Contemporary Engineering Sciences, Vol. 9, 2016, no. 30, 1447 - 1462
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ces.2016.69161

A Preliminary Analysis Model of Big Data for Prevention of Bioaccumulation of Heavy Metal-Based Pollutants: Focusing on the Atmospheric Data Analyses for Smart Farm

Jun-Ho Huh and Kyungryong Seo[1]

Department of Computer Engineering, Pukyong National University at Daeyeon, Busan, Republic of Korea

Abstract

It is expected that the building-type fish or agricultural farms will appear in towns or suburbs of the Republic of Korea in the near future. The advanced water-circulating systems and LED technology are already in use helping the farmers to settle in these areas. Nevertheless, there are some other requirements for a successful operation of the farm. One of the surrounding factors is the atmospheric conditions which often have an impact on the growth of fishes or produce. This research describes a process of obtaining the analytical results by visualizing the data of atmospheric environment of the Gangnam District, one of the busiest areas in the capital city of Seoul. The data was used to create a model for the preliminary big data analysis which will be appropriate for the establishment of adequate city-oriented farms. Also, this model can be used to establish countermeasures to suppress bioaccumulation of heavy metals in the farm-grown fishes or produces. The basic research process involves visualization

[1]Corresponding author

of data obtained from the univariate, simple and multiple regression analyses so that the data can be viewed conveniently. This also facilitates the procedures of finding a log-transformed model and modeling of overall characteristics through categorization of the explanatory variables.

Keywords: Pollutant Monitoring, Preliminary Analysis Modeling, Optimal Site Selection, Aquaculture, Agriculture, R, Smart Farm

1 Introduction

Led by the Ministry of Maritime Affairs and Fisheries of the Republic of Korea (ROK), the researches on the building –type fish farms are progressing rapidly and some promising results are starting to appear [1-8]. This will no doubt advance the emergence of city or suburb-based farms in many densely populated areas. Moreover, the method of using the LED lights in the crop farms are already spreading due to the consumers' preference of organic produce. Both factors facilitate convenient construction of farms in the cities or suburbs. Additional factors would be the technological advancement of farm-related technologies such as the efficient water circulation systems which minimize the quality-water requirements, or the nutrient supply system that can add required nutrients automatically. Thus, this study focuses on the other factor that is essential for growing fishes or crops within the cities and suburbs where the atmospheric environments are much inferior to those in rural areas. Different from other factors, atmospheric environments cannot be controlled artificially but farmers still have to find a way to deal with them.

The ROK is bordering with China where industries are booming and many air-polluting gases and heavy metals are being increasingly emitted for the factories. These pollutants are mixed into the yellow sands or micro-dusts that are carried by the winds toward the ROK and accumulated in the foods that Koreans take. Because of the geographical nature, most of Korean territories are influenced by these pollutants and some countermeasures have to be taken. This study is to provide a convenient way to visualize an accurate atmospheric data obtained through the statistical analyses. To achieve this purpose, a data sample of Gangnam District, which was provided by the Seoul Metropolitan Government, was used for the analysis.

2 Related Research

The scale of big data is much more larger than that of the data generated from the analog environment of the past, shorter in generation cycles, and not only the

numerical data but the character and image data are included in the big data as well. Since the use of PC, internet or mobile devices has become people's daily routine, the volume of data left behind by them is increasing rapidly [3-12].

Along with the fact that the volume of big data has increased explosively, the types of data have been also diversified such that people's behaviors, as well as their thoughts and opinions can be anticipated through positional information and SNS services. Many countries and companies are attempting to construct and utilize the big data system now.

Accordingly, the market for the big data is becoming larger over time and the data is being used in different areas of our daily lives and much information is shared by the general population. However, since the analysis of big data is very complicated and difficult that sometimes it is quite hard to recognize its meaning and direction, the visualization of big data has come into the picture. Recently, the big data analysis is shifting from AMOS to R [9-18].

3 Scope of Research

The big data used in this study is the atmospheric data measured by the Seoul Metropolitan Government. The data has been extracted from "Daily average atmospheric information in respective periods" published in "Seoul Open Data Plaza" and contains several variables (i.e., fine dust, ultra-fine dust, concentrations of ozone, nitrogen dioxide, carbon monoxide and sulfur dioxide) related to the atmospheric conditions in the one-year-long daily measurements.

