The Impact of $PM_{2.5}$ on Lung and Bronchial Cancers: Regression and Time Series Analysis in the U.S. from 1999 to 2014

Jing Kersey[a], Jingjing Yin[a]*, Atin Adhikari[b], Xiaolu Zhou[c], Weitian Tong[d] and Lixin Li[d]

a Department of Biostatistics, Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, Georgia 30460. Email: jingkersey@gmail.com; jyin@georgiasouthern.edu

b Department of Epidemiology & Environmental Health Sciences, Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, Georgia 30460. Email: aadhikari@georgiasouthern.edu

c Department of Geology and Geography, College of Science and Mathematics, Georgia Southern University, Statesboro, Georgia 30460. Email: xzhou@georgiasouthern.edu

d Department of Computer Science, College of Engineering and Computing, Georgia Southern University, Statesboro, Georgia 30460. Email: wtong@georgiasouthern.edu; lli@georgiasouthern.edu

* Correspondence should be addressed to: Dr. Jingjing Yin, Georgia Southern University, PO Box 8015, Statesboro, Georgia 30460-8015, USA. Email: jyin@georgiasouthern.edu.

Abstract. Particulate matter 2.5 ($PM_{2.5}$) are fine particles can penetrate deeply into our lungs and other airways areas because of their small sizes. Sometimes these fine particles may even enter the bloodstreams. Only a few researches studied the relation between $PM_{2.5}$ and lung cancers. In this paper, innovative machine learning and spatiotemporal interpolation methods were used to compute historical $PM_{2.5}$ interpolation data in the contiguous United States. Time series analysis (including seasonal ARIMA models, lagged regressions, generalized estimating equations) is then applied to lung and bronchial cancers and $PM_{2.5}$ data. Based on our current data covering a 15-year span (1999-2014), $PM_{2.5}$ doesn't have a strong effect on lung and bronchial cancer rates in the United States at either the national or state level. However, the most urban state, New Jersey, and highest $PM_{2.5}$ state, California, have a relatively greater tendency to have significant $PM_{2.5}$ effect among all contiguous U.S. states.

Keywords: $PM_{2.5}$, lung and bronchial cancers, spatiotemporal interpolation, time series analysis, regression analysis

1 Introduction

According to the ISO 23210:2009 standard of the International Standards Organization, particulate matter 2.5 or $PM_{2.5}$ are fine particles which pass through a size-selective air sampling inlet with a 50% efficiency cut-off at 2.5 μm aerodynamic diameter size. Generally particulate matter with a mass median aerodynamic diameter of $\leq$ 2.5 μm are often termed as $PM_{2.5}$. $PM_{2.5}$ can penetrate deeply into our lungs and other airways areas because of their small sizes.

MOBIMEDIA 2018, June 21-22, Qingdao, People's Republic of China

DOI 10.4108/eai.21-6-2018.2276588

Sometimes these fine particles may even enter the bloodstreams. $PM_{2.5}$ can absorb gases or carry other finer toxic and carcinogenic chemicals due to their large surface areas[1][2]. The findings regarding the carcinogenicity of outdoor particulate matter are highly consistent in epidemiological research [3][4], studies of lung cancers in experimental animal models[5], and a wide range of in vitro studies focusing mechanisms related to cancer [6]. Increased risks of lung cancers were consistently reported in large-scale cohort studies as well as case-control studies involving millions of human subjects and thousands of lung cancer cases from different countries from Europe, North America, and Asia [3][4]. However, most of these studies addressed PM_{10} except a few epidemiological studies on the relationship between $PM_{2.5}$ and lung cancer mortality and incidences, which reported relative risks ranging from 1.04 to 1.43 with 95% CI ranging between 0.85 and 2.41[3][7][8][9]. Therefore, more information is required on the associations between $PM_{2.5}$ and lung cancer mortality using new analytical approaches. In this study, we have explored innovative machine learning, spatiotemporal interpolation and statistical methods to address this research issue.

