**REPORT ON PROPOSED RESEARCH PLAN 1**

**Written by: Dinh Thi Hong Hue**

**Proposed Research:** Using Machine Learning forecasting the PM 2.5 and Carbonaceous components concentration three days in advance, predicting the sources and pollutants distribution in Gwangju, Korea.

1. **Introduction**

Particulate matter, a mixture of particles and liquid droplets, has been considering as an important air quality issue in urban area due to their effect not only on animal and plant grow, but also the human health. According to US EPA, particulate matter is on the list of six common pollutants divided into two categories through their size including: inhalable coarse and fine particles. The inhalable coarse particles are particulate matter with aerodynamic diameter larger than 2.5 µm and smaller than 10 µm. The coarse particles are come from industrial operation (grinding and crushing process) and near the road with stirred-up dust from vehicles. Fine particulate matter, is particulate matter with aerodynamic diameters smaller than 2.5 µm, emitted from combustion sources including: vehicles, forest fires, wood burning, industrial operation, agricultural burning and power plant etc. Many studies reported the adverse human health effect related to aerosol particles size (US EPA; Anderson et al., 2012; Mengfei et al., 2015). The coarse particles can get into lung; while the fine particles even easily penetrate deeper into bloodstream and cause of variety health problems (US EPA; Anderson et al., 2012; Mengfei et al., 2015). Besides, the condensation of pollutants in ambient air with particles, especially fine particle increases the carcinogenic potential in respiratory system. According to an abundance of studies, the particulate matter combined with microscopic solid or liquid, especially fine particles which enter deeply into lung and heart cause serious health problems including (US EPA): irregular heartbeat, reduced lung function, increased respiratory symptoms (Irritated air way, coughing, breathing troubles etc.), premature death in heart and lung diseases, aggravated asthma and heart attacks. Therefore, it’s necessary to evaluate fine particulate matter and combined contaminants to investigate the human health impact.

Carbonaceous aerosol is a substantial factor in climate forcing, impaired visibility and adverse health effects (US EPA, 2004), contributing significant fraction to fine particulate matter (PM 2.5) (Seinfeld and Pandis, 1998). Carbonaceous aerosol is a crucially constituent of fine particles, especially in urban area. It is accounted for 30 – 70 % of PM 2.5 which high fraction of OC to TC (OC + EC). Indeed, the roadside pollution is considering as a particular concern from the risk of vehicle emission source. The primary source from transportation affect significantly portion to OC and EC concentration. In addition, the distribution of carbonaceous aerosols from vehicle emission source spread widely up to 200 to 300 m from the roadside (Young et al., 2016). Therefore, the measurement of carbonaceous species is essential focus on this study to better understand the level and evaluate the emission source as well as the effect of human activity to atmospheric environment.

Machine Learning is concentrated on the utilization of specialized algorithms in order to extract the trend and information from date sets (Catherine et. al, 2013). In the Environmental Engineering field, there is a large amount of data sets such as temperature, rainfall, contaminants concentrations etc. over various time (Catherine et. al, 2013). Any error information in data collection can reduce the data clarity as well as provide incomplete picture of environmental events. (Catherine et. al, 2013). In addition, the abundant quantity of data collection causes the analysis difficulty for traditional methods. By applying the machine learning techniques, the problems caused by the size of database and the efficiency of analysis can be solved (Catherine et. al, 2013; Babovic, 2005).

The tradition statistic models e.g. multi linear regression can calculate the cause-and-effect relationship, however, it is difficult to estimate the exposure degree in certain forecasting types (Mishra et al., 2015). The prediction of pollutants concentration in ambient air can be considered as a non-linear regression issue in relationship between predictors and predictands. By using artificial neural network (ANN) in ambient air quality, the non-linear data can be handled with self-learning capabilities (Hornik, 1991; Mishra. et al., 2015). A study of Kukkonen et al. (2003) predicted the concentration of PM10 and NO2 in Helsinki, Finland comparing five ANN models, a linear model and a deterministic model (Mishra. et al., 2015). The results showed that ANN models forecasted better result than other, especially for ANN models having non-constant variance. Mishra and Goyal (2015) forecasted the PCA-ANN model in Agra. The ANN model presented a surpassed result compared with MLR model.

