Using Machine Learning Approach to Evaluate the PM2.5 Concentrations in China from 1998 to 2016

Li Lin, Liping Di*, Ruixin Yang, Chen Zhang, Eugene Yu, Md. Shahinoor Rahman, Ziheng Sun, Junmei Tang
Center for Spatial Information Science and Systems (CSISS), George Mason University
Fairfax, VA 22030, USA.
{llin2,ldi*}@gmu.edu

Abstract— **Pollution is one of the main negative outcomes for rapid economic growth without sustainable development in China. Different types of pollutions are harming people's health and the impacts of pollution on environment and people's health could last for decades. Fine particulate matter(PM2.5), which is one of most common types of air pollutions in China, could penetrate and sediment in human's respiratory system and cause different kind of respiratory diseases. Research has shown the strong association between Aerosol Optical Depth (AOD) and PM2.5. For this reason, remote sensing imagery could be used to estimate the level of PM2.5 concentration near ground. With utilizing PM2.5 dataset estimated by Socioeconomic Data and Applications Center (SEDAC) and machine learning approach, this paper is aimed to provide spatiotemporal comparison of PM2.5 concentrations in China. Result from this analysis could help people to better understand the recent history and current status of PM2.5 pollution in China.**

Keywords—Remote Sensing; MODIS; PM2.5, Air Quality

I. Introduction

Pollution is one of the main negative outcomes for rapid economic growth in China [1]. China started suffering from air pollution since early 2000s. Although it is not easy to identify all sources of PM2.5, research has grouped PM2.5 into two categories based on its formation processes: primary and secondary sources [1]. Primary source includes all sources release PM2.5 into air directly, and all other non-direct PM2.5 emissions are considered as secondary sources [1]. Despite PM2.5 came from diverse sources, the outcomes are similar and research on the impact of PM2.5 has taken place in developed countries for many years. On the contrast, PM2.5 related studies in developing counties did not bring enough attrition until recent years.

Recent research summarized the influences of excessive PM2.5 pollution in China for past few decades [1]. First of all, visibility may be reduced when the concentration of PM2.5 is high in atmosphere [1]. Some scientists also suggested that high PM2.5 concentrations in air is also associated with regional and global climate change [1]. Significant health threat is one of worst impacts from PM2.5 pollution. According to a research in 2015, air pollution leaded to 1.6 million mortality each year in China [2]. Although it is an estimation based on statistical models, the impact to people's health from air pollution is unignorably. [1] summarized studies on PM2.5 and human health and concluded the existence of association between PM2.5 and mortality. Fine particulate matter(PM2.5) could sediment in human's respiratory system and cause serious health problems [1], [3], [4].

Remote sensing technique has been widely utilized in many different fields [5], [6], [7], [8], [9], [10], [11], [12], [13]. It is also used to monitor and analysis PM2.5 became popular since the ground measurement of PM2.5 concentration was not existing until early 2000s. With utilizing PM2.5 dataset generated by National Aeronautics and Space Administration Socioeconomic Data and Applications Center (NASA SEDAC), this paper will conduct spatial and temporal analysis on the PM2.5 concentration for China. 400 cities were evaluated to see if any spatial or temporal variation and trend exist. The paper also aims to group cities by PM2.5 concentration and temporal variation using machine learning approaches.

II. Data

Most early studies of air pollution in China were not started until the beginning of twenty-first century [1]. The monitoring and collecting of air pollution data began in less than 10 years, and massive studies on air pollution were initiated after the establishment of China's PM2.5 air quality standard [1]. Many cities provided air quality observation data has been widely used in research to study air pollution in China. However, research has found evidences showing more than fifty cities' self-reported PM2.5 data from 2000 to 2010 were manipulated [14]. Alternatively, remote sensing imageries provides objectively data source to estimate air pollution in China.