According to the data, the units used for the ozone concentrations and both fine dusts and ultra-fine dusts are *ppm* and ㎍/㎥, respectively. This project was preceded with the data gathered tracing back a year from December, 2015 and as there were many districts, one site was chosen for data analysis.

4 Data Analysis

To perform the analysis against each data, several variables in a certain area will be the target of analysis as there are different data values for respective districts and dates.

	A	B	C	D	E	F	G
1	day	dust	hdust	ozone	no2	co	so2
2	20151221	48	32	0.002	0.042	0.7	0.005
3	20151220	47	28	0.004	0.044	0.7	0.004
4	20151219	56	32	0.004	0.045	0.7	0.005
5	20151218	41	21	0.005	0.04	0.6	0.005
6	20151217	19	9	0.017	0.019	0.3	0.004
7	20151216	26	12	0.024	0.017	0.3	0.005
8	20151215	53	32	0.012	0.034	0.6	0.005
9	20151214	46	28	0.002	0.044	0.6	0.005
10	20151213	39	23	0.006	0.047	0.6	0.005
11	20151212	31	15	0.013	0.038	0.5	0.005
12	20151211	32	20	0.014	0.033	0.6	0.005
13	20151210	46	27	0.003	0.045	0.7	0.005
14	20151209	67	40	0.003	0.058	0.8	0.006
15	20151208	69	36	0.004	0.066	1	0.006
16	20151207	50	27	0.004	0.048	0.7	0.004
17	20151206	36	19	0.012	0.038	0.5	0.005
18	20151205	34	16	0.015	0.035	0.5	0.005
19	20151204	40	15	0.013	0.03	0.4	0.004
20	20151203	19	11	0.016	0.02	0.4	0.004
21	20151202	52	29	0.008	0.038	0.7	0.005
22	20151201	52	32	0.004	0.045	0.7	0.005
23	20151130	41	28	0.01	0.037	0.6	0.005
24	20151129	39	29	0.004	0.034	0.5	0.004
25	20151128	42	27	0.008	0.039	0.5	0.005
26	20151127	37	25	0.019	0.027	0.4	0.004
27	20151126	33	23	0.024	0.019	0.4	0.004
28	20151125	12	8	0.014	0.026	0.4	0.004
29	20151124	19	11	0.009	0.031	0.4	0.004
30	20151123	3	2	0.018	0.02	0.3	0.004
31	20151122	30	19	0.01	0.033	0.6	0.005
32	20151121	46	31	0.012	0.04	0.7	0.005
33	20151120	35	22	0.009	0.035	0.6	0.005
34	20151119	14	7	0.009	0.029	0.3	0.004
35	20151118	9	6	0.016	0.025	0.3	0.004

Fig. 1 The daily atmospheric information of Gangnam District.

[Fig. 1] shows the daily atmospheric information of Gangnam District and its variables include fine dusts, ultra-fine dust, concentrations of ozone, nitrogen dioxide, carbon monoxide and sulfurous gas. The number of the samples (i.e., the number of days for the measurements) observed using a length function was 166, and analyses for each row variable were performed after reading the data using the R studio. The results are shown in [Fig. 2].