Kwangju city, the sampling site of carbonaceous in PM 2.5, is located in between the mountainous and plain areas in the southwestern of South Korea. Kwangju is the sixth metro cities in South Korea with 1.49 million people in 501.18 km2. The climate of Kwangju is affected strongly by west coastal weather with warm weather and adequate precipitation. The city has cold winter effected from cold continental high-pressure system in Mongolia and hot summer from North Pacific Ocean high pressure system. Kwangju is regarded as a famous agricultural area as well as an industrial development area. Due to the increasing of economy and technology, the decreasing of air quality is considering. In statistic of Kwangju government, there were 47 new cars registered everyday (as Dec. 31st, 2013) (Statistic, [www.gwangju.go.kr](http://www.gwangju.go.kr)), ringing a bell for ambient air quality in Kwangju. Thus, vehicular emissions are evaluating as a large contributor to aerosol level in this city. In addition, many studies were reported the influence of local pollution and regional long-range pollution to air quality in Kwangju. Therefore, it’s necessary to consider the dependence of weather patterns and biomass burning emission on sampling site.

1. **The purposes**

* Predict the sources (primary sources or/and secondary sources) of atmospheric contaminants in order to reduce the pollution concentration
* Forecast the pollutants distribution and predict three days in advance.

1. **Proposed Methodology**

* The input includes the concentration of PM2.5, carbonaceous components (water-soluble organic carbon, organic carbon, elemental carbon, total carbon), oxalate and eight ion species (Na+, K+, NH4+, Ca2+, NO3+, Cl-, PO43-, SO42-), temperature (min and max), wind direction, wind speed, sunlight hour, **input of the two-next days**.
* The desired output will be the information of sources such as primary or/and secondary sources, the modeling of distribution as well as the predicted three-days in advance.
* Utilization of air mass trajectory based geographic model. The air mass trajectory model can be utilized to detect the location and direction of a variety of air contaminant sources. In order to identify the distribution of PM2.5 transporting to the sites, the Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) with backward model was applied in this study.
* A multi linear regression (MLR) model and ANN model were applied to predict the daily pollutant concentration in the study as well as determine the efficiency of techniques.

1. ***Database preparation:***

* The PM 2.5 samples were collected consecutively from urban and road site in Kwangju, Korea. The sampling periods were August 10th to September 10th, 2015 and December 8th, 2015 to January 15th, 2016. Sampling began from 9:00 AM to 9:00 AM of the next day.
* The sampling at urban site was carried out on the rooftop of three-story building in Chonnam National University campus (CNU) at 350 11’N, 1260 54’E. The building is placed approximately 150 m near Uchi – a busy road and 0.5 km southern of Honam express highway. The site is mainly surrounded by residential areas.
* The road site in the Bukgu Office (BK) is located about 0.5 km away from CNU site, sampling set at rooftop of three-story building in front of the Uchi round-about. Surrounding the sampling site, there are many main roads with high concentration of cars, trucks and buses, etc. especially on the rush hour. There is no other combustion source except vehicular source around the sampling site. Field blank filters in each site were also collected to correct their background. An hourly air monitoring station was measuring PM 2.5, NO2, CO, SO2 and O3 from Ministry of the Environment during sampling collection. The version of dataset for experiment is 70 days for summer and winter time. The range and average of PM 2.5 and carbonaceous components in two sites are listed as table 1 and table 2. While the concentration of water-soluble inorganic species is shown in the table 3. The Figure 1 illustrates the variation of PM 2.5 in summer and winter time in compared with Korean and US EPA standard.
* Data collection was divided into two sets including 80% to 85% for training process and 15% to 20% for testing. In the input layer, there are 18 variables applied into the models.

Table 1: Average and range of PM 2.5 concentration during the sampling periods

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Sampling  Site |  | Total | | Exceedance  Korean Standard | | Exceedance  US. EPA Standard | |
|  |  | Average  (µg/m3) | Range | Number | Average  (µg/m3) | Number | Average  (µg/m3) |
| CNU | Summer | 19.73 | 4.82–57.77 | 2 | 55.83 | 3 | 51.03 |
|  | Winter | 34.64 | 10.97–93.71 | 7 | 64.52 | 18 | 50.04 |
| BKO | Summer | 24.22 | 6.82–62.65 | 2 | 58.10 | 6 | 41.06 |
|  | Winter | 34.93 | 9.5–91.1 | 7 | 66.50 | 15 | 49.41 |



Figure 1: The PM 2.5 variation during summer (a) and winter season (b)

Table 2: Average and range concentration of carbonaceous compositions at sampling sites