Aerosol Optical Depth (AOD), which could be used to measure PM2.5 concentrations by calculating scattering of light in atmosphere, could be obtained from multiple satellites [4]. PM2.5 annual average grid data (1998 to 2016) used in this research was derived from MODIS, MISR and SeaWiFS Aerosol Optical Depth (AOD) with Geographically Weighted Regression GWR [15]. The product was implemented and distributed by National Aeronautics and Space Administration Socioeconomic Data and Applications Center (NASA SEDAC) [15]. Near surface PM2.5 was estimated using GEOS-Chem chemical transport model and adjusted by Geographically Weighted Regression at 0.01degree resolution for most of the world [15].

Some critical values of PM2.5 concentration discussed in this paper were observed from two official agencies: United States Environmental Protection Agency (EPA), and Ministry of

978-1-5386-5038-7/18/$31.00 ©2018 IEEE

Environmental Protection (MEP, China). Both agencies provide national standards on annual average PM2.5 concentration (ug/m3). EPA revised the standard in 2012 (15 ug/m3 prior to 2012), and now the standard for annual PM2.5 is 12 ug/m3. Meanwhile, the China's standard is 15 ug/m3 for national park and 35 ug/m3 for all other places. PM2.5 standards from World Health Organization [16] is 10 ug/m3 which is stricter, but "WHO Air Quality Guidelines" also stated that annual average PM2.5 for developing cities is about 35 ug/m3 (World Health Organization, 2016).

III. Method

PM2.5 grid data from 1998 to 2016 for China was subtracted from the entire dataset using ArcGIS. Three-year-mean PM 2.5 between 1998-2000 and 2014-2016 were calculated to demonstrate the overall air pollution status (Figure 1&2). From two maps produced from previous step, a map of PM2.5 difference was calculated to quantify the net increase of PM2.5 for China (Figure 3). Results were classified into different categories according to EPA standard and China Air Quality Standard.

Information of locations and rank-by-size for 400 cities in China was downloaded from Natural Earth (Hongkong, Macao and Taiwan were not included in this study), an open map dataset available to public. Depended on the size ranking, 400 cities were grouped into eight classes. Spatial average of PM2.5 concentrations for 400 cities was calculate for year 1998 to 2016. Result was tested to see if temporal trend exists for cities using Mann-Kendall (MK) method. Temporal mean (1998 to 2016) PM2.5 concentrations were calculated by the size of city, from largest to smallest. Result was evaluated to see if PM2.5 concentrations are different when cities' sizes change. Net PM2.5 increases for different groups of cities then calculated. Statistical analyzes were conducted to see if PM2.5 concentration increase speed changes for different size of cities.

Hierarchical clustering approach was adopted in order to find pattern on PM2.5 concentration and temporal variation using open source Scipy library in Python. 31 cities were selected from 400 cities to avoid spatial autocorrelation. A dendrogram was generated to compare the pattern distance between cities. Different cluster numbers were tested to find the best classification result. Patterns for classes from hierarchical clustering method were visually compared.

IV. Result and Discussion

Few three-year- mean PM2.5 concentrations maps for China were generated and classified using old EPA and MEP's standards. Three-years-mean PM2.5 map from 1998 to 2000 shown that most regions of China met MEP's PM2.5 standard (<35 ug/m3) while about half of China were within EPA's PM2.5 standard (<15 ug/m3) (Figure 1). Three-years-mean PM2.5 map from 2014 to 2016 shown that total area met old EPA's PM2.5 standard (<15 ug/m3) decreased significantly (Figure 2). Meanwhile, areas which disqualifying both EPA and MEP's standards increased dramatically. From these maps, it is clear to see two large regions suffered from PM2.5 pollution: 1) Northeast China provinces, 2) Mid-east China (provinces between two mega cites: Beijing and Shanghai). In addition, many high PM2.5 concentration hotspots could be identified from the maps.

A net PM2.5 increase map was generated from three-years-mean PM2.5 for 1998 to 2000 and 2014 to 2016 (Figure 3). Result shows that air quality for most regions of China was decreasing between 1998 to 2016 with only few exceptions. The increase of PM2.5 concentrations level for two regions mentioned earlier in this paper (Northeast and Mid-east China) is more than two time over current EPA standard (12 ug/m3).