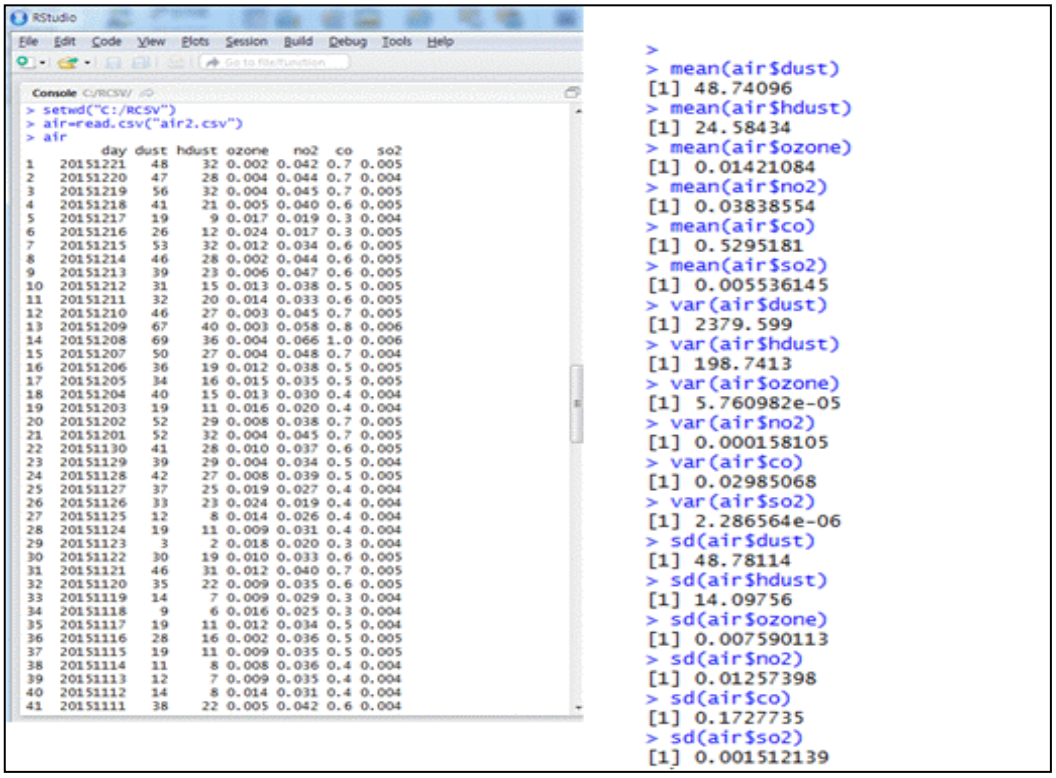


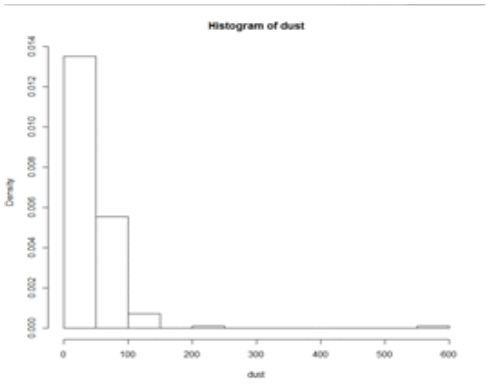
Fig 2. The analysis results using R studio against the variables

[Table. 1] Values of Averages, Variances and Standard Deviation for the variables

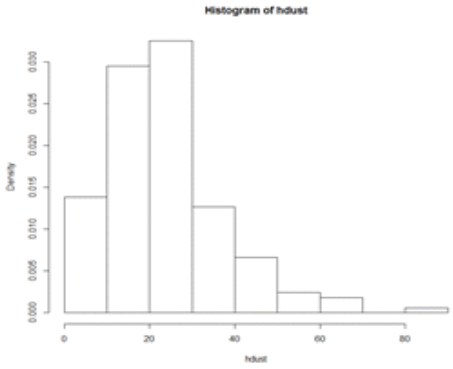
Variable	Average	Variance	Standard deviation
Fine dusts	48.74096	2379.599	48.78114
Ultra-fine dusts	24.58434	198.7413	14.09756
Ozone	0.01421084	5.760982e-05	0.007590113
NO2	0.03838554	0.000158105	0.01257398
CO	0.5295181	0.02985068	0.1727735
SO2	0.005536145	2.286564e-06	0.001512139

[Table. 1] shows the respective values of averages, variances and standard deviations for the variables involved. Each daily variable value can be checked

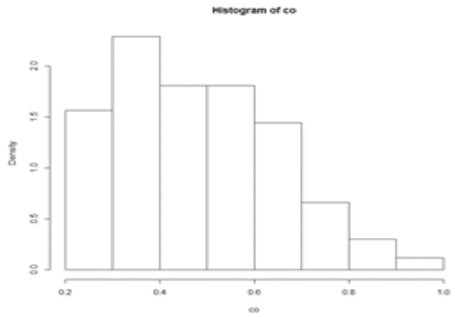
with the R studio and their averages, variances and standard deviations can be calculated with the mean, var and sd functions, respectively.



