Leukemias & Lymphomas Analysis in Illinois

Gabi Capone

8 March 2023

The leukemia and lymphoma data comes from Illinois Department of Public Health's Illinois State Cancer Registry via the Illinois State Cancer Registry.

The Illinois Zip Code data comes from the American Community Survey, part of the US Census Bureau.

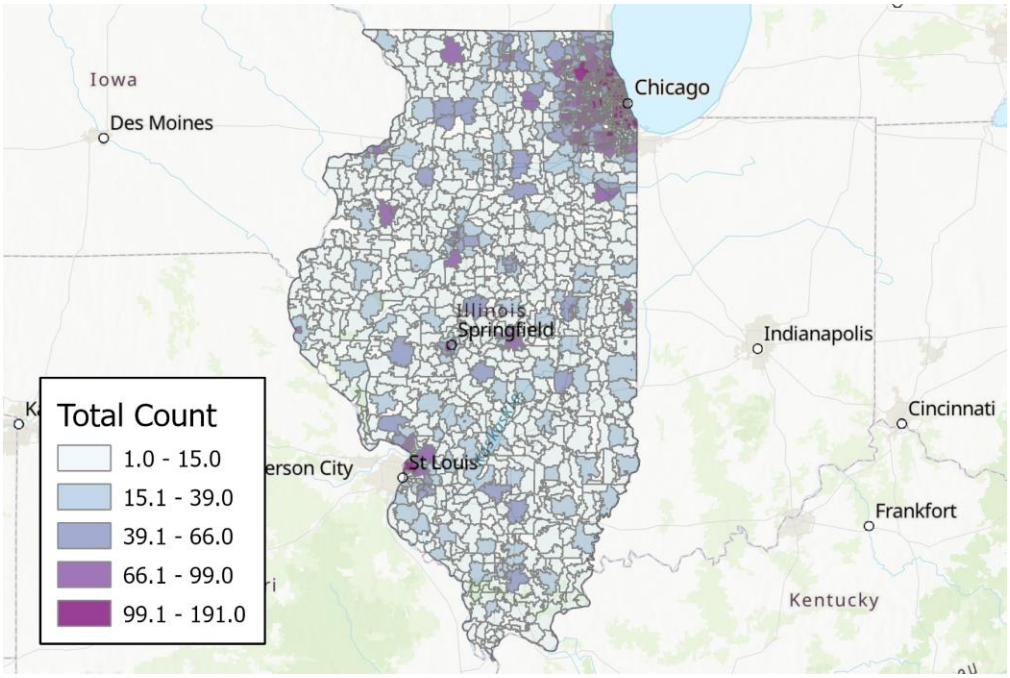

Figure 1: Leukemias and lymphomas counts by Illinois zip code 2015-2019

“Of greater interest in epidemiology is the rate of cancer rather than the absolute count, since we usually want to know where there are more cancer cases than you would expect for the size of the population” (https://michaelminn.net/tutorials/arcgis-pro-regression/index.html)

There appears to be a slightly higher crude rate of leukemia and lymphoma per 100,000 people in the southwest part of Illinois.