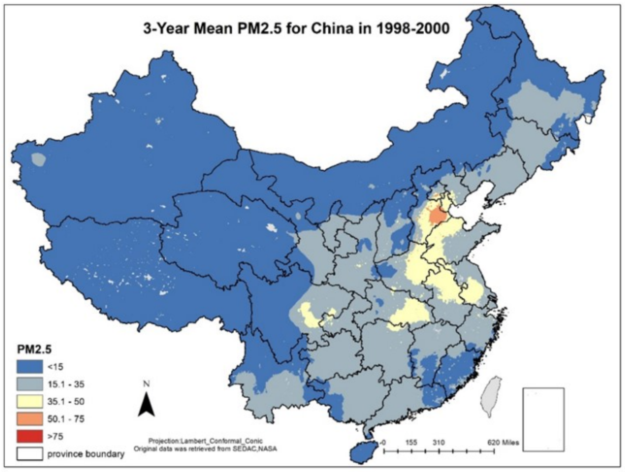
3-Year Mean PM2.5 for China in 1998-2000
PM2.5
<15
15.1 - 35
35.1 - 50
50.1 - 75
>75
province boundary
N
Projection:Lambert_Conformal_Conic
Original data was retrieved from SEDAC,NASA
0
155
310
620 Miles

Fig. 1. 3-year mean PM2.5 concentration for 1998-2000.

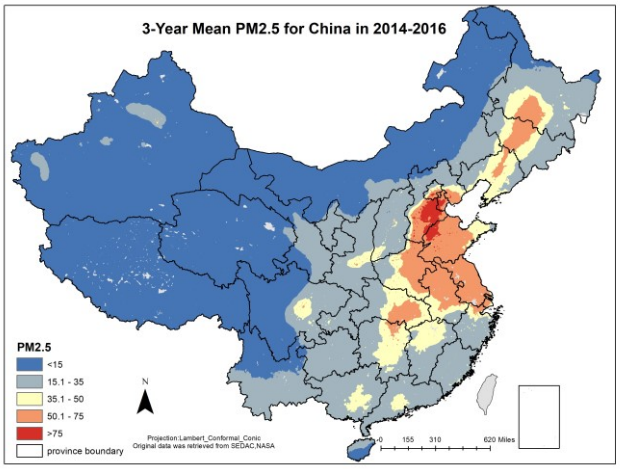
3-Year Mean PM2.5 for China in 2014-2016
PM2.5
<15
15.1 - 35
35.1 - 50
50.1 - 75
>75
province boundary
N
Projection:Lambert_Conformal_Conic
Original data was retrieved from SEDAC,NASA
0
155
310
620 Miles

Fig. 2. 3-year mean PM2.5 concentration for 2014-2016.

978-1-5386-5038-7/18/$31.00 ©2018 IEEE

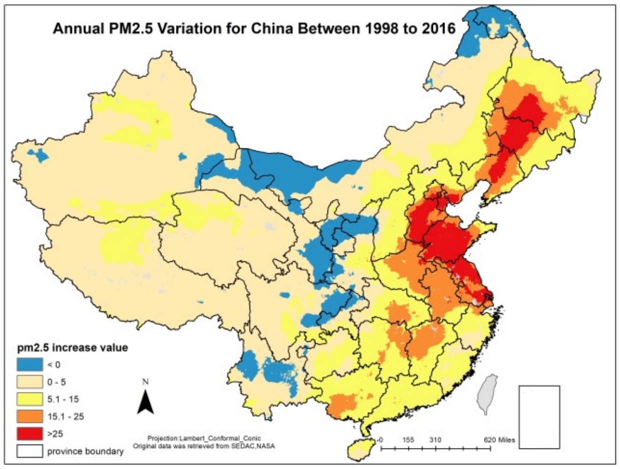
Annual PM2.5 Variation for China Between 1998 to 2016
pm2.5 increase value
< 0
0 - 5
5.1 - 15
15.1 - 25
>25
province boundary
N
Projection:Lambert_Conformal_Conic
Original data was retrieved from SEDAC,NASA
0 155 310 620 Miles

Fig. 3. Net PM2.5 increase from 1998 to 2016.