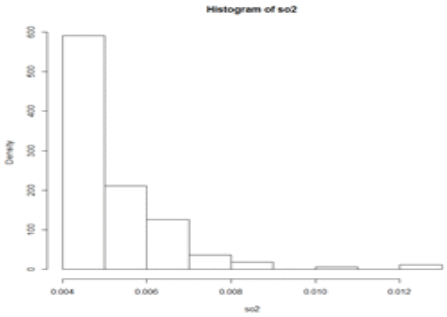
(Histogram of fine dust)



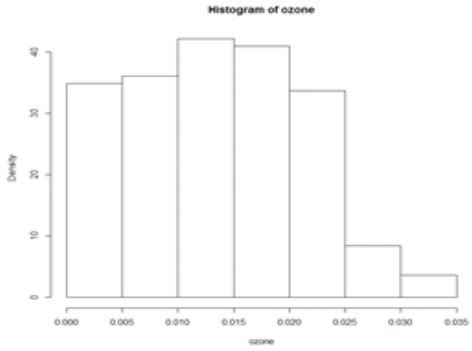
(Histogram of ultra-fine dust)



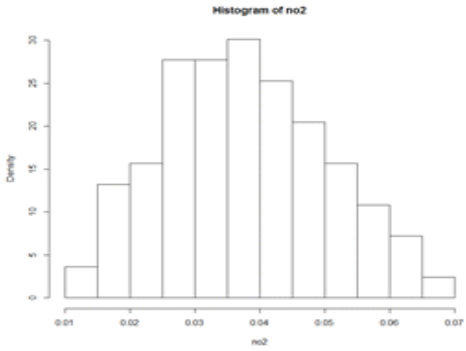
(Histogram of ozne)



(Histogram of NO_2)



(Histogram of Co_2)



(Histogram of SO_2)

Fig 3. The histograms of respective parameters

[Fig. 3] shows the histograms of each parameter, and in observing the histogram of fine dust, it is clear that the distribution is concentrated on the right

side. Along with the volume imbalance, it is distributed at the larger points of 200 and 500. As for the ultra-fine dust, its histogram reveals a form ascending to the right side and the data is evenly distributed when compared to the histogram of fine dust. The histogram of ozone shows quite evenly distributed data and as the peaks show little differences at the top, it is possible to grasp that there are many larger values. The histogram of NO_2 has a shape of bell and a symmetrical form. In CO's case is the data is biased to the left, as SO_2 , which has a higher peak instead.

By this time, it is necessary to analyze the data with a simple regression analysis to understand the correlation of each variable with ozone. The ozone concentrations will be considered as the dependent variables and others, as the independent variables. The first analysis is carried out for the fine dust.

```
>
> out=lm(ozone ~ dust, data=air)
>
> out

Call:
lm(formula = ozone ~ dust, data = air)

Coefficients:
(Intercept)         dust
  1.445e-02   -4.867e-06

> summary(out)

Call:
lm(formula = ozone ~ dust, data = air)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0123118 -0.0059937 -0.0003386  0.0059547  0.0190532

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.445e-02  8.362e-04  17.279   <2e-16 ***
dust        -4.867e-06  1.214e-05  -0.401    0.689
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.007609 on 164 degrees of freedom
Multiple R-squared:  0.0009784, Adjusted R-squared:  -0.005113
F-statistic: 0.1606 on 1 and 164 DF,  p-value: 0.6891

>
```

Fig 4. A simple regression analysis conducted against the fine dust using R studio.

[Fig. 4] shows the result of a simple regression analysis performed against the fine dusts using the R studio. Here, the ozone concentrations are the dependent variables and the fine dusts are the independent variables. Different from the common regression analyses which interpret the correlations between the dependent and independent variables, the simple regression analysis has only a single independent variable, the fine dusts. The same analysis was conducted for the rest of the variables.

5 Result of Modelling of Preliminary Analysis for Big Data

The simple regression analysis by the R studio is to grasp the linear relationship between a dependent (ozone) and an independent (fine dust) variables. It is important that how well the regression equation derived from the analysis result can explain the relationship between these two variables. [Fig. 5] shows the result of simple regression analysis for the fine dusts variable.

```
        Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.445e-02 8.362e-04 17.279 <2e-16 ***
dust -4.867e-06 1.214e-05 -0.401 0.689
F Value: 0.1606, p-value: 0.6891, R-Square: 0.0009784
```

Fig 5. The result of simple regression analysis for the fine dust variable.

The result is displayed in [Fig. 6].

```
Inclination:  -4.867e-06  p-value : 0.689
R²        :  0.0009784
R²(a)     : -0.005113
F(df1=1,df2=164) : 0.1606     p-value : 0.6891
```

Fig 6. The result of dust.

In order to identify the strength of the relationship, one can check it with an F-value which indicates the adequacy of the model in use.