2 Methods

2.1 Data Description

Lung and bronchial cancer data set and $PM_{2.5}$ data set are used in this project. Lung and bronchial cancer rates from 1999 to 2014 in the United States were downloaded from the National Program of Cancer Registries (NPCR) website at the national and states levels. Rates are per 100,000 persons and are age-adjusted to the 2000 U.S. standard population [10].

The $PM_{2.5}$ data was spatiotemporally interpolated and then aggregated at various spatial and temporal levels. We downloaded daily $PM_{2.5}$ data (1997-2015) in the contiguous U.S. from the Environmental Protection Agency (EPA) Air Quality System and aggregated them into monthly data with the schema (*longitude*, *lattidue*, *month*, *year*, *mean* $PM_{2.5}$). Attributes longitude and latitude are to locate the centroids of a U.S. census block group, and *mean* $PM_{2.5}$ is the average of daily values of $PM_{2.5}$ in the month at the centroid. Using innovative machine learning and spatiotemporal interpolation methods[11] [12], we trained the aggregated monthly data to find the optimal interpolation parameters, then interpolated at the centroids of census blocks. We then aggregated the mean $PM_{2.5}$ values at every centroid in each state to obtain the mean $PM_{2.5}$ value for each state at each month. Since the lung and bronchial cancer rates were collected annually, we also computed average of the mean $PM_{2.5}$ values at each month to get annual $PM_{2.5}$ value for each state. Finally, we took average across all states to obtain national $PM_{2.5}$ values.

The data of urban area was collected from U.S. Census Bureau. We used the ratio between urban area of cartographic boundary and the overall state area to represent the urban percentage [13].

2.2 Statistical Analysis

The main goal of this study is to examine the relation between lung and bronchial cancer annual rates and $PM_{2.5}$ values. Time series analysis including ARIMA models and seasonal ARIMA models [14] were used to predict the trend of lung and bronchial cancer rates (annually) and $PM_{2.5}$ values (both monthly and annually). The generalized estimating equation (GEE) [15] was used to estimate the parameters of the generalized linear regression models to determine possible association of urban percentage with lung and bronchial cancer rates and $PM_{2.5}$ values, respectively. Furthermore, lagged regression model [16] regressing lung and bronchial cancer rates on its past time-series and $PM_{2.5}$ time-series was used.

3 Results

3.1 $PM_{2.5}$

Average national $PM_{2.5}$ has a trend over time and it decreases in the long run. The optimal model by automatically model selection based on Bayesian Information Criterion (BIC) is determined to be seasonal ARIMA (2,1,1) (2,0,0) [12]. Consistently in each year, $PM_{2.5}$ is low

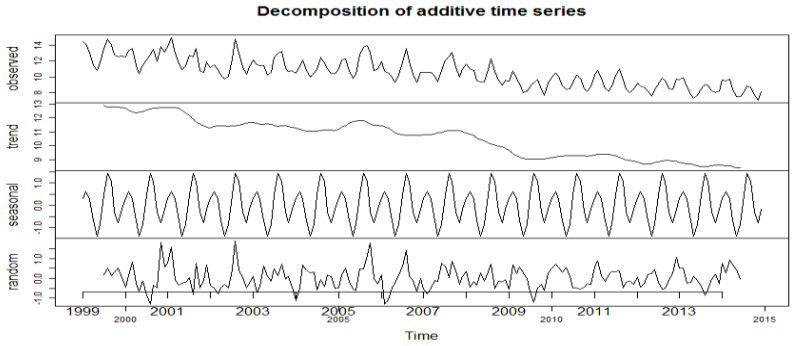

Figure 1.Trend and Seasonal Effects of National $PM_{2.5}$

in the summer months and peaks between fall and winter (See **Figure 1**).

In the generalized linear regression model, annual $PM_{2.5}$ value is strongly associated with year and its value decreases consistently each year (estimate = -0.0272, p-value < .0001). This is consistent with the ARIMA result. The regression analysis also discovered that annual $PM_{2.5}$ for individual state is significantly related to the urban percentage of that state - for every 1% increasing in urban percentage, the $PM_{2.5}$ value increases by 29.87% (with p-value = 0.0336). As years passing by, the effect of urban percentage becomes lesser (estimate = -0.0148, p-value = 0.0346).