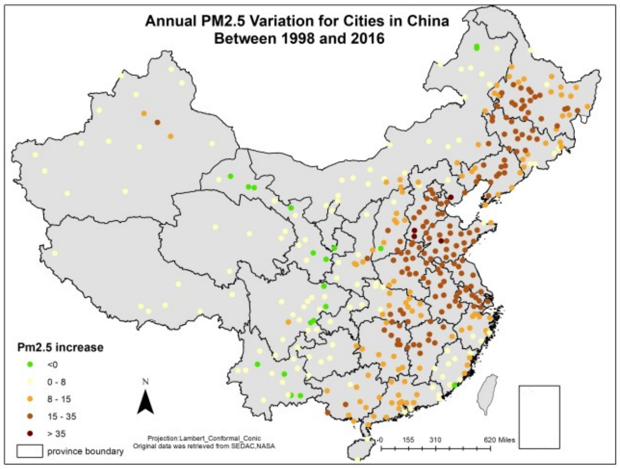
Annual PM2.5 Variation for Cities in China
Between 1998 and 2016
Pm2.5 increase
<0
0 - 8
8 - 15
15 - 35
> 35
province boundary
N
Projection:Lambert_Conformal_Conic
Original data was retrieved from SEDAC,NASA
0 155 310 620 Miles

Fig. 4. Net PM2.5 increase for cities in China.

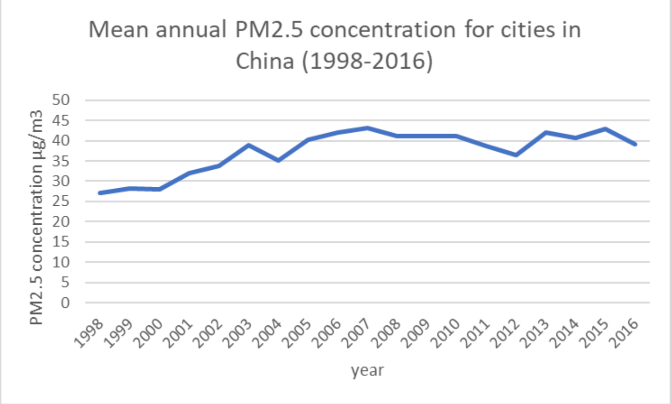

Fig. 5. Temporal trend of PM2.5 increase for cities in China.

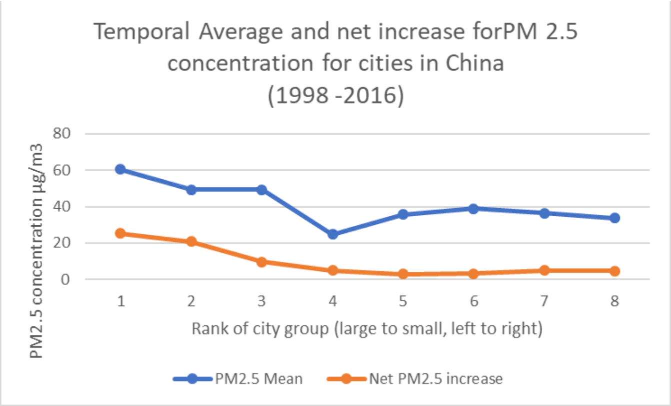

Fig. 6. Temporal average and net increase of PM2.5 concentration for different scale of cities. (large to small city rank, left to right)