The F-value in this case showed the level of 0.1606, which is significant. On the other hand, the value of the determination coefficient R^2, which explains how well the regression line can explain the actual data, is 0.0009784. This very low

value suggests that the ozone concentration can hardly be explained or determined by the fine dust. The results of the other same analyses to identify the impact are shown in [Fig. 7].

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.723e-02	1.159e-03	14.863	< 2e-16 ***
hdust	-1.228e-04	4.093e-05	-2.999	0.00313 **

F-value : 8.993, p-value: 0.003133, R-squared: 0.05198

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.029376	0.001439	20.41	<2e-16 ***
no2	-0.395077	0.035638	-11.09	<2e-16 ***

F-value : 122.9, p-value: < 2.2e-16, R-squared: 0.4284

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.028782	0.001489	19.33	<2e-16 ***
co	-0.027517	0.002674	-10.29	<2e-16 ***

F-value : 105.9, p-value: < 2.2e-16, R-squared: 0.3923

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.020708	0.002187	9.470	< 2e-16 ***
so2	-1.173562	0.381090	-3.079	0.00243 **

F-value : 9.483, p-value: 0.002432, R-squared: 0.05466

Fig 7. The regression analysis for the respective variables.

The F-value of the ultra-fine dust was 8.993, confirming that the model was adequate. But similar to the fine dust, its determination coefficient was calculated as 0.05, which was too small to be significant. As for NO_2 , an F-value was 122.9 and the determination coefficient was 0.4284. A significant adequacy and the best coefficient among the variables. CO had the F-value of 105.9 so that the adequacy of its model was high enough but the determination coefficient was rather low showing a value of 0.3923. Finally, for SO_2 , the F-value was 9.483 and the determination coefficient value was 0.05466, which in this case showed significant adequacy level and poor explanatory power, respectively.

[Fig. 8] shows an indication of correlations between ozone and other variables using the plotting method. [Fig. 9] shows R Studio plot. [Fig. 10] shows regression analyses of independent variables against the ozone concentration (dependent variable).

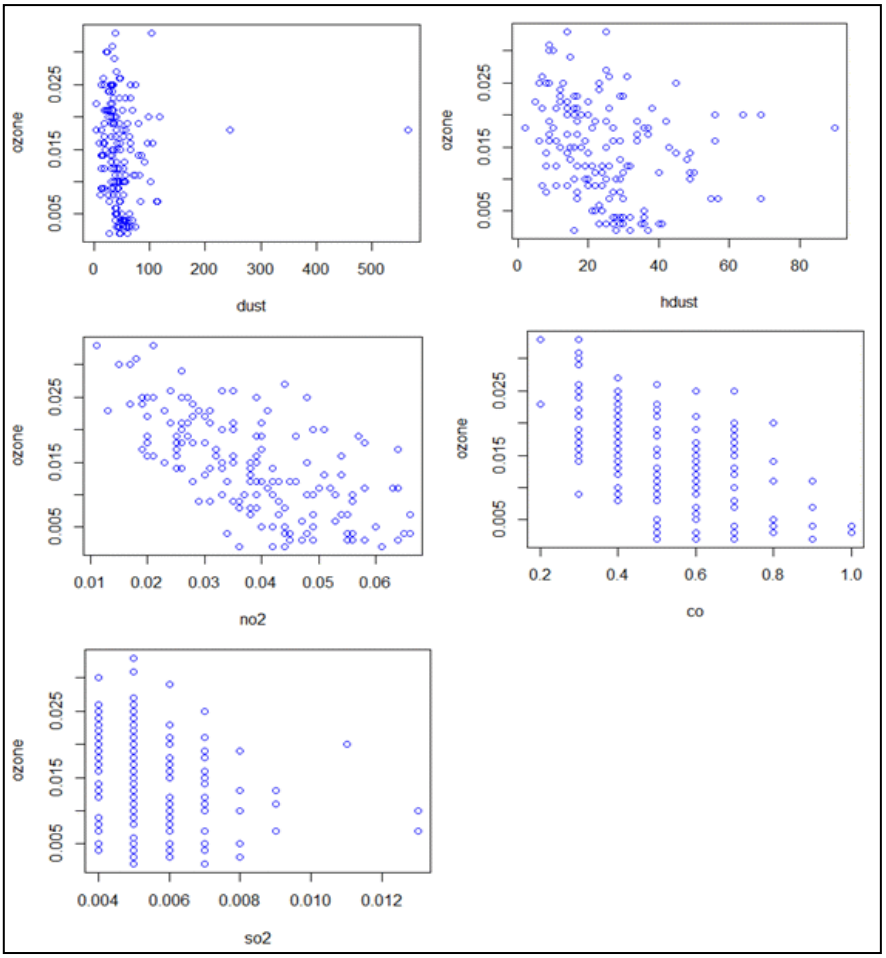


Fig 8. An indication of correlations between ozone and other variables using the plotting method.