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sampling site |  | CNU |  |  |  |  | BKO |  |  |  |
|  | Summer | | Winter | |  | Summer | | Winter | |
| Parameters |  | Mean | Range | Mean | Range |  | Mean | Range | Mean | Range |
| OC | µg/m3 | 4.07 | 1.97–6.84 | 5.02 | 1.37–12.2 |  | 5.41 | 3.09–8.76 | 5.48 | 1.21–12.18 |
| EC | µg/m3 | 1.45 | 0.69–2.58 | 1.0 | 0.44–2.02 |  | 1.58 | 0.87–2.29 | 1.05 | 0.39–1.82 |
| WSOC | µg/m3 | 1.91 | 1.0–2.69 | 2.49 | 0.95–5.98 |  | 2.82 | 1.81–4.23 | 2.49 | 1.02–5.01 |
| OC/EC | – | 2.84 | 2.05–4.29 | 5.07 | 2.44–9.12 |  | 3.43 | 2.59–4.94 | 5.22 | 2.75–9.46 |
| WSOC/OC | – | 0.38 | 0.22–0.46 | 0.54 | 0.32–0.82 |  | 0.42 | 0.24–0.59 | 0.49 | 0.33–0.84 |
| WSOC/ EC | – | 1.36 | 0.8–2.25 | 2.67 | 1.0–4.88 |  | 1.82 | 1.39–2.44 | 2.46 | 1.13–4.57 |

Table 3: Average of water-soluble inorganic species in sampling sites

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sampling site | CNU | | | | | |  | BKO | | | | | |
| Summer | | | Winter | | |  | Summer | | | Winter | | |
| Parameters | Range  (µg/m3) | Mean  (µg/m3) | %1 | Range  (µg/m3) | Mean  (µg/m3) | % |  | Range  (µg/m3) | Mean  (µg/m3) | % | Range  (µg/m3) | Mean  (µg/m3) | % |
| Na+ | 0.05–2.75 | 0.84 | 4.24 | 0.04–0.62 | 0.28 | 0.79 |  | 0.03 –2.96 | 0.62 | 2.56 | 0.006–0.61 | 0.16 | 0.45 |
| NH4+ | 0.13–11.08 | 2.38 | 12.07 | 0.89–11.82 | 3.63 | 10.46 |  | 0.18–9.30 | 2.58 | 10.64 | 0.71–9.85 | 3.42 | 9.80 |
| K+ | 0.03–0.61 | 0.15 | 0.77 | 0.07–0.60 | 0.21 | 0.59 |  | 0.003–0.60 | 0.14 | 0.59 | 0.07–0.54 | 0.21 | 0.61 |
| Cl- | 0.005–0.30 | 0.05 | 0.24 | 0.25–1.56 | 0.52 | 1.49 |  | 0.001–0.25 | 0.05 | 0.22 | 0.20–1.60 | 0.48 | 1.36 |
| NO3- | 0.16–10.34 | 1.79 | 9.07 | 1.10–24.36 | 7.55 | 21.80 |  | 0.50–8.84 | 2.03 | 8.36 | 1.08–20.35 | 7.19 | 20.58 |
| SO42- | 0.59–24.56 | 8.22 | 41.67 | 1.19–20.18 | 5.22 | 15.08 |  | 0.66–29.38 | 7.92 | 32.68 | 1.07–16.99 | 5.32 | 15.22 |
| Oxalate | 0.002–0.26 | 0.14 | 0.69 | 0.004–0.32 | 0.14 | 0.41 |  | 0.002–0.42 | 0.19 | 0.78 | 0.02–0.28 | 0.14 | 0.39 |
| nss K+ | 0.008–0.58 | 0.12 | 0.62 | 0.06–0.58 | 0.2 | 0.56 |  | 0.03–0.55 | 0.13 | 0.54 | 0.07–0.54 | 0.21 | 0.59 |
| ∑SI2 | 1.22–45.82 | 12.39 | 57.27 | 3.21–56.36 | 16.40 | 44.51 |  | 1.29–42.46 | 12.37 | 45.75 | 3.09–46.75 | 15.93 | 43.58 |
| NO3-/SO42- | 0.05–0.79 | 0.29 |  | 0.57–4.06 | 1.55 |  |  | 0.06–1.09 | 0.38 | - | 0.51–4.0 | 1.47 | - |

1The percentage (%) of species with PM 2.5 mass concentration is shown in the column (%).

2∑SI is abbreviation for Total Secondary Ions including NH4+, NO3-, and SO42- concentration.

1. ***Multi linear regression (MLR) model:***

* In MLR model, a function includes a dependent variable predicted from a variety of independent variables. The equation of MLR can be presented as:

Y = b1 + b2X1 + b3X2 + … + bmXm + e

Whereas b1, b2, b3, …, bm are linear parameters, X1, X2, …, Xm are independent variables, Y is dependent variable, e is estimated error. The estimating of b1, b2, b3, …, bm can be calculated by the least square error techniques (Mishra et al., 2015).