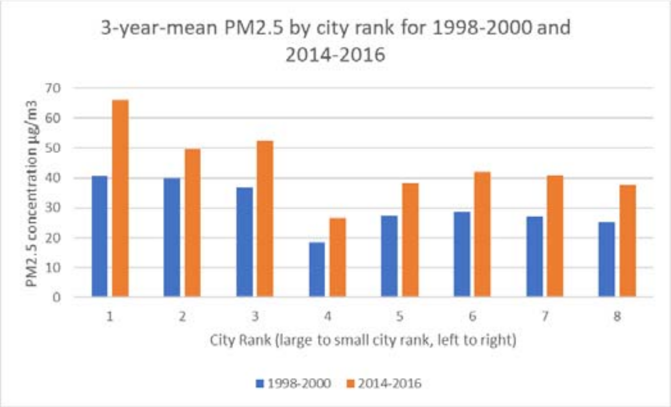

Fig. 7. PM2.5 concentration for different scale of cities (1998-2000 and 2014-2016) (large to small city rank, left to right)

Temporal average PM2.5 concentration suggested that 48 cities met the EPA standard, 136 cities failed in EPA standard but passed MEP standard, and 216 of China's cities failed in both standards (Figure 4). Mean PM2.5 concentrations level from 1998 to 2016 for all cities was calculated and compared. Result shows there was an increase trend of PM2.5 concentration from 1998 to 2016 (Figure 5). Mann-Kendall (MK) test suggested the trend is significant with p-value <0.01. 400 cities were grouped into 8 classed based on their size. For each class, average annual PM2.5 concentrations were calculated, and result implied that concentrations in larger cities are not different from smaller cities (95% confidence level) (Figure 6). However, PM2.5 concentration increase speed in larger cities are significant faster than smaller city(p<0.05). Figure 7 summarized the average PM2.5 concentrations between 1998 to 2000 and 2014 to 2016, the result shows cities in rank 4 have low annual PM2.5 concentration in both 1998-2000 and 2014-2016 compare with other cities from visual comparison.

PM2.5 concentration has strong spatial autocorrelation due to the air circulation. To avoid spatial autocorrelation, 31 were selected from 400 cities: 22 provinces' capitals, 5 capitals of municipalities and 4 municipality cities. Hierarchical clustering was used to classify cities into various categories base on their temporal trend. After tested results with 2 to 10 classes, we found best result exist when data divided into three class and the temporal patterns were meaningful (figure 8). Cities in first group (Aggressive growth) has a high PM2.5 concentration level in 1998 and significantly increased during study period (Figure 9). Cities in second group (Moderate growth) has a relative lower PM2.5 concentration level in 1998 and relative slow growth rate in PM2.5 compared with first group (Figure 9). Third group (No significant growth) has low PM2.5 concentration level across the study period (Figure 9). Cities in Aggressive growth class are spatial close to each other, while other two classes do not seem to have geographic correlation. Table 1 listed classification result for all 31 studied cities.

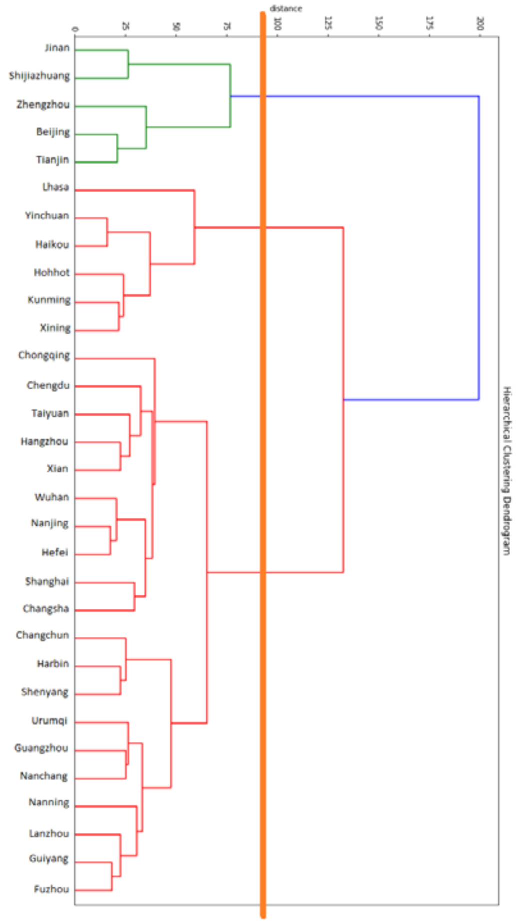

Fig. 8. Hierarchical Clustering Dendrogram with 3 classes.