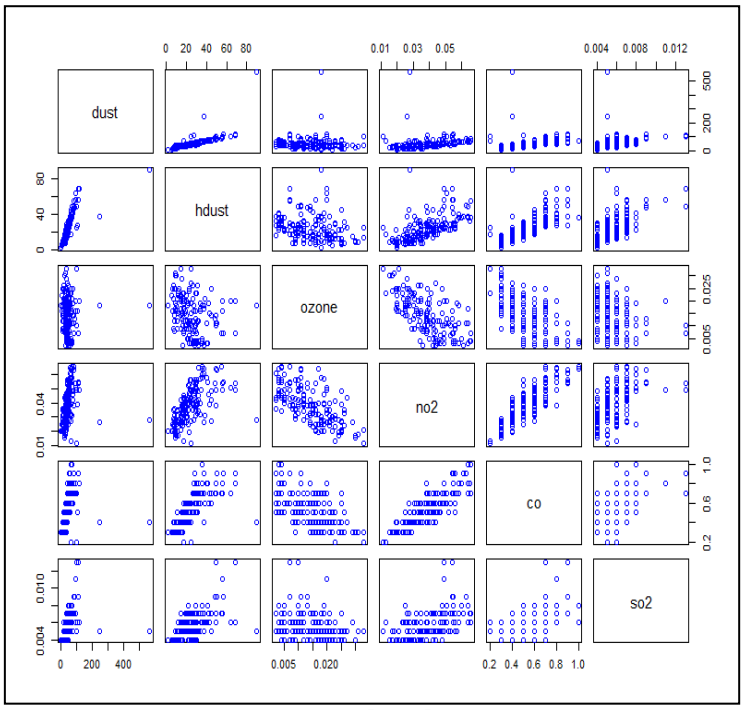


Fig 9. R Studio plot

```
>
> out = lm(ozone ~ ., data=air)
> summary(out)

Call:
lm(formula = ozone ~ ., data = air)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0143057 -0.0033864 -0.0000208  0.0030951  0.0122863

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.261e-02  1.770e-03  18.423  < 2e-16 ***
dust        -5.399e-05  1.371e-05  -3.937 0.000123 ***
hdust        4.298e-04  6.827e-05   6.297 2.81e-09 ***
no2         -2.706e-01  5.978e-02  -4.526 1.17e-05 ***
co          -3.357e-02  5.328e-03  -6.301 2.75e-09 ***
so2          3.298e-01  3.371e-01   0.979 0.329274
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.004962 on 160 degrees of freedom
Multiple R-squared:  0.5855,    Adjusted R-squared:  0.5725
F-statistic:  45.2 on 5 and 160 DF,  p-value: < 2.2e-16

>
```

Fig 10. Regression analyses of independent variables against the ozone concentration (dependent variable).

As shown in [Fig. 11], the regression analyses of several variables were conducted for the dependent variable (ozone concentration) by using the lm function. The adequacy of the model was significant and the regression coefficient was also statistically significant.

Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.261e-02	1.770e-03	18.423	< 2e-16 ***
dust	-5.399e-05	1.371e-05	-3.937	0.000123 ***
hdust	4.298e-04	6.827e-05	6.297	2.81e-09 ***
no2	-2.706e-01	5.978e-02	-4.526	1.17e-05 ***
co	-3.357e-02	5.328e-03	-6.301	2.75e-09 ***
so2	3.298e-01	3.371e-01	0.979	0.329274

F-value : 45.2, p-value: < 2.2e-16, R-squared: 0.5855

Fig 11. Execution of multiple regression analysis.

In this analysis, the F-value was 45.2 and the determination coefficient (R^2) was 0.5855. This higher value compared to the value obtained from the single regression shows that the explanatory power of the model is better.