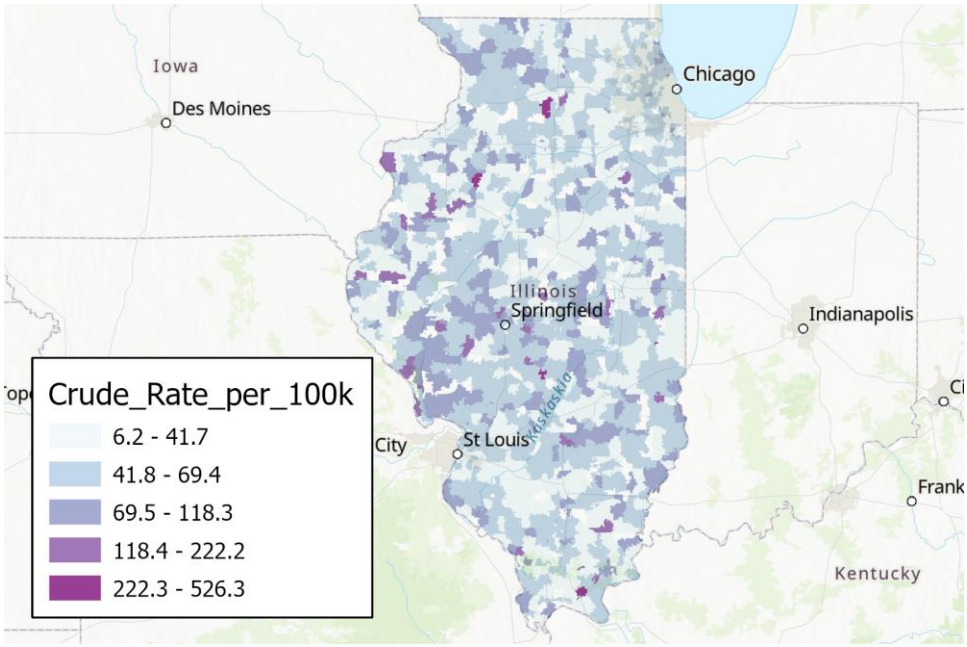

Figure 2: Leukemias and lymphomas rate of cancer per 100,000 residents by Illinois zip code 2015-2019

“The US EPA's Toxics Release Inventory (TRI) is a program that collects information reported annually by U.S. facilities in different industry sectors about quantities of toxic chemical released to the air, water, or land disposal, and/or managed through recycling, energy recovery and treatment” (https://michaelminn.net/tutorials/arcgis-pro-regression/index.html).

There appears to be a high concentration of carcinogenic releases in the Chicagoland area, and a few significant concentrations in Southern Illinois.

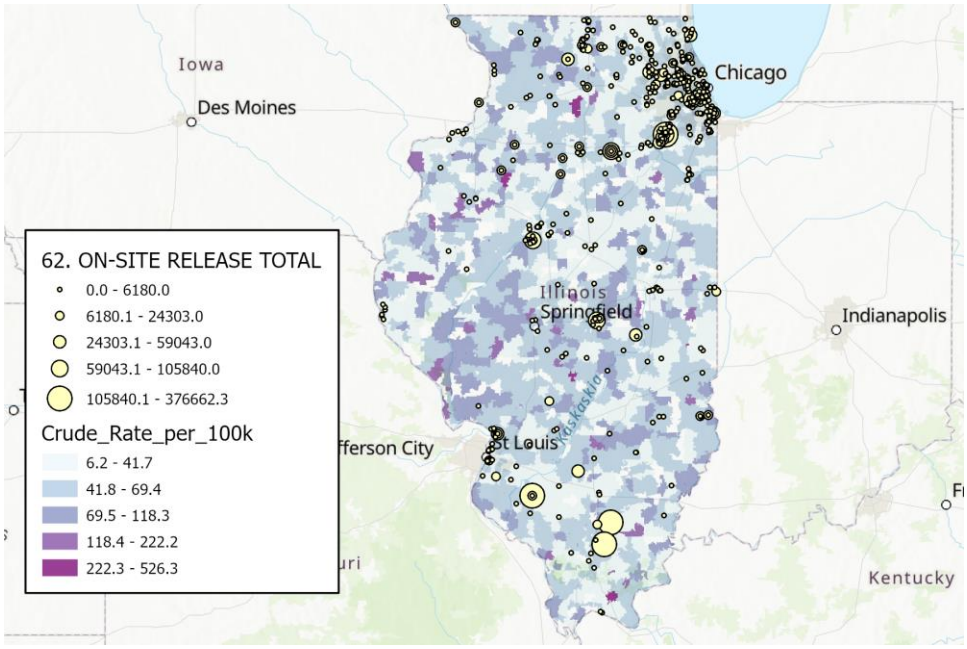

Figure 3: Carcinogenic emissions in Illinois, 2021

“In this exploratory analysis, we will make an arbitrary distance decay assumption that living within five kilometers of a release source presents the possibility for exposure” (https://michaelminn.net/tutorials/arcgis-pro-regression/index.html).

“These choices operate on an assumption that toxics disperse evenly around the facility and that all different carcinogens are equally carcinogenic” (https://michaelminn.net/tutorials/arcgis-pro-regression/index.html).

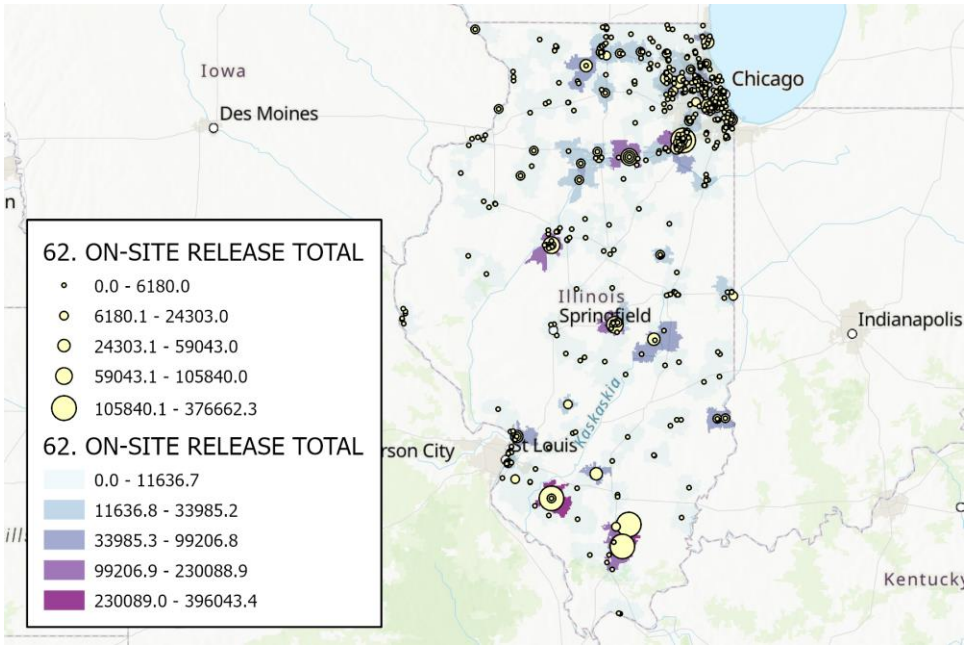

Figure 4: Estimated exposure to carcinogenic emissions by zip code in Illinois, 2021

“Unlike simple observation or techniques like kernel density analysis, the Getis-Ord GI* algorithm uses statistical comparisons of all areas to create *p-values* indicating how probable it is that clusters of high values (hot spots) or clusters of low values (cold spots) in specific areas could have occurred by chance” (https://michaelminn.net/tutorials/arcgis-pro-regression/index.html).

There is a higher quantity of cold spots in the Chicagoland region, while there is a significant cluster of hotspots just southwest of Springfield.

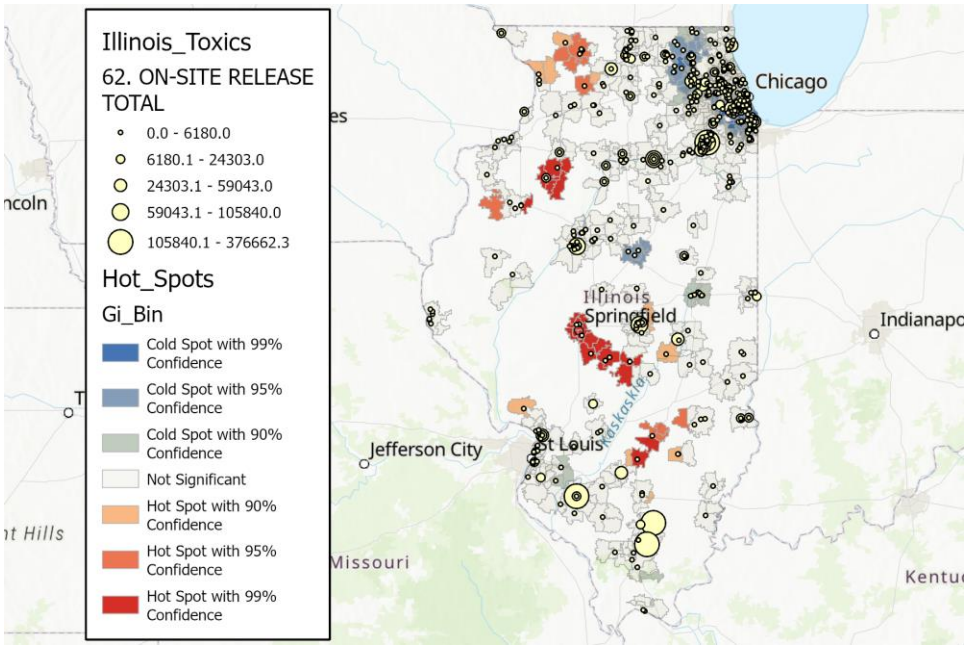

Figure 5: Hot spot map of leukemia and lymphoma cancer rates in Illinois, 2015 – 2019

"Correlation is 'a relation existing between phenomena or things or between mathematical or statistical variables which tend to vary, be associated, or occur together in a way not expected on the basis of chance alone'" (https://michaelminn.net/tutorials/arcgis-pro-regression/#r-squared).

"When there is no correlation, the X/Y scatter chart dots have no clear upward or downward pattern" (https://michaelminn.net/tutorials/arcgis-pro-regression).

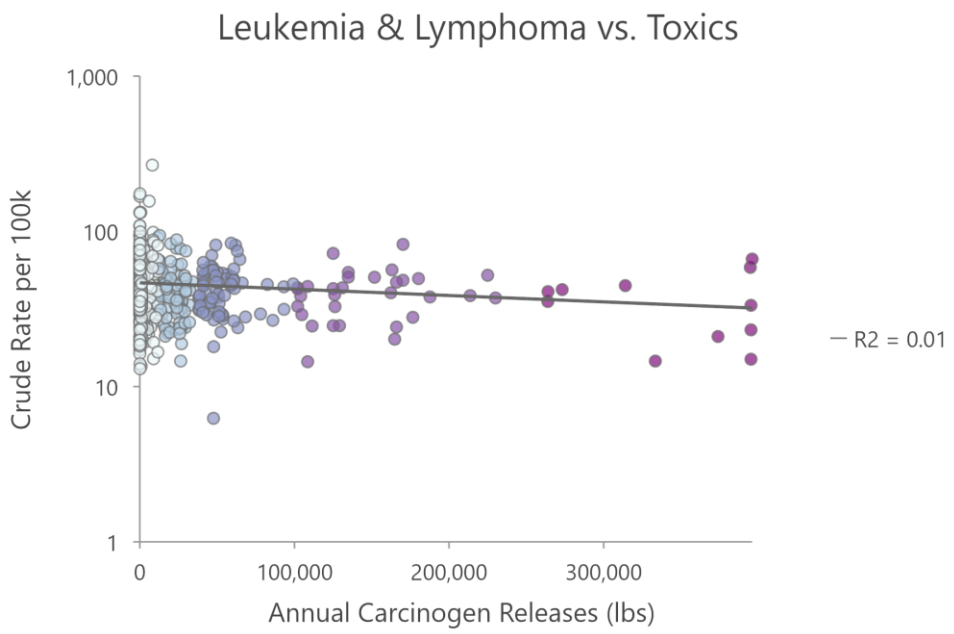

Figure 6: X/Y scatter chart comparing leukemia and lymphoma cancer rates and carcinogen releases in Illinois, 2015 – 2019

"Adjusted R-squared is calculated to compensate for increases in R-squared that will occur as additional variables are added but do not necessarily add explanatory power to the model" (https://michaelminn.net/tutorials/arcgis-pro-regression/#r-squared).

In this model, the value of adjusted R-Squared is 0.194754, which indicates that there is a relatively strong fit, meaning 19% of the variation in leukemia and lymphoma cases can be attributed to the selected independent variables.

OLS Diagnostics

Input Features	Zip_Code_Toxics	Dependent Variable	CRUDE_RATE_PER_100K
Number of Observations	606	Akaike's Information Criterion (AICc)['d']	5318.755209
Multiple R-Squared['d']	0.201409	Adjusted R-Squared['d']	0.194754
Joint F-Statistic['e']	30.264684	Prob(>F), (5,600) degrees of freedom	0.000000*
Joint Wald Statistic['e']	124.618694	Prob(>chi-squared), (5) degrees of freedom	0.000000*
Koenker (BP) Statistic['f']	23.777009	Prob(>chi-squared), (5) degrees of freedom	0.000240*
Jarque-Bera Statistic['g']	16651.658373	Prob(>chi-squared), (2) degrees of freedom	0.000000*

Notes on Interpretation

* An asterisk next to a number indicates a statistically significant p-value ($p < 0.01$).

[a] Coefficient: Represents the strength and type of relationship between each explanatory variable and the dependent variable.

[b] Probability and Robust Probability (Robust_Pr): Asterisk (*) indicates a coefficient is statistically significant ($p < 0.01$); if the Koenker (BP) Statistic [f] is statistically significant, use the Robust Probability column (Robust_Pr) to determine coefficient significance.

[c] Variance Inflation Factor (VIF): Large Variance Inflation Factor (VIF) values (> 7.5) indicate redundancy among explanatory variables.

[d] R-Squared and Akaike's Information Criterion (AICc): Measures of model fit/performance.

[e] Joint F and Wald Statistics: Asterisk (*) indicates overall model significance ($p < 0.01$); if the Koenker (BP) Statistic [f] is statistically significant, use the Wald Statistic to determine overall model significance.