1. ***Artificial neural network (ANN) model:***

* In ANN model, a multi-layer perceptron of back-propagation neural network was utilized to forecast the next-three days concentration of contaminants. A logistic sigmoid function was calculated to transfer the values in hidden layer (Feng et al., 2015).
* A set of links are created by input neurons as synaptic weights of W1, W2, W3, …, Wm. A summing function computing weights of input as an adder function

The action function and the output function as where bj and bk are bias. The process of ANN is illustrated in Figure 2.

* There are one input layer, one output layer and **two proposed hidden layer**. The MLP is in hidden layer with 11 neurons and a square activation function (Mishra et al., 2015). The input layer receives the outside information fed-forward to the hidden layer. The activation function in hidden layer, is a non-linear function of the sum of weights, plays an important role on detecting input signal and fed-forward to the output layer. The output layer, in turn, releases the output as a linear function of activation function in the former layer.
* The error measurement of two models were conducted to compare the effectiveness such as the mean absolute error (MAE), the root mean square error (RMSE), and the index of agreement (IA), the correlation coefficient (R), fractional bias (FB). The model was developed via ANN toolbox in MATLAB.

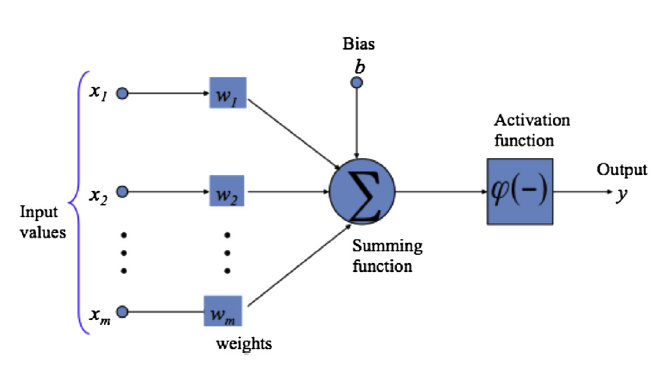


Figure 2: Schematic structure of artificial neural network (Mishra et al., 2015)

**References:**

1. US EPA (United States Environmental Protection Agency), Particulate matter
2. Mengfei Z., Ting Q., Zhongsi H., Mengya Z., Wei X., Guangli X., Jun T., Shuncheng L., “Comparison of ionic and carbonaceous compositions of PM 2.5 in 2009 and 2012 in Shanghai, China”, Science of the Total Environment 536 (2015) 695–703.
3. Seinfeld J. H., Pandis S. N., “Atmospheric chemistry and physics: from air pollution to climate change,” Wiley, New York, 1998.
4. Young S. G., Won S. R., Choi Y., Chang Y. S., Jin H. C., Kim Y. P., Kang C. H., “Concentration variations in primary and secondary particulate matter near a major road in Korea,” Asian Journal of Atmospheric Environment, vol. 10-1, pp 32 – 41, 2016.
5. Catherine W., Wei L. W., Zeyar A., “Applications of machine learning in environmental Engineering”, Technical Report DNA 2013-03.
6. Babovic V., “Data mining in hydrology,” Hydrological processes, vol. 19, no. 7, pp. 1511–1515, 2005.
7. Dhirendra Mishra, P. Goyal, Abhishek Upadhyay, “Artificial intelligence-based approach to forecast PM2.5 during haze episodes: A case study of Delhi, India”, Atmospheric Environment 102, pp. 239 – 248, 2015.
8. Xiao Feng, Qi Li, Yajie Zhu, Junxiong Hou, Lingyan Jin, Jingjie Wang, “Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation”, Atmospheric Environment 107, pp. 118 – 128, 2015.
9. Mishra, D., Goyal, P., “Development of artificial intelligence based NO2 forecasting models at Taj Mahal, Agra”, Atmos. Pollut. Res., 2015, <http://dx.doi.org/10.5094/apr.2015.012>.
10. Hornik, K., “Approximation capabilities of multilayer feed-forward networks”, Neural Network, vol. 4, pp. 251 – 257, 1991.
11. Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorling, S., Chatterton, T., Foxall, R., Cawley, G., “Extensive evaluation of neural network models for the prediction of NO2 and PM10 concentrations, compared with a deterministic modeling system and measurements in central Helsinki”, Atmos. Environ. vol. 37, pp. 4549 – 4550, 2003.