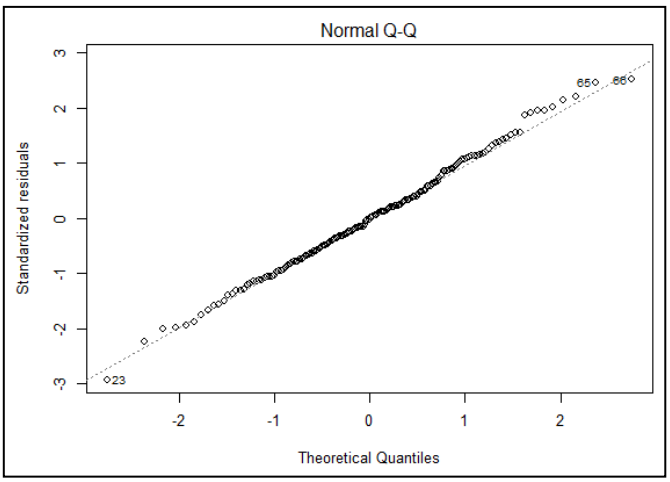


Fig 12. Normal Q-Q plot for the residuals using R-studio.

Also, after representing the residuals as shown in [Fig. 12], it was possible to determine that there was no problem in the error terms since the diagram of residuals had shown an evenly distributed residuals.

Among the five other variables analyzed against the ozone concentration, the variable which has had the best explanatory power (i.e., having the largest impact) was NO_2 , followed by in the order of CO, SO_2 , ultra-fine dust and fine dust. The residuals diagram which indicates the difference (error) between estimated value of the dependent variable from the model and the actual value observed from the dependent variable was produced. Then, it was possible to acquire the information about some uncertainties which cannot be explained with a statistical model, as well as the precision levels of the estimations calculated with the regression coefficients.

6 Conclusion and Future Work

More and more fish and agricultural farms are being established near cities and in suburbs as the farming technologies are developing faster every year and it is expected that they will appear in the central areas of cities eventually. In most of the cases, the problems of clean water supply and other needs are being taken cared of or expected to be cleared in the near future but the problems concerning the atmospheric conditions remain as uncontrollable factors, especially when they are caused by the neighboring countries. Although the ROK is regarded as an industrialized country and China being a developing country, the amount of pollutants produced by both countries exceed the global standards harming people's health and giving much negative impact on the food production industry as a substantial amount of these pollutants are accumulated in the produce and marine products. The described research in this paper adopts the method of visualizing the environmental data that includes several pollutants. The sample data provided by the Seoul Metropolitan Government was used to obtain analytical results through statistical analyses and visualized for easier viewing. The analyses continued over a period of one year and the causalities between the pollutants and ozone concentration was studied as well. All the correlations have been indicated through visualization, histograms and plotting diagrams, all of which showed a respective significance level.

The analysis methodology in this paper also included the univariate, simple regression and multiple regression analyses for visualization, the log transformation to establish an analytic model, and categorization of the explanatory variables.

The authors hope that this study will contribute in establishing an effective farm site selection method and our future tasks involve further analyses for other meteorological elements and the atmospheric conditions in other areas.

Acknowledgements. The first draft of this paper [1] was presented Oral Session in The 9th 2016 International Interdisciplinary Workshop Series, JEJU, April, 19-21 (2016). I am grateful to 3 anonymous commentators who have contributed to the enhancement of the paper's completeness with their valuable suggestions at the Workshop.

This work was supported by a Research Grant of Pukyong National University (2016 year).

References

[1] Jun-Ho Huh, Han-Byul Kim, Kyungryong Seo, Preliminary Analysis Model of Big Data for Prevention of Bioaccumulation of Heavy Metal-Based Pollutants: Focusing on the Atmospheric Data Analyses, *Advanced Science and Technology Letters*, **129** (2016), 159-164. http://dx.doi.org/10.14257/astl.2016.129.32

[2] Jun-Ho Huh, Sugarbayar Otgonchimeg, Kyungryong Seo, Advanced metering infrastructure design and test bed experiment using intelligent agents: focusing on the PLC network base technology for Smart Grid system, *The Journal of Supercomputing*, **72** (2016), no. 5, 1862-1877. http://dx.doi.org/10.1007/s11227-016-1672-4

[3] D. Battré, M. Hovestadt, B. Lohrmann, A. Stanik, D. Warneke, Detecting bottlenecks in parallel dag-based data flow programs, In: *2010 3rd Workshop on Many-Task Computing on Grids and Supercomputers*, (2010). http://dx.doi.org/10.1109/mtags.2010.5699429

