Air quality and lung cancer: Analysis via Local Control

Mithun Kumar Acharjee[1], Kumer Das[2], S. Stanley Young[3]
[1]Department of Mathematics, Lamar University, Beaumont, TX-77710
[2]Department of Mathematics, Lamar University, Beaumont, TX-77710
[3]CGStat LLC, Raleigh, North Carolina

Abstract
The possible association between PM2.5 and lung cancer mortality can be partitioned into components, within similar observational units, and across different observational units. Within unit's covariates are very similar and across unit's covariates can be, and usually are, very different. Hence, there is a need to understand the possible effect of PM2.5 on mortality taking into account within and between observational units. To know the important covariates, our idea is to use Local Control Analysis (LCA) to estimate these two components and determine how much of the variation in estimates can be attributed. For the purpose of analysis, we calculated Local Treatment Difference (LTD) for LTD approach and slope and intercept for Local Control Regression (LCR) approach, to determine if the treatment (PM2.5) effects vary significantly across clusters. For that evaluation, we used Recursive Partitioning (RP). The benefit of this study is twofold. First, we use a reliable strategy (LCA) for observational data. Second and importantly, there is subgroup heterogeneity in the effect of PM2.5 on lung cancer mortality and this heterogeneity is largely explained by factors other than air quality.

Keywords: PM2.5, Lung Cancer Mortality, Local Control Analysis

1. Introduction

Epidemiological studies conducted on particles PM2.5 indicate that PM2.5 has substantially greater toxicity than larger particles [1-3]. A few study make the case that if potential bias is carefully taken into account then there is no association between air quality and deaths [4-8]. Although results have been mixed, overall, an association between PM2.5 (particles having aerodynamic diameter ≤2.5 μm) exposure and lung cancer mortality have been reported by recent studies [7, 9, 10]. PM2.5 is associated with greater increases in daily mortality than larger particles, and are of greater public health concern [1, 3], suggesting that size is not the only indicator of PM-related health effects.

The empirical estimation of cumulative environmental exposures, from different environmental sources, is difficult. As a result, epidemiologic research has traditionally focused on single environmental exposures [11, 12]. However, it is impossible to get total picture of the environmental effects on health by measuring a single environmental exposure. Rather, various environmental exposures (including social exposures) occur simultaneously, to engender poor health upshots, including cancer. The Environmental Quality Index (EQI) was developed to capture multidimensional ambient environmental exposures. The publically available EQI [13] is a county-level measure of cumulative ambient environmental exposures for the United States for the period 2000- 2005 [11, 14].

We examined county-level lung cancer incidence rates for the period 2010-2014 in association with the EQI, which represents the period 2000-2005. To assess which environmental domains drive the associations with cancer incidence, we also considered three domain-specific indices namely land EQI, sociodemographic EQI and built EQI. The magnitude of this association varies with geographic location [15].

We also investigated associations between lung cancer incidence and the PM2.5. The possible association between PM2.5 and lung cancer mortality can be partitioned into components, within similar observational units and across different observational units. Within unit's covariates are very similar and across unit's covariates can be, and usually are, very different. Hence, there is a need to understand the possible effect of PM2.5 on mortality taking into account within and between observational units. In this study, our idea is to use a relatively new statistical strategy known as Local Control Analysis (LCA) to estimate these two components and determined how much of the variation in estimates can be attributed to know the important covariates. The basic idea of LCA is to cluster experimental units, in our case counties, into subgroups that have similar socio-economic characteristics but are not necessarily spatially contiguous. We then make statistical comparisons primarily within these clusters of counties that are relatively homogeneous in terms of important confounding factors.

To determine if the treatment (PM2.5) effect varies significantly across clusters, we calculated Local Treatment Difference (LTD) for LTD approach and slope and intercept for Local Control Regression (LCR) approach. Recursive Partitioning (RP) is used for that purpose. The benefits of this study are twofold. First, we use a reliable strategy (LCA) for observational data. Second and importantly, there is subgroup heterogeneity in the effect of PM2.5 on lung cancer mortality and this heterogeneity is largely explained by factors other than air quality.

2. Methodology

2.1 Data and Variables

Data comes from two different sources. Population-based lung cancer incidence rates for the period 2010-2014 (most updated data) were abstracted from National Cancer Institute state cancer profiles [16]. This national county-level database of cancer data is collected by state public health surveillance systems. The domain specific county level environmental quality index (EQI) data for the period 2000-2005 were abstracted from United States Environmental Protection Agency (USEPA) profile. Complete descriptions of the datasets used in the EQI are provided in Lobdell's paper [17]. Data were merged based on the Federal Information Processing Standards (FIPS) code.Out of 3144 counties in United States this study has available information for 2602 counties: Data was not available for four states namely Kansas, Michigan, Minnesota and Nevada due to state legislation because of state legislation and regulations which prohibit the release of county-level data to outside entities, county whose lung cancer mortality information is missing were omitted from the data set, the Union county, Florida is an outlier in terms of mortality information which was deleted from the data set, in the process of local control analysis this study experiences two non-informative clusters (non-informative cluster is one for which either treatment or control group information is missing) (cluster 28 and 29). For the purpose of analysis, non-informative clusters information was deleted from the data set.

Three types of variables was used in this study: (i) lung cancer mortality as an outcome variable (ii) binary treatment indicator is the PM2.5 high (greater than 10.59 mg/m^3) vs. low (less than 10.59 mg/m^3) (iii) three potential X confounder for clustering namely land EQI, sociodemographic EQI and built EQI. For each index, higher values correspond to poorer environmental quality [18].

