.

We also observed differences in the maps when using SaTScan as compared to ArcMap. The clusters detected by SaTScan were smaller than those generated from local Moran's I and we even observed contradictory results – low-low clusters in local Moran's I in north Queens and high clusters in Kulldorff's spatial scan statistics at the same region. SaTScan determines whether there is clustering or not and then gives the location, size of the cluster. LISAs provides a full map with each region colored by how similar its value is to the values of its neighbors. With the second technique, we were able to predict clustering and outliers, but this will also lead to multiple comparisons and inaccurate results (in our case, we fixed this problem with FDR correction). The methodologies in modeling are different; therefore, it is not surprising to observe slightly different results. In terms of choosing an appropriate analysis tool for a certain study, it is not always a right or wrong answer but depends on which model is

a good fit to the objectives – whether the research focuses more on central clustering or neighborhood features/outliers.

Our regression model showed that smoking was the only predictor for lung cancer and this result is consistent with the fact that smoking is the most important behavioral risk factor for lung malignancies. However, the Behavioral Risk Factor Surveillance System (BRFSS) survey collects only data at the individual level, which is less satisfactory and reliable than the Census Tract data. Moreover, when using aggregate mean (the percentage of smokers) as a surrogate in the analyses, we are at risk of losing details at various geographic locations within each county. Although the sample size of BRFSS in NYS was relatively large (906,420), the overall response rate was only 36.3%. Amongst these responded individuals, they are generally more concerned about their health conditions, which means that with higher awareness, smoking-related diseases – COPD and lung cancer are more likely to be detected among them as compared to non-responders. This could substantially lead to a selection bias in our work.

Surprisingly, air pollution, specifically PM_{2.5} level, was not correlated with the relative risk for lung cancer incidence. Several possibilities could have contributed to this result. Firstly, the air pollution in each county in NYS was not severe and in fact, the average PM_{2.5} level was only 8.5 ug/m³ across the state in 2014, which equals to a “good” category in Air Quality Index system – the highest possible class that we could achieve. This low PM_{2.5} level might be less influential to lung cancer development compared to smoking; while in countries like China, severe air pollution with extremely high particulate matter readings caused by fossil fuel usage is more likely to be a contributor to lung cancer. Secondly, the PM_{2.5} data were retrieved from the Centers for Disease Control and Prevention at the county level. However, the website did not indicate the exact locations of data collection and whether these locations were randomly distributed in and representative for each county. Similarly, we are not sure if the numbers were average through consistent observations at each location over a calendar year.

Therefore, the accuracy and reliability are compromised, although this was the best option after our comprehensive research on currently available datasets. Finally, the common sense is that there is a long period of latency between exposures and the development of malignancy. Exposure data (both smoking status and air pollution in our case) collected decades before the diagnosis of lung cancer should be used in the analyses. However, NYS county ranking data before 2013 was not available; therefore, we could only assume that the smoking status and air particulate matter remains unchanged over time, which is not perfectly true in the real-world considering migrations, car ownership, and industrial structure change in the past three decades.

In conclusion, our study has identified significant hot spots of lung cancer incidence across NYS, which could serve as a guidance for resource relocation and optimization to achieve better patient management and clinical outcomes. Moreover, this work has validated that smoking is a major contributor to lung cancer, supporting a more restrictive regulation on cigarettes and further public education in smoking cessation.