[f] Koenker (BP) Statistic: When this test is statistically significant ($p < 0.01$), the relationships modeled are not consistent (either due to non-stationarity or heteroskedasticity). You should rely on the Robust Probabilities (Robust_Pr) to determine coefficient significance and on the Wald Statistic to determine overall model significance.

[g] Jarque-Bera Statistic: When this test is statistically significant ($p < 0.01$) model predictions are biased (the residuals are not normally distributed).

Figure 7: Multiple regression diagnostics for leukemia and lymphoma cancer rates in Illinois, 2015 - 2019

From the model, it appears that the intercept and median age are statistically significant independent variables.

Median age was the most statistically significant variable—increasing the median age by one yields 1.52 new cancer cases per 100k residents to the model.

Summary of OLS Results - Model Variables

Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]	Robust_SE	Robust_t	Robust_Pr [b]	VIF [c]
Intercept	-39.960523	19.864375	-2.011668	0.044694*	19.623698	-2.036340	0.042146*	--------
PERCENT_VETE	0.175231	0.253417	0.691472	0.489532	0.394808	0.443837	0.657334	1.296191
MEDIAN_HOUSE	0.000052	0.000032	1.644892	0.100528	0.000054	0.971167	0.331844	1.374408
PERCENT_HEAL	0.212635	0.231266	0.919441	0.358221	0.264419	0.804160	0.421611	1.461715
MEDIAN_AGE	1.515967	0.153026	9.906592	0.000000*	0.200633	7.555924	0.000000*	1.208016
F62__ON_SITE	-0.000024	0.000014	-1.687842	0.091971	0.000012	-1.944685	0.052275	1.032080

Figure 8: Multiple regression results for leukemia and lymphoma cancer rates in Illinois, 2015 – 2019

"Geographical phenomena are very commonly clustered together in space, which means that geospatial variables very commonly exhibit spatial autocorrelation that causes multiple regression model coefficients and outputs to be biased so that model outputs are untrustworthy" (https://michaelminn.net/tutorials/arcgis-pro-regression/#r-squared).

As the z-score of 3.613 is greater than 2.58, there is high spatial autocorrelation in the model residuals.

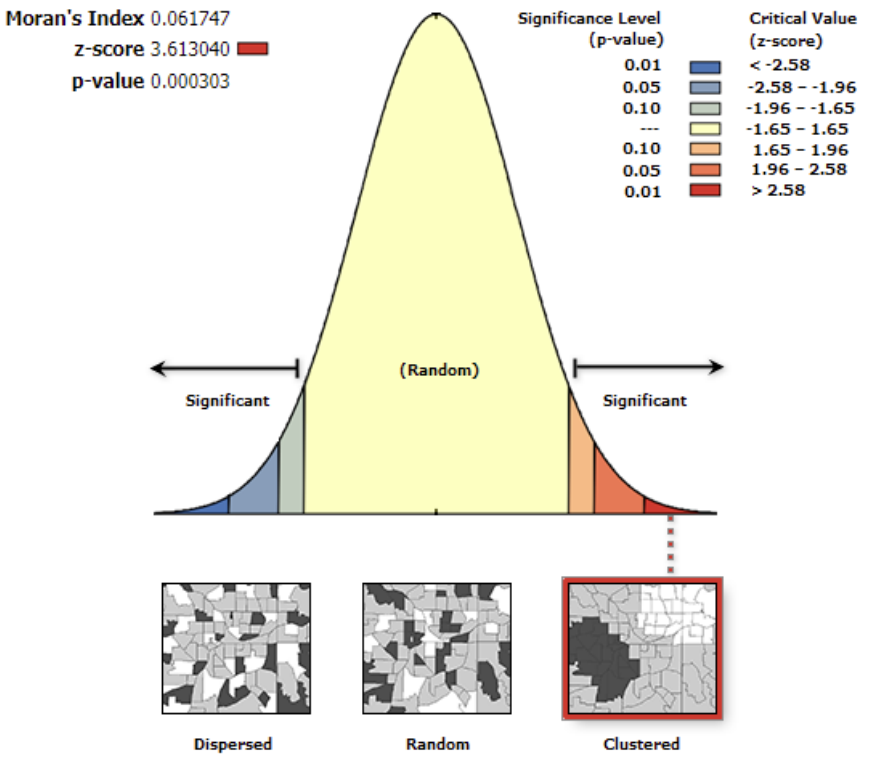

Figure 9: Residuals autocorrelation analysis

"Exploratory regression is a data mining technique that involves trying all possible combinations of explanatory variables to find the best models" (https://michaelminn.net/tutorials/arcgis-pro-regression/#r-squared).

As the adjusted R-Squared values are all 0.29, it appears that there is a weak positive correlation between the median household income, percent 65+, percent work at home, percent single mothers, and rates of leukemia and lymphoma, and a weak negative correlation exists between the percent no vehicles and leukemia and lymphoma.

Highest Adjusted R-Squared Results

AdjR2	AICc	JB	K(BP)	VIF	SA	Model
0.29	4933.71	0.00	0.00	1.96	0.00	+MEDIAN_HOUSEHOLD_INCOME* +PERCENT_65_PLUS*** +PERCENT_WORK_AT_HOME* +PERCENT_SINGLE_MOTHERS** -PERCENT_NO_VEHICLE***
0.29	4933.73	0.00	0.00	1.50	0.00	+MEDIAN_HOUSEHOLD_INCOME*** +PERCENT_PRE_WAR_UNITS** +PERCENT_65_PLUS*** +PERCENT_SINGLE_MOTHERS** -PERCENT_NO_VEHICLE***
0.29	4935.40	0.00	0.00	2.85	0.00	+PERCENT_65_PLUS*** +PERCENT_TRANSIT_TO_WORK** +PERCENT_WORK_AT_HOME** +PERCENT_SINGLE_MOTHERS** -PERCENT_NO_VEHICLE***

Figure 10: Exploratory regression results for leukemia and lymphoma cancer rates in Illinois, 2015 – 2019

“In a situation like this where the overall effect of carcinogenic releases is not clearly present in a model covering the general area, it may be more useful to focus on identifying and further investigating specific outlier areas where there is a confluence of both high cancer rates and high carcinogenic releases” (https://michaelminn.net/tutorials/arcgis-pro-regression/#outlier-analysis).

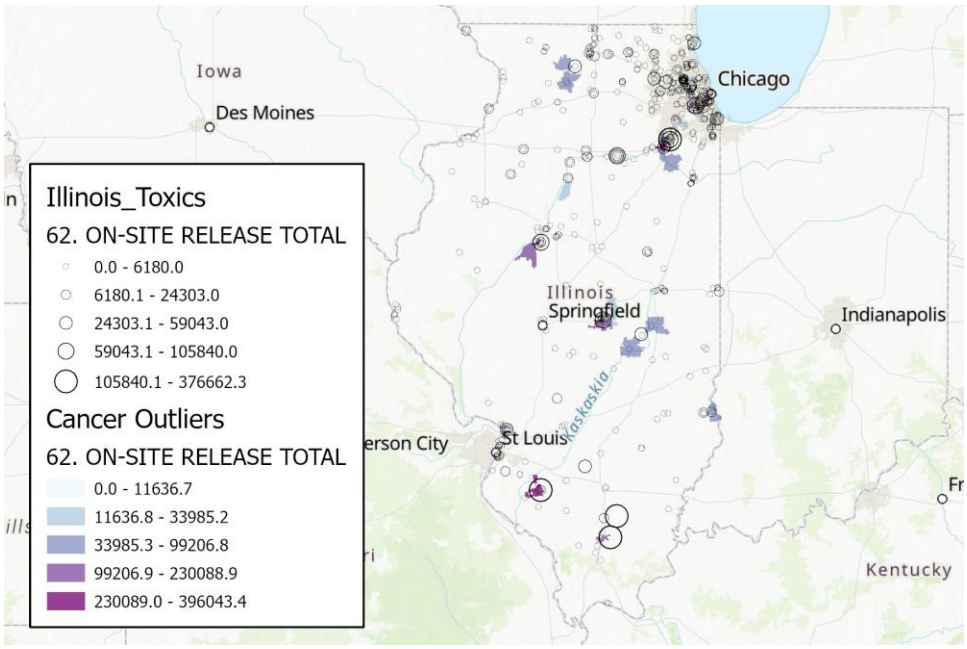

Figure 11: High cancer and exposure zip codes in Illinois

Zip code 60410 is in located in Channahon, IL, part of the larger Will County.

For zip code 60410, the crude rate of leukemia and lymphoma per 100k residents is 58.62 and the estimated annual carcinogen exposure in pounds is 394870.77

The largest emitter sources deal primarily with Styrenics, plastics, and oil refining.

“Workers exposed to high levels of vinyl chloride for several years have higher rates of liver cancer” (http://www.idph.state.il.us/cancer/factsheets/vinyl-chloride.htm).

Figure 12: High cancer and exposure zip code 60410 in Will County, Illinois