Modelling of a Base Big Data Analysis Using R Method for Selection of Suitable Vertical Farm Sites: Focusing on the Analysis of Pollutants

Jun-Ho Huh*, Kyungryong Seo**

ABSTRACT

The problem of food deficiency is a major discouragement to many low-income developing countries. Most of these countries experience constant danger of hunger, malnutrition and diseases as they are unable to maintain their food supplies mainly due to lack of arable lands and modern crop, livestock and fishery production technologies. In addition, the pollutants resulting from the secondary industries are becoming another serious issue in their food problems. The pollutants mixed in the sands blowing from the mainland China and the toxic waters flowing in the farm land form the industrialized zones are some of the examples. The Vertical Farm, or Plant Factory, proposed in this study could be the best alternative food production system for them. Vertical farm is an efficient food production system that yields relatively a large volume of food materials without environmental risks. The system does not require a large open space and manpower and can minimize the possibility of infiltration of pollutants. This research describes a basic model of the system focusing on determining the optimal sites for it based on the meteorological data concentrating on the atmospheric pollutants. The types and volume of pollutants are analyzed and identified through the big data obtained, followed by visualization of analysis results and their comparisons for better understanding.

Key words: Modelling, R, Big Data, Vertical Farm, Vertical Farm Sites.

1. INTRODUCTION

Despite of international efforts to solve the problem of overpopulation, the population in the developing countries who are experiencing a serious food shortage is increasing faster. What makes worse is that they do not have adequate skills or lands to maintain enough food for their people. Aside from these problems, consuming inedible or contaminated foodstuffs is causing a grave problem for them as well. The Vertical Farm was proposed in 1999 as an alternative efficient food production system which can solve such problems and increase the production volume. Since then many researches have been carried out and recently, the government of Republic of Korea (ROK) is supporting the construction of pilot vertical farms in the country. The Ministry of Maritime Affairs and Fisheries is leading the project along with some private corporations with the hope of building actual vertical farms in near future. Basically, the indoor crop-growing vertical farms commonly require water-recycling equipments, application-specific LED lights and nutrient solutions. But recent rapid increase in the number of health-conscious consumers demands the farm produce to be contaminant-free and healthy. The demand is high as Koreans are living next to the mainland China.
from where most of pollutants are flowing in from through both air and sea and result in bio-accumulation, especially heavy-metals. This is common in every other country where industries are growing fast and the pollution prevention measures cannot catch up with the speed of their growth. Although it may be possible to reduce in-flow of contaminants by constructing a closed vertical farm, it will not be able to keep pollutants or contaminants out completely so that selection of an optimal site for a vertical farm is essential in lowering the possibility of contaminations. On that account, a basic analysis model was designed by performing a big data analysis against major contaminants found in the Korean meteorological data. R-language was used for the construction of the model.

The results obtained from the univariate, simple regression, and multi regression analyses have been visualized for easier viewing and a model was established using the log transformation and other means. A base research was performed for the selection of suitable sites for the vertical farms by modelling the involved characteristics based on the categorized explanatory variables.

2. RELATED STUDIES

2.1. Vertical Farm

The Professor of Columbia university, Dick Despommier conceived an idea of Vertical Farm system in 1999 [3] while he was trying to find the ways to solve the problem of food shortages mainly due to lack of arable lands. Also being called as the Plant Farm, this system produces relatively a larger volume of crops and plants in a limited space by controlling temperature, humidity, water level and lights. The weather conditions do no influence its output much. Commercialization of the system has progressed much after going through various test bed experiments in US, EU, Japan and other developing countries [1-5]. Even the livestock industry is studying the system in cooperation with some pharmaceutical companies [6-9]. An example of vertical