3.2 National Lung and Bronchial Cancer Rates

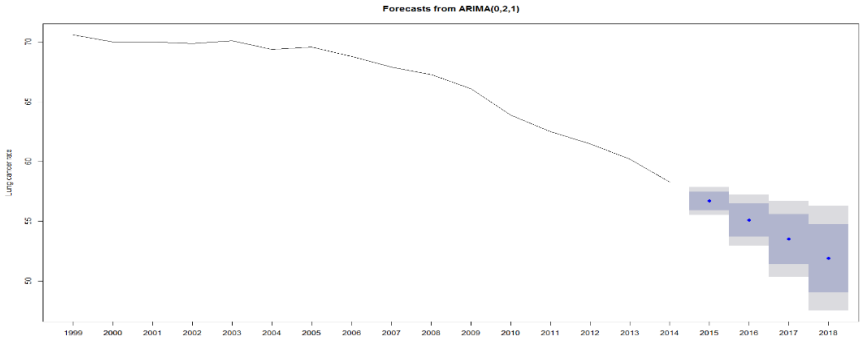

Figure 2. Lung Cancer Rates and Prediction for 2015-2018

National lung and bronchial cancer rates decreased over time. Particularly, it dropped greatly since 2005 (See **Figure 2**). Interestingly, this observed decrease is matching with the implementation of the fine particle ($PM_{2.5}$) National Ambient Air Quality Standards by the US Environmental Protection Agency (EPA) in 2005. The optimal time series model selected is ARIMA (0,2,1) and we give the prediction of rates for year 2015-2018 based on this model (See **Figure 2**). The national lung cancer rates seem to have strong association with national $PM_{2.5}$, lags of cancer rate, and lags of $PM_{2.5}$ (See **Figure 3** and **Figure 4**). However, further lagged regression analysis with the two time series suggested that $PM_{2.5}$ does not have strong causal effect on lung and bronchial cancer. Even though lags of cancer rates and lags of $PM_{2.5}$ are individually significant, the final selected optimal model only includes lag1 of cancer rate (cancer rate from previous year), lag1 of $PM_{2.5}$, and lag5 of $PM_{2.5}$, suggesting the time trend of adjacent year for lung and bronchial cancer rates plays a major role leading to its current rate.