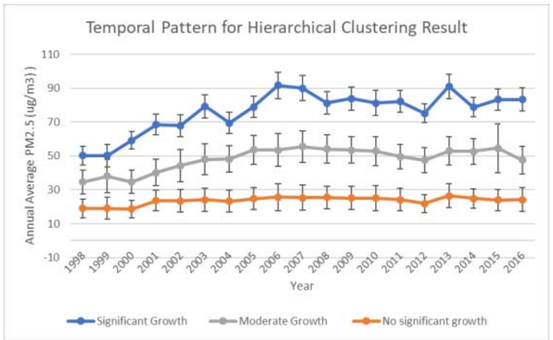

Fig. 9. Temporal pattern for 3 classes from unsupervised classification.

Group	Cities
Aggressive growth	Beijing, Tianjin, Jinan, Shijiazhuang, Zhengzhou
Moderate growth	Shanghai, Urumqi, Chengdu, Lanzhou, Guiyang, Nanning, Fuzhou, Changchun, Guangzhou, Harbin, Nanchang, Changsha, Taiyuan, Wuhan, Nanjing, Hangzhou, Hefei, Shenyang, Chongqing, Xian
No significant growth	Kunming, Lhasa, Yinchuan, Xining, Hohhot, Haikou

Table 1: Detail list of cities in each classification results

V. Conclusion and Future Work

The association between PM2.5 and mortality is proved in many studies. Reducing the PM2.5 concentration will help to improve human's living quality and reduce mortality. World Health Organization estimated that there would be 15% less air-pollution-related death if PM2.5 concentrations reduced from 35ug/m3 to 10ug/m3 (World Health Organization, 2016). The understanding of PM2.5 concentrations' spatial and temporal distribution is critical and urgent for developing counties which are suffering from air pollution such as China.

Remote sensing based air quality monitoring and analyzing became more important and valuable given the unreliability of PM2.5 ground measurements which are self-reported by local governments. Using serval methods and multiple satellites data, NASA Socioeconomic Data and Applications Center (SEDAC) provided annual average PM2.5 data from 1998 to 2016 [15]. PM2.5 standards discussed in this research were observed from United States Environmental Protection Agency (EPA), Ministry of Environmental Protection (MEP, China), and World Health Organization [16]. This paper utilized these datasets and standard to study spatial and temporal variation of PM 2.5 concentration in China from 1998 to 2016. In addition, analyses were performed over 400 cities in China to test the if there is a temporal trend for air pollution in China. This paper also tried to find if there is any association between size of cities and PM2.5 concentrations.

Result suggested that PM 2.5 concentration increased significantly from 1998 to 2016 for 400 cities in China ($p<0.01$) regardless of the size of city. Moreover, two heavy air pollution impacted regions were identified: north-east provinces of China (Heilongjiang, Jilin, and Liaoning) and mid-east provinces of China (Beijing, Tianjin, Hebei, Henan, Shandong, Hubei, Jiangsu, Anhui, and Shanghai). Result also implied that PM2.5 concentration increasing speed is higher when a city is larger($p<0.1$) based on the data from 1998 to 2016.

Unsupervised classification method was utilized in this research to detect temporal trend patter for the PM2.5 concentration variation between 1998 and 2016. To avoid spatial autocorrelation, 31 cities were selected in the experiment including 22 provinces’ capitals, 5 capitals of municipalities and 4 municipality cities. Three types of PM2.5 temporal patterns were identified in the result: 1) Aggressive growth (5 cities), 2) Moderate growth (20 cities), 3) No significant growth (6 cities). Following observations could be found from these three categories: cities with aggressive PM2.5 growths are spatially closer to each other compare with other two classes. Aggressive growth class has much rapid growth rate. Other two classes do not have spatial similarities. Cities in third class do not have significant economic growth during study period so the stable PM2.5 concentrations may be related with stable economic environment. Further research is needed on following aspects: 1) further evaluate the classification results; 2) identity the cause of these different temporal patterns; 3) find approaches to reduce PM2.5 concentration to meet MEP standard.