[4] A. Behm, V.R. Borkar, M.J. Carey, R. Grover, C. Li, N. Onose, R. Vernica, A. Deutsch, Y. Papakonstantinou, V.J. Tsotras, Asterix: towards a scalable, semistructured data platform for evolving-world models, *Distrib. Parallel Databases*, **29** (2011), no. 3, 185-216. http://dx.doi.org/10.1007/s10619-011-7082-y

[5] K.S. Beyer, V. Ercegovac, R. Gemulla, A. Balmin, M.Y. Eltabakh, C.C. Kanne, F. Özcan, E.J. Shekita, Jaql: a scripting language for large scale semistructured data analysis, *Proceedings of the VLDB Endowment*, **4** (2011), no. 12, 1272-1283.

[6] C. Boden, M. Karnstedt, M. Fernandez, V. Markl, Large-scale social media analytics on stratosphere, *Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion*, (2013). http://dx.doi.org/10.1145/2487788.2487916

[7] V.R. Borkar, M.J. Carey, R. Grover, N. Onose, R. Vernica, Hyracks: a flexible and extensible foundation for data-intensive computing, *2011 IEEE 27th International Conference on Data Engineering*, (2011), 1151-1162. http://dx.doi.org/10.1109/icde.2011.5767921

[8] N. Bruno, S. Agarwal, S. Kandula, B. Shi, M.C. Wu, J. Zhou, Recurring job optimization in scope, *Proceedings of the 2012 international conference on Management of Data - SIGMOD '12*, (2012), 805-806. http://dx.doi.org/10.1145/2213836.2213959

[9] M. Cha, H. Haddadi, F. Benevenuto, P.K. Gummadi, Measuring user influence in twitter: the million follower fallacy, *Fourth International AAAI Conference on Weblogs and Social Media*, (2010).

[10] H. Chafi, Z. DeVito, A. Moors, T. Rompf, A.K. Sujeeth, P. Hanrahan, M. Odersky, K. Olukotun, Language virtualization for heterogeneous parallel computing, *OOPSLA '10 Proceedings of the ACM International Conference on Object Oriented Programming Systems Languages and Applications*, (2010), 835-847. http://dx.doi.org/10.1145/1932682.1869527

[11] B. Chattopadhyay, L. Lin, W. Liu, S. Mittal, P. Aragonda, V. Lychagina, Y. Kwon, M. Wong, Tenzing a sql implementation on the mapreduce framework, *Proceedings of VLDB*, **4** (2011), no. 12, 1318–1327.

[12] Jun-Ho Huh, Taehoon Koh, Kyungryong Seo, NMEA2000 Ship Area Network Design and Test Bed Experiment using Power Line Communication with the 3-Phase 3-Line Delta Connection Method, *International Journal of Applied Engineering Research, Research India Publications*, **10** (2015), no. 11, 27789-27797.

[13] S. Chaudhuri, K. Shim, Including group-by in query optimization, *Proceeding of the 20th VLDB Conference*, (1994), 354-366.

[14] J. Cohen, Graph twiddling in a mapreduce world, *Computer. Sci. and Engineering*, **11** (2009), no. 4, 29-41. http://dx.doi.org/10.1109/mcse.2009.120

[15] Jun-Ho Huh, Kyungryong Seo, Development of Competency-oriented Social Multimedia Computer Network Curriculum, *Journal of Multimedia and Information System*, **1** (2014), no. 2, 133-142.

[16] Jun-Ho Huh, Taehoon Koh, Kyungryong Seo, NMEA2000 Ship Area Network (SAN) design and Test Bed using Power Line Communication (PLC)

with the 3-Phase 3-Line Delta Connection Method, *Advanced Science and Technology Letters*, **94** (2015), 57-63.

http://dx.doi.org/10.14257/astl.2015.94.13

[17] Jun-Ho Huh, Kyungryong Seo, Hybrid AMI Design for Smart Grid Using the Game Theory Model, *Advanced Science and Technology Letters*, **108** (2015), 86-92. http://dx.doi.org/10.14257/astl.2015.108.19

[18] Jun-Ho Huh, Namjug Kim, Kyungryong Seo, Design and Implementation of Mobile Push Service-Based Mobile Medication-Hour Notification System, *Advanced Science and Technology Letters*, **117** (2015), 92-96. http://dx.doi.org/10.14257/astl.2015.117.22

Received: June 15, 2016; Published: October 14, 2016