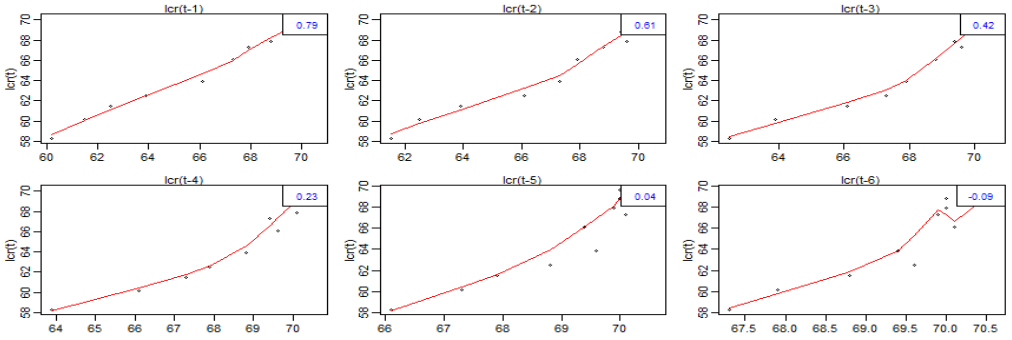

Figure 4. Lags of Lung Cancer Rates

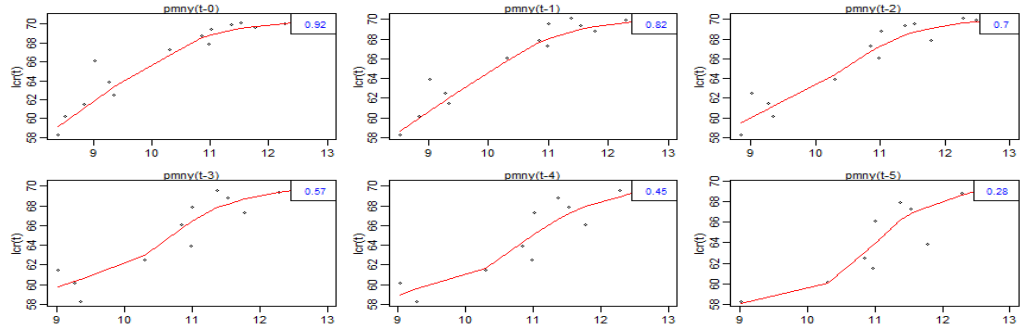

Figure 3. Lags of PM2.5 over Lung Cancer Rates

3.3 State Level Analysis

Since $PM_{2.5}$ is strongly associate with the urban-percentage of states, we continue the analysis for most rural state-Wyoming (WY), most urban state - New Jersey (NJ), Fully urban- District of Columbia (DC), and the state with highest $PM_{2.5}$ - California (CA).

Statistical analysis for those four areas carries similar results as the national level analysis. Lung cancer rates in all the four areas have significant trend over time, but they weren't

significantly affected by $PM_{2.5}$. Notice that for the most urban state-new Jersey (NJ) and highest $PM_{2.5}$ state-California (CA), the p-value is 0.0572 and 0.17, respectively, for $PM_{2.5}$ effect when considering trend effect and $PM_{2.5}$ effect.

4 Discussion

According to the U.S. EPA factsheet, on September 8, 2005, the EPA proposed requirements that state and local governments have to meet as they implement the national ambient air quality standards for fine particulate matter ($PM_{2.5}$). EPA established the $PM_{2.5}$ standards in 1997 and designated areas as attainment or nonattainment in December 2004. This proposed rule was the next step toward improving particle pollution air quality for the U.S. population. The proposed rule described the implementation framework and requirements that state, local, and tribal governments must and EPA instructed that implementation plan must show how an area that is not attaining the $PM_{2.5}$ standards will reduce air pollutant emissions in order to meet the standards as soon as possible.

Although a few previous US and European studies reported associations between particulate matter and lung cancer in some selected populations groups, at both national level and state level, we found that $PM_{2.5}$ doesn't have strong effect on lung and bronchial cancer rates in the United States based on data of 15 years span (1999-2014) in the current study. Historic data of $PM_{2.5}$ and lung and bronchial cancers over a longer period of time would help future research to see if there is an association between $PM_{2.5}$ and lung cancer.

The most urban state - New Jersey (NJ) and highest $PM_{2.5}$ state - California (CA) have relatively more tendency to have significant $PM_{2.5}$ effect among all contiguous U.S. states. It indicates when $PM_{2.5}$ level is higher, there might be some association between $PM_{2.5}$ and lung and bronchial cancer rates. Therefore, it would be very interesting to conduct similar research in developing countries, such as in China and India, which have significantly higher $PM_{2.5}$ concentration levels.

One limitation of the study is that the monthly and yearly $PM_{2.5}$ were calculated at each state in the contiguous U.S. using arithmetic mean of $PM_{2.5}$ concentration values at the centroids of census block groups for each state. We calculated these national estimates assuming the centroid data in our study is a stratified random sample from the contiguous U.S. which may not be the case thus the averages may not be accurate. Some weighted mean, such as based on the density of centroids or population would be a better approach.

Urban percentage is found to be strongly related to $PM_{2.5}$. Current study only looked at the national and state level. We can also group states with similar characteristics, such as by urban percentage, population sizes, or/and geographical adjacency for future work.

The analyses can also be extended to finer-scale areas, for example by counties in United states. While $PM_{2.5}$ data are available at this scale, the availability of lung and bronchial cancer data may be a problem.