REFERENCES

[1] Y. Lin, J. Zou, W. Yang, and C.-Q. Li, “A Review of Recent Advances in Research on PM2. 5 in China,” *International journal of environmental research and public health*, vol. 15, no. 3, p. 438, 2018.

[2] R. A. Rohde and R. A. Muller, “Air pollution in China: mapping of concentrations and sources,” *PloS one*, vol. 10, no. 8, p. e0135749, 2015.

[3] E. Butt *et al.*, “Global and regional trends in particulate air pollution and attributable health burden over the past 50 years,” *Environmental Research Letters*, vol. 12, no. 10, p. 104017, 2017.

[4] S. Munir, S. Gabr, T. M. Habeebullah, and M. A. Janajrah, “Spatiotemporal analysis of fine particulate matter (PM2. 5) in Saudi Arabia using remote sensing data,” *The Egyptian Journal of Remote Sensing and Space Science*, vol. 19, no. 2, pp. 195–205, 2016.

[5] L. Guo, Z. Sun, L. Di, and L. Lin, “Spatial distribution and variation analysis of Lyme disease in the Northeastern United States,” in *Agro-Geoinformatics (Agro-Geoinformatics), 2016 Fifth International Conference on*, 2016, pp. 1–4.

[6] L. Lin et al., “A review of remote sensing in flood assessment,” in Agro-Geoinformatics (Agro-Geoinformatics), 2016 Fifth International Conference on, 2016, pp. 1–4.

[7] L. Lin *et al.*, “Extract flood duration from Dartmouth Flood Observatory flood product,” in *Agro-Geoinformatics, 2017 6th International Conference on*, 2017, pp. 1–4.

[8] L. Lin, L. Di, C. Zhang, L. Hu, J. Tang, and E. Yu, “Developing a Web service based application for demographic information modeling and analyzing,” in *Agro-Geoinformatics, 2017 6th International Conference on*, 2017, pp. 1–5.

[9] M. S. Rahman *et al.*, “Comparison of selected noise reduction techniques for MODIS daily NDVI: An empirical analysis on corn and soybean,” in *Agro-Geoinformatics (Agro-Geoinformatics), 2016 Fifth International Conference on*, 2016, pp. 1–5.

[10] M. S. Rahman *et al.*, “Agriculture Flood Mapping with Soil Moisture Active Passive (SMAP) Data: A Case of 2016 Louisiana Flood.”

[11] R. Shrestha *et al.*, “Regression based corn yield assessment using modis based daily ndvi in iowa state,” in *Agro-Geoinformatics (Agro-Geoinformatics), 2016 Fifth International Conference on*, 2016, pp. 1–5.

[12] R. Shrestha *et al.*, “Crop Fraction Layer (CFL) datasets derived through MODIS and LandSat for the Continental US from.”

[13] Q. Wu, M. Liu, X. Wang, L. Di, L. Kang, and L. Lin, “Assessing the water environmental capacity of pollution consumption in Jiulong River Basin,” in *Agro-Geoinformatics (Agro-geoinformatics), 2015 Fourth International Conference on*, 2015, pp. 318–323.

[14] D. Ghanem and J. Zhang, “‘Effortless Perfection:’Do Chinese cities manipulate air pollution data?,” *Journal of Environmental Economics and Management*, vol. 68, no. 2, pp. 203–225, 2014.

[15] A. van Donkelaar *et al.*, “Global Annual PM2.5 Grids from MODIS, MISR and SeaWiFS Aerosol Optical Depth (AOD) with GWR, 1998-2016.” NASA Socioeconomic Data and Applications Center (SEDAC), 2018.

[16] W. H. Organization (WHO) and others, “WHO Expert Consultation: Available Evidence for the Future Update of the WHO Global Air Quality Guidelines (AQGs),” *WHO: Geneva, Switzerland*, 2016.

978-1-5386-5038-7/18/$31.00 ©2018 IEEE