References

- Boscoe, F. P., McLaughlin, C., Schymura, M. J., & Kielb, C. L. (2003). Visualization of the spatial scan statistic using nested circles. *Health Place*, 9(3), 273-277. doi:10.1016/s1353-8292(02)00060-6
- Bureau, U. C. (2010). TIGER/Line Shapefiles. Retrieved from <https://www.census.gov/cgi-bin/geo/shapefiles/index.php>
- Dirk U. Pfeiffer, T. P. R., Mark Stevenson, Kim B. Stevens, David J. Rogers, and Archie C. A. Clements. (2008). *Spatial Analysis in Epidemiology*: Oxford University Press.
- Health, N. Y. S. D. o. (2018). Environmental Facilities and Cancer Map Data from 2011-2015. Retrieved from https://apps.health.ny.gov/statistics/cancer/environmental_facilities/mapping/map/
- Henschke, C., Boffetta, P., Gorlova, O., Yip, R., DeLancey, J., & Foy, M. (2010). Assessment of lung-cancer mortality reduction from CT Screening. *Lung Cancer* (Amsterdam, Netherlands), 71(3), 328–332. <https://doi.org/10.1016/j.lungcan.2010.10.025>
- JAMA. Proportion of never smokers among men and women with lung cancer in 7 US states. *JAMA Oncology*, 7(2), 302-304. doi:10.1001/jamaoncol.2020.6362
- Juster, T. (2017). Trends in Lung Cancer Mortality by County, New York State 1994-2013 [map]. New York State Department of Health. Accessed from the Centers for Disease Control and Prevention's Chronic Disease GIS Exchange <http://www.cdc.gov/dhdsp/maps/gisx/mapgallery/>
- Li, R., Zhou, R., & Zhang, J. (2018). Function of PM2.5 in the pathogenesis of lung cancer and chronic airway inflammatory diseases. *Oncology Letters*, 15(5), 7506–7514. <https://doi.org/10.3892/ol.2018.8355>
- McDowell, S. (2020). Study: More Than 12% of People Newly Diagnosed with Lung Cancer Never Smoked Cigarettes. Retrieved from <https://www.cancer.org/latest-news/study-more-than-twelve-percent-of-people-newly-diagnosed-with-lung-cancer-never-smoked.html>

National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health. (2014). The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General. *Centers for Disease Control and Prevention* (US)

NYSR. (2017). Cancer Incidence by Census Tract. Retrieved from
<https://www.health.ny.gov/statistics/cancer/registry/tract/index.htm>

NYS Department of Environmental Conservation. (n.d.). PM2.5 monitoring. *NYS Department of Environmental Conservation*. <https://www.dec.ny.gov/chemical/8539.html>

NYS Department of Health. (n.d.). About lung cancer. *New York State Department of Health*.
<https://health.ny.gov/statistics/cancer/registry/abouts/lung.htm>.

Roadmaps, C. H. R. (2019). New York Rankings Data. Retrieved from
<https://www.countyhealthrankings.org/app/new-york/2019/downloads>

SaTScan. (2005). SaTScan: Software for the spatial, temporal, and spacetime scan statistics. Retrieved from <https://www.satscan.org/>

Siegel, D. A., Fedewa, S. A., Henley, S. J., Pollack, L. A., & Jemal, A. (2021). Proportion of Never Smokers Among Men and Women With Lung Cancer in 7 US States. *JAMA Oncology*, 7(2), 302-304.
doi:10.1001/jamaoncol.2020.6362

Wang, H., Li, J., Gao, M., Chan, T., Gao, Z., Zhang, M., Li, Y., Gu, Y., Chen, A., Yang, Y., & Ho, H. (2020). Spatiotemporal variability in long-term population exposure to PM2.5 and lung cancer mortality attributable to PM2.5 across the Yangtze River Delta (YRD) region over 2010–2016: A multistage approach. *Chemosphere* (Oxford), 257, 127153–127153.
<https://doi.org/10.1016/j.chemosphere.2020.127153>

Wei, H., Liang, F., Cheng, W., Zhou, R., Wu, X., Feng, Y., & Wang, Y. (2017). The

mechanisms for lung cancer risk of PM2.5: Induction of epithelial-mesenchymal transition and cancer stem cell properties in human non-small cell lung cancer cells. *Environmental Toxicology*, 32(11), 2341–2351. <https://doi.org/10.1002/tox.22437>

Wolfe, A., Bornstein, K., Lee, Y., Mirsaeidi, M., & Holt, G. (2019). Relationship between PM2.5 levels and lung cancer incidence and mortality by US states. *Chest*, 156(4), A405–A405. <https://doi.org/10.1016/j.chest.2019.08.442>