It takes decades to manifest lung cancer and there are many other factors (such has air pollutants other than $PM_{2.5}$, family genetics and medical history, as well as behavior, indoor environmental exposures, etc.) contributing to the development of the cancer. To further investigate how air pollution contributes to lung cancer, mixture of different pollutants combined other factors over a long-time span should be examined.

References

[1] Lonati, G., Giugliano, M., Butelli, P., Romele, L. and Tardivo, R.: Major chemical components of PM2. 5 in Milan (Italy). Atmospheric Environment. 39(10):1925-34 (2005)

[2] Zheng, M., Cass, G.R., Schauer, J.J. and Edgerton, E.S.: Source apportionment of PM2. 5 in the southeastern United States using solvent-extractable organic compounds as tracers. Environmental science & technology, 36(11), pp.2361-2371 (2002)

[3] Raaschou-Nielsen, O., Andersen, Z.J., Beelen, R., et al.: Air pollution and lung cancer incidence in 17 European cohorts: prospective analyses from the European Study of Cohorts for Air Pollution Eff ects (ESCAPE). 14: 813–22. Lancet Oncol (2013)

[4] Krewski, D., Jerrett, M., Burnett, R.T., et al.: Extended follow-up and spatial analysis of the American Cancer Society study linking particulate air pollution and mortality. Res Rep Health Eff Inst 2009, 140: 5–114 (2009)

[5] Zeidler-Erdely, P.C., Meighan, T.G., Erdely, A., et al.: Lung tumor promotion by chromium-containing welding particulate matter in a mouse model. Particle Fibre Toxicol, 10: 45 (2013).

[6] Knaapen, A.M., Borm, P.J., Albrecht, C. and Schins, R.P.: Inhaled particles and lung cancer. Part A: Mechanisms. Int J Cancer. 109:799-809 (2004).

[7] Pope, C.A. 3rd, Burnett, R.T., Thun, M.J., et al.: Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA, 287:1132-41(2002)

[8] Krewski, D., Burnett, R.T., Goldberg, M., et al.: Reanalysis of the Harvard Six Cities Study, part I: validation and replication. Inhalation Toxicology 17:335-42 (2005)

[9] Carey, I.M., Atkinson, R.W., Kent, A.J., et al.: Mortality associations with long-term exposure to outdoor air pollution in a national English cohort. Am J Respir Crit Care Med. 187:1226-33 (2013)

[10] U.S. Cancer Statistics Working Group: United States Cancer Statistics: 1999-2014 Incidence and Mortality Web-based Report. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute. Available at: www.cdc.gov/uscs (2017)

[11] Tong,W., Franklin, J. , Zhou, X., Li, L.,and Besenyi, G. : Machine learning on spark for the optimal IDW-based spatiotemporal interpolation. In Proceedings of the 9th International Conference on Geographic Information Science (GISscience), 336-339 (2016).

[12] Tong, W., Li, L. , Zhou, X. , Franklin, J., Besenyi, G. , and Yates, H.: Learning with spark for the optimal IDW-based spatiotemporal interpolation. Applied Geomatics. Under review. Submitted on Oct. 26, 2017.

[13] US Census Bureau: https://www.census.gov/geo/maps-data/data/cbf/cbf_ua.html (2018)

[14] Robert H. Shumway, R.H. and Stoffer, D. s.: Time Series Analysis and Its Applications. Chapter 3. eBook ISBN: 978-3-319-52452-8. 4th Edition. Springer International Publishing, (2017)

[15] Hardin, J. W., & Hilbe, J. M.: *Generalized estimating equations*. John Wiley & Sons, Inc. (2003)

[16] Robert H. Shumway, R.H. and Stoffer, D. s.: Time Series Analysis and Its Applications. pp.218-222. eBook ISBN: 978-3-319-52452-8. 4th Edition. Springer International Publishing, (2017)

[17] EPA:https://www3.epa.gov/pmdesignations/1997standards/documents/Sep05/factsheet.htm (accessed on March 6, 2018).