2.2 Local Control Analysis

Local Control Analysis first clusters the data set and then a simple analysis within each cluster. The statistics (LTD, Slope and Intercept) coming from each cluster are then analyzed further. Once the descriptive statistics are computed, we determine if these treatment effects vary significantly across clusters. For that evaluation, we use recursive partitioning (RP). There is a SAS JMP Add-In, Local Control, for automating steps in the analysis process [19].

Local control, LC, analysis strategy [20, 21, 22], for large, observational data sets is easily explained as it is a series of simple steps that together provide a coherent analysis strategy.

Step1: LC starts by dividing data points without regard for their either status as treated or control, into many subgroups. The point is to assure that objects within a cluster are as alike as possible for their observed baseline x-characteristics. They used hierarchical clustering.

Step2: Next, a simple difference between the two treatments is computed within each cluster:

$$E[\ (Y|t=1) - (Y|t=0)\ |X],$$

so that they have a single degree of freedom comparison, given X. That is known as “fair treatment comparison” and a local treatment difference, LTD or LC effect size estimates. These LC effect-sizes are (continuous) measures of within-cluster association between local observed y-outcomes and t-exposures.

Step3: Next, they display the LTDs in a histogram.

Step4: Once the descriptive statistics are computed, we determine if these treatment effects vary significantly across clusters using RP.

Therefore, the LCA starts by designating just one of the available X-confounders as the treatment t-exposure, which is the "main cause" of observed variation in the *Y*-outcome variable. LC then repeats through its three nonparametric preprocessing phases (Aggregate, Confirm and Explore) to “design” new variables LTDs, Slopes or Intercept. Finally, we use Reveal phase to determine whether LTSs, Slope or Intercept appear to be either truly heterogeneous or most random. In this study, we used two approaches of LC method:

A. Local Treatment Difference (LTD) approach: Local Treatment Difference, LTD (∂), defined as, LTD (∂) = E[(Y|t=1) – (Y|t=0)/ X within ∂], where ∂ denotes a cluster of subjects who are relatively well-matched in *X* confounder space.

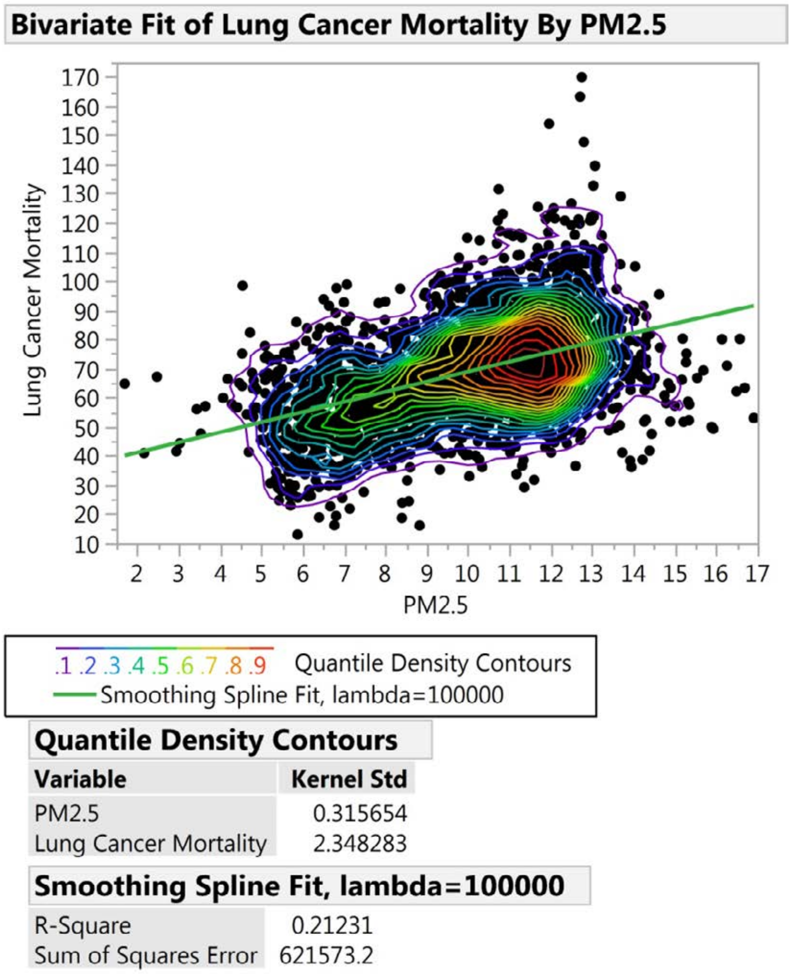
B. Local Control Regression (Slope and Intercept) approach: We performed simple linear regression for 50 clusters, got slope and intercept for each cluster, found 48 non-informative clusters. Here slope and intercept takes the place of LTD. Finally, repeated the 4-phase process of LTD approach for slope and intercept.

2.3 Clustering Technique

Hierarchical clustering technique is used to cluster the county level lung cancer data. Hierarchical cluster analysis begins by separating each object into a cluster by itself. Wards method is used as a linkage criterion. In this study, we produced 50 subgroups or clusters using wards method. Within each subgroup, counties are relatively well matched based on the 3 potential X confounders. JMP software is used to cluster the data.

3. Results

The EQI was linked to county-level annual age-adjusted cancer incidence rates from the Surveillance, Epidemiology, and End Results (SEER) Program state cancer profiles. Figure 1 shows a positive association between PM2.5 and lung cancer mortality.



Quantile Density Contours

Variable	Kernel Std
PM2.5	0.315654
Lung Cancer Mortality	2.348283

Smoothing Spline Fit, lambda=100000

R-Square	0.21231
Sum of Squares Error	621573.2

Figure 1: Association between PM2.5 and lung cancer mortality

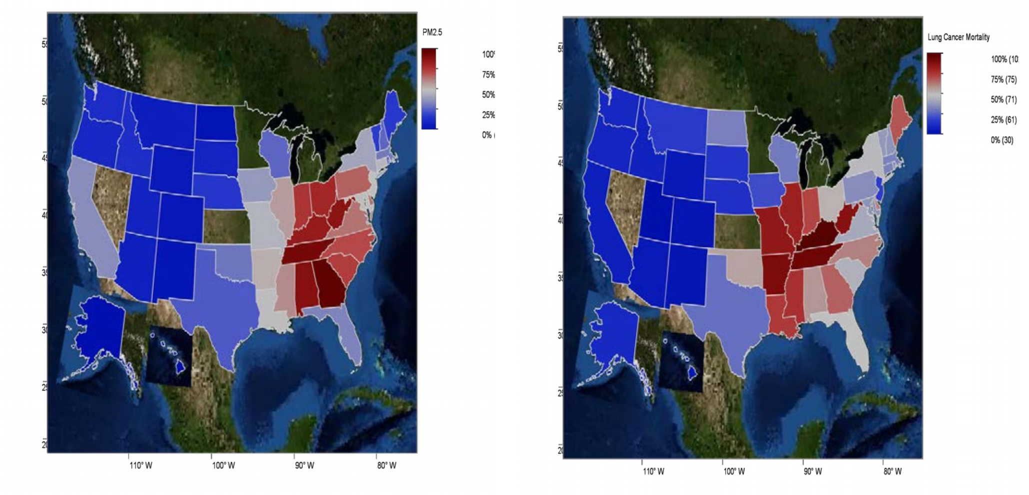
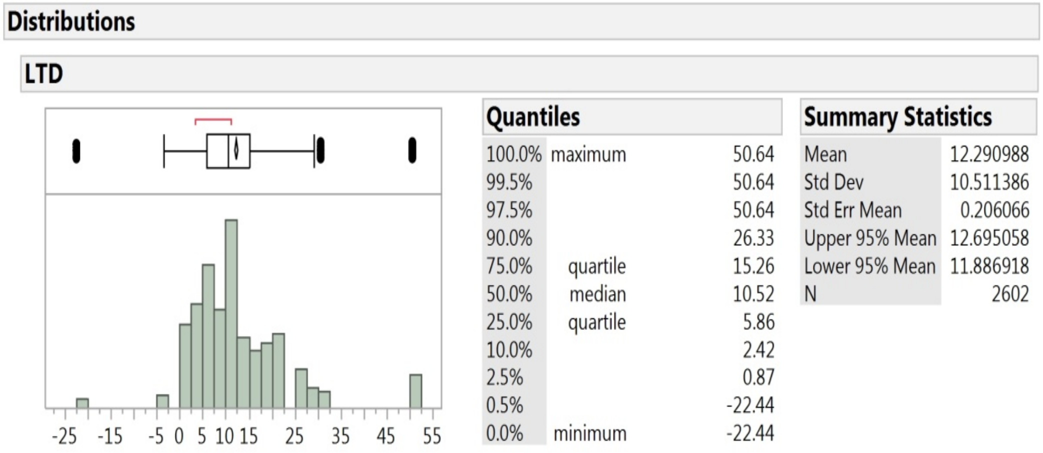


Figure 2: State colored by PM2.5 (left figure) and lung cancer mortality (right figure)

Figure 2 shows that data is missing for 4 states namely Kansas, Michigan, Minnesota and Nevada. Due to high PM2.5 in the different North East and South East central states, lung cancer is also very high. PM2.5 is high in Alabama, Georgia, Tennessee, Kentucky, Virginia, West Virginia, Illinois, Indiana, South Carolina, North Carolina and Pennsylvania. Lung cancer is almost high in the same states plus Arkansas, Missouri, Iowa and Maine. Most of the manufacturing industries are in Alabama, Arkansas, Illinois, Indiana, Iowa, Kansas, Kentucky, Louisiana, Michigan, Mississippi, North Carolina, Ohio, Pennsylvania, South Carolina and Texas. As a result, the rate of PM2.5 rate is high in these states, which leads to the high lung cancer mortality. Only exception is Maine. Maine's lung cancer mortality rate is 30% higher than the national average. Smoking and tobacco use was identified as the major risk factor for lung cancer.

3.1 Distribution and recursive partitioning tree for LTD using JMP



Distributions

LTD

Quantiles

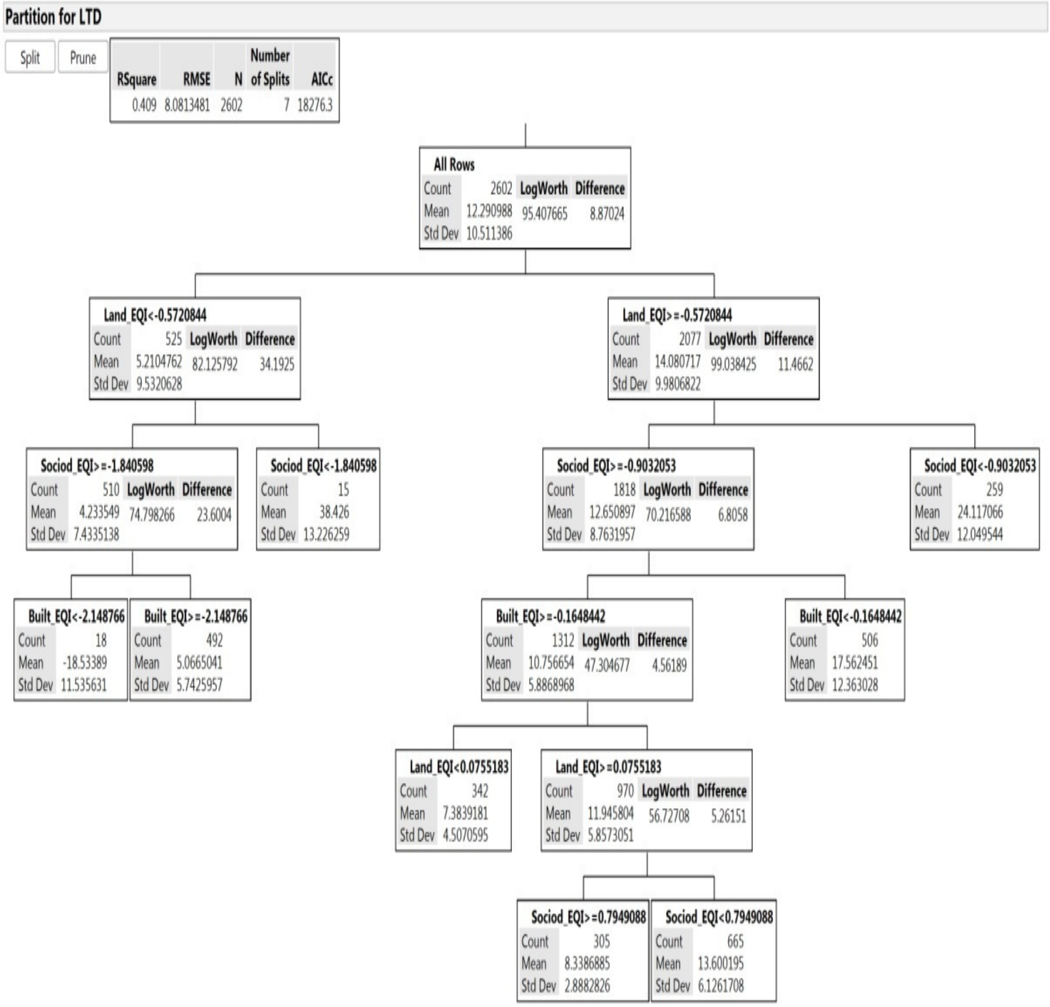
100.0%	maximum	50.64
99.5%		50.64
97.5%		50.64
90.0%		26.33
75.0%	quartile	15.26
50.0%	median	10.52
25.0%	quartile	5.86
10.0%		2.42
2.5%		0.87
0.5%		-22.44
0.0%	minimum	-22.44

Summary Statistics

Mean	12.290988
Std Dev	10.511386
Std Err Mean	0.206066
Upper 95% Mean	12.695058
Lower 95% Mean	11.886918
N	2602

Figure 3: Local Treatment Difference (LTD) distribution

These histograms display distributions of treatment effect sizes. The histogram displays the LTD estimates for 48 informative subgroups of lung cancer patients relatively well matched in x spaces. Figure 3 shows that LTD is positive which indicates that the lung cancer mortality increases with the increase of PM2.5.



Partition for LTD

RSquare	RMSE	N	Number of Splits	AICc
0.409	8.0813481	2602	7	18276.3

All Rows

Count	2602	LogWorth	Difference
Mean	12.290988	95.407665	8.87024
Std Dev	10.511386		

Land_EQI<-0.5720844

Count	525	LogWorth	Difference
Mean	5.2104762	82.125792	34.1925
Std Dev	9.5320628		

Land_EQI>=-0.5720844

Count	2077	LogWorth	Difference
Mean	14.080717	99.038425	11.4662
Std Dev	9.9806822		

Sociod_EQI>=-1.840598

Count	510	LogWorth	Difference
Mean	4.233549	74.798266	23.6004
Std Dev	7.4335138		

Sociod_EQI<-1.840598

Count	15
Mean	38.426
Std Dev	13.226259

Sociod_EQI>=-0.9032053

Count	1818	LogWorth	Difference
Mean	12.650897	70.216588	6.8058
Std Dev	8.7631957		

Sociod_EQI<-0.9032053

Count	259
Mean	24.117066
Std Dev	12.049544

Built_EQI<-2.148766

Count	18
Mean	-18.53389
Std Dev	11.535631

Built_EQI>=-2.148766

Count	492
Mean	5.0665041
Std Dev	5.7425957

Built_EQI>=-0.1648442

Count	1312	LogWorth	Difference
Mean	10.756654	47.304677	4.56189
Std Dev	5.8868968		

Built_EQI<-0.1648442

Count	506
Mean	17.562451
Std Dev	12.363028

Land_EQI<0.0755183

Count	342
Mean	7.3839181
Std Dev	4.5070595

Land_EQI>=0.0755183

Count	970	LogWorth	Difference
Mean	11.945804	56.72708	5.26151
Std Dev	5.8573051		

Sociod_EQI>=0.7949088

Count	305
Mean	8.3386885
Std Dev	2.8882826

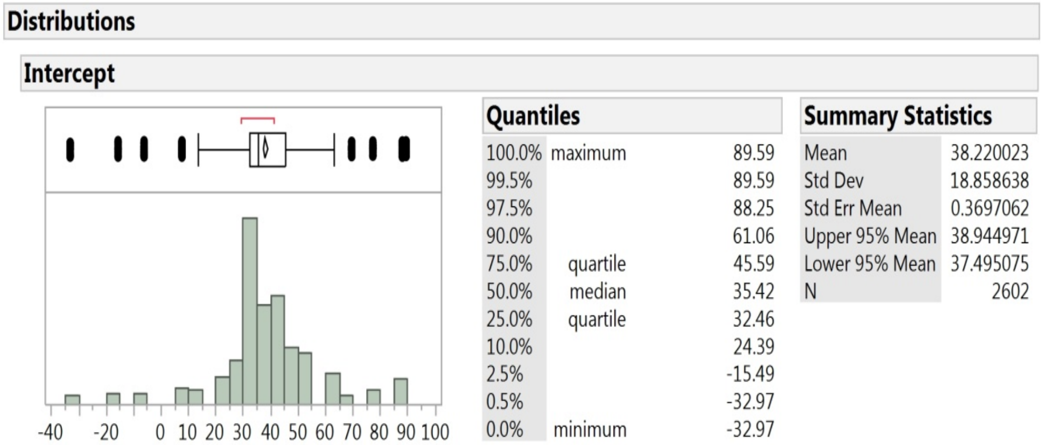
Sociod_EQI<0.7949088

Count	665
Mean	13.600195
Std Dev	6.1261708

Figure 4: A "tree model" for prediction of observed local treatment differences

Since land domain separates off 525 counties with much lower LTD (low rate of lung cancer incidence) where the environmental quality is high (less negative EQI), it is no wonder that residents with land EQI is a very good predictor of across cluster variation in LTDs. Similarly, socio-demographic EQI is used in three highly significant splits (high logworth value (logworth is negative log of the p-value for the split)), with LTDs expected to be higher where the land EQI is high. On the other hand, built EQI is used in two highly significant splits, with high LTD. For both splits, the land EQI is high. So, all three x confounders are important predictors. Note that the SD does not decrease much indicating heterogeneity. The high logworth value and standard deviation indicates that LTD values differ significantly across clusters.

3.2 Distribution and recursive partitioning tree for intercept using JMP



Quantiles

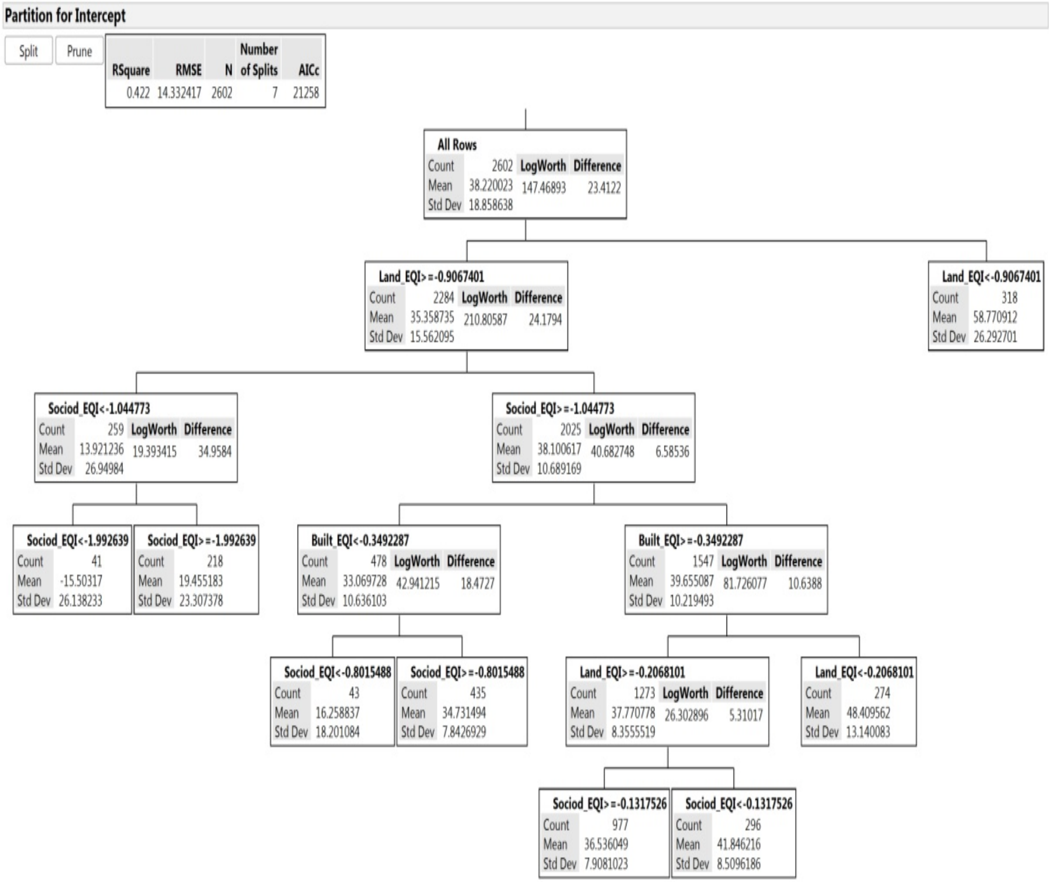
100.0%	maximum	89.59
99.5%		89.59
97.5%		88.25
90.0%		61.06
75.0%	quartile	45.59
50.0%	median	35.42
25.0%	quartile	32.46
10.0%		24.39
2.5%		-15.49
0.5%		-32.97
0.0%	minimum	-32.97

Summary Statistics

Mean	38.220023
Std Dev	18.858638
Std Err Mean	0.3697062
Upper 95% Mean	38.944971
Lower 95% Mean	37.495075
N	2602

Figure 5: Distribution of intercept for non-informative clusters

Point to histogram of intercepts indicates that there are major differences across the counties related to covariates. The overall differences are positive and are not random.



RSquare	RMSE	N	Number of Splits	AICc
0.422	14.332417	2602	7	21258

All Rows

		LogWorth	Difference
Count	2602		
Mean	38.220023	147.46893	23.4122
Std Dev	18.858638		

Land_EQI>=-0.9067401

		LogWorth	Difference
Count	2284		
Mean	35.358735	210.80587	24.1794
Std Dev	15.562095		

Land_EQI<-0.9067401

Count	318
Mean	58.770912
Std Dev	26.292701

Sociod_EQI<-1.044773

		LogWorth	Difference
Count	259		
Mean	13.921236	19.393415	34.9584
Std Dev	26.94984		

Sociod_EQI>=-1.044773

		LogWorth	Difference
Count	2025		
Mean	38.100617	40.682748	6.58536
Std Dev	10.689169		

Sociod_EQI<-1.992639

Count	41
Mean	-15.50317
Std Dev	26.138233

Sociod_EQI>=-1.992639

Count	218
Mean	19.455183
Std Dev	23.307378

Built_EQI<-0.3492287

		LogWorth	Difference
Count	478		
Mean	33.069728	42.941215	18.4727
Std Dev	10.636103		

Built_EQI>=-0.3492287

		LogWorth	Difference
Count	1547		
Mean	39.655087	81.726077	10.6388
Std Dev	10.219493		

Sociod_EQI<-0.8015488

Count	43
Mean	16.258837
Std Dev	18.201084

Sociod_EQI>=-0.8015488

Count	435
Mean	34.731494
Std Dev	7.8426929

Land_EQI>=-0.2068101

		LogWorth	Difference
Count	1273		
Mean	37.770778	26.302896	5.31017
Std Dev	8.3555519		

Land_EQI<-0.2068101

Count	274
Mean	48.409562
Std Dev	13.140083

Sociod_EQI>=-0.1317526

Count	977
Mean	36.536049
Std Dev	7.9081023

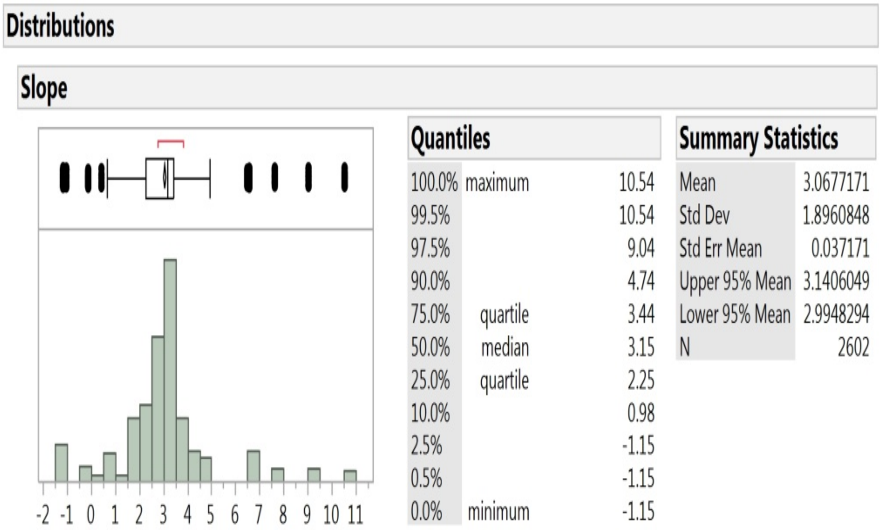
Sociod_EQI<-0.1317526

Count	296
Mean	41.846216
Std Dev	8.5096186

Figure 6: A “tree model” for prediction of intercept

Again, the lung cancer mortality is low in case of high land EQI (high environmental quality), sociodemographic EQI and built EQI. Mean LTD is comparatively low in case of high environmental quality means that 3 confounders are good predictors. The land and sociodemographic domain dominates the intercept tree. In overall, the logworth value is very high means the splits are highly significant. Therefore, we can conclude that heterogeneity exists across the counties related to covariates.

3.3 Distribution and recursive partitioning tree for slope using JMP



Distributions

Slope

Quantiles

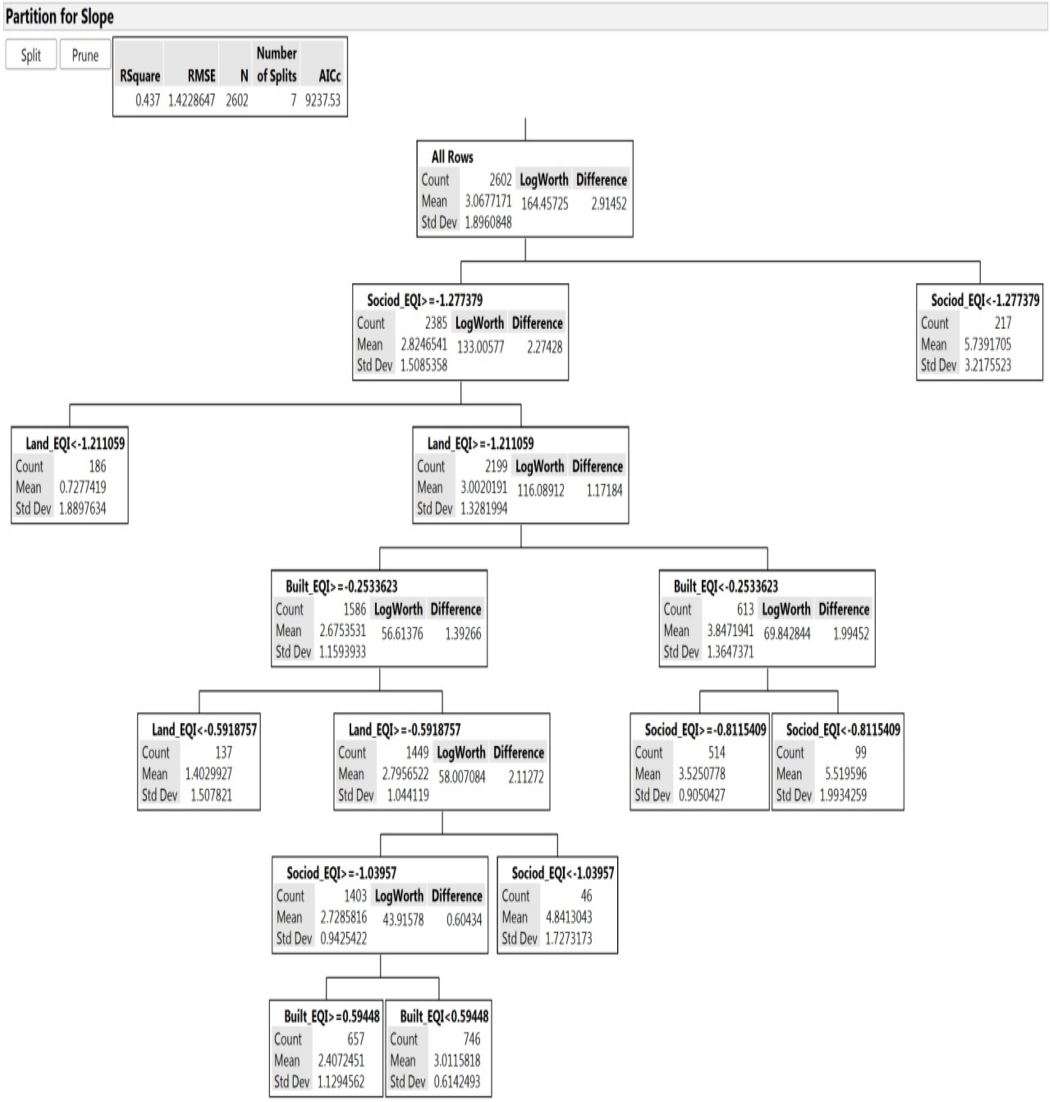
100.0%	maximum	10.54
99.5%		10.54
97.5%		9.04
90.0%		4.74
75.0%	quartile	3.44
50.0%	median	3.15
25.0%	quartile	2.25
10.0%		0.98
2.5%		-1.15
0.5%		-1.15
0.0%	minimum	-1.15

Summary Statistics

Mean	3.0677171
Std Dev	1.8960848
Std Err Mean	0.037171
Upper 95% Mean	3.1406049
Lower 95% Mean	2.9948294
N	2602

Figure 7: Distribution of slopes for non-informative clusters

The slope histogram display distributions of treatment effect sizes. Slope did the same thing that LTD done where the only difference is that these two methods are used for two different approaches (LTD approach and Local Linear Regression approach) of Local Control. The histogram displays the slope estimates for 48 informative subgroups of lung cancer patients relatively well matched in x spaces. Figure 7 shows that slope is positive which indicates that the lung cancer mortality increases with the increase of PM2.5.



Partition for Slope

RSquare	RMSE	N	Number of Splits	AICc
0.437	1.4228647	2602	7	9237.53

Figure 8: A “tree model” for prediction of slope

Figure 8 shows that the sociodemographic EQI is most important as it appears in 1st split and often in the tree. The overall logworth value is high indicates the higher heterogeneity. The large heterogeneity says that there is no constant/ consistent effect of PM2.5.

4. Conclusion

LC strategy gives importance on local effect-size estimates and their empirical distribution, rather than upon the statistical significance of some of their values [23]. Local Control approach is so straightforward simple that one can easily grasp the idea by understanding of clustering, simple differences, single predictor linear regression and histograms. Histograms are drawn robustly estimate LTD, intercept and slope and display their distribution across clusters.

In case of LTD approach, we consider a binary treatment variable. Based on that binary variable we compute the mean of lung cancer incidence rate for each non-informative cluster. This study considers PM2.5 as binary treatment variable. On the other hand, the LCR approach does not need any binary treatment variable. Instead of that LCR consider the PM2.5 as continuous independent variable to perform the simple linear regression for each 50 clusters, which gives the slope, and intercept for each cluster. LCR developed by Young and Obenchain [23] is very similar to what Janes et al [6] and Greven et al [7] have done; they look within and across locations, and Young and Obenchain work within and across clusters of locations. In this study, intercepts represent baseline effects while slopes can be taken to be the adjusted effects of lung cancer incidence versus PM2.5 exposures within clusters.

In this study, the bivariate graph shows the positive association between lung cancer incidence and PM2.5. In addition, the histogram shows that the overall distribution of LTD and slope is positive. This positive distribution depicts that the lung cancer mortality increases with the increase of PM2.5. Point to histogram of intercepts indicates that there are major differences across the counties related to covariates. The decision tree for LTD shows that the land domain is the most important as it appears in first split and appears often in the tree. Sociodemographic and land is the most important domain in case of slope and intercept decision tree respectively. For each tree, the logworth value is high. However, with logworths this large, the splits are still highly significant. The large heterogeneity says that there is no constant/ consistent effect of PM2.5. One possible explanation for this high heterogeneity is that there may well be important unmeasured covariates.

References

[1] Schwartz, J., Dockery, D., and Neas, L. 1996. Is daily mortality associated specifically with fine particles? JAWMA, 46:927–939

[2] Burnett, R., Brook, J., Dann, T., Delocla, C., Philips, O., Cakmak, S., Vincent, R., Goldberg, M.S., and Krewski, D. 2000. Association between particulate- and gas-phase components of urban air pollution and daily mortality in eight Canadian cities. Inhal Toxicol, 12(4):15–39

[3] Cifuentes, L., Vega, J., Ko¨pfer K., and Lave, L. 2000. Effect of the fine fraction of particulate matter versus the coarse mass and other pollutants on daily mortality in Santiago, Chile. JAWMA, 50:1287–1298.

[4] Chay, K., and Dobkin, C. 2003.The Clean Air Act of 1970 and Adult Mortality. The Journal of Risk and Uncertainty, 27:279–300.

[5] Enstrom, J.E. 2005. Fine particulate air pollution and total mortality among elderly Californians, 1973-2002. Inhalation Toxicology, 17: 803-816.

[6] Janes, H, Dominici, F, Zeger, and S. L. 2007. Trends in air pollution and mortality: an approach to the assessment of unmeasured confounding. Epidemiology. 18:416–423.

[7]Greven, S., Dominici, F., and Zeger, S. 2011.An Approach to the Estimation of Chronic Air Pollution Effects Using Spatio-Temporal Information. Journal of the American Statistical Association, 106:494, 396-406.

[8] Cox, P. M., Pearson, D., Booth, B.B., Friedlingstein, P., Huntingford, C., Jones, D.C., and Luke, C.M. 2013. Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability. Nature, 494, 341-344.

[9] Krewski, D, Lubin, J.H., Zielinski, J.M., Alavanja, M., Catalan, V.S., Field, R.W., Klotz, J. B., Létourneau, E.G, Lynch, C.F., Lyon, J.L., Sandler, D.P., Schoenberg, J.B., Steck, D.J., Stolwijk, J.A., Weinberg, C., and Wilcox, H. B. 2006. A combined analysis of North American case-control studies of residential radon and lung cancer. Journal of Toxicology and Environmental Health, 69(7):533–597

[10] Beelen, R., Hoek, G., Piet A., Brandt, R. Goldbohm, A., Fischer, P, Leo J. Schouten, J.L., Jerrett, M., Hughes, E., Armstrong, B., and Brunekreef, B. 2008. Long-term effects of traffic-related air pollution on mortality in a Dutch Cohort (NLCS-AIR Study). Environ health perspect, 116(2): 196–202.

[11] Messer, L.C., Jagai, J.S., Rappazzo, K. M., and Lobdell, D.T. 2013. Construction of an environmental quality index for public health research. Environ Health, pp.13:39.

[12] Peters, A., Hoek, G., and Katsouyanni, K. 2012.Understanding the link between environmental exposures and health: does the exposome promise too much? J Epidemiol Community Health,66:103-105.

[13] US Environmental Protection Agency. 2015. Environmental Quality Index. Washington, DC: US Environmental Protection Agency.

[14] Lobdell, D.T., Jagai, J. S., Rappazzo, K., and Messer, L.C. 2011. Data sources for an environmental quality index: availability, quality, and utility. Am J Public Health,101(suppl 1):S277-S285.

[15] Franklin, M., Zeka, A., and Schwartz, J. 2007. Association between PM2.5 and all-cause and specific-cause mortality in 27 US communities. J Expo Sci Environ Epidemiology,17:279–287.

[16] Siegel, R., Naishadham, D., Jemal, A. 2014. Cancer statistics, 2013. CA Cancer J Clin,63:11-30.

[17] Lobdell, D.T., Jagai, J.S., Rappazzo, K., and Messer, L.C. 2011.Data sources for an environmental quality index: availability, quality, and utility. Am J Public Health,101(suppl 1):S277–S285.

[18] Jagai, J. S., Messer, L. C., Rappazzo, K. M., Gray, C. L., Grabich, S. C. and Lobdell, D. T. 2017. County-level cumulative environmental quality associated with cancer incidence. Cancer, 123: 2901–2908.

[19] Wolfinger and Obenchain. 2014.https://community.jmp.com/docs/DOC-7453.

[20] Obenchain, R. L. 2009. SAS macros for local control (phases one and two), Observational Medical Outcomes Partnership (OMOP), Foundation for the National Institutes of Health (Apache 2.0 License). http://members.iquest.net/~softrx

[21] Obenchain, R. L. 2010. The local control approach using JMP. In Analysis of observational health care data using SAS, ed. D. E. Faries, A. C. Leon, J. M. Haro, and R. L. Obenchain, 151– 192. Cary, NC, SAS Press.

[22] Obenchain, R.L., and Young, S. S. 2013. Advancing Statistical Thinking in Observational Health Care Research, Journal of Statistical Theory and Practice 7:2:456-469.

[23] Young, S., Obenchain, L., and Kristic, G. 2015. Bias and response heterogeneity in an air quality data set. JMP discovery summit proceedings 2015. https://community.jmp.com/t5/Discovery-Summit-2015/Bias-Adjustment-in-Data-Mining-Local-Control-Analysis-of-Radon/ta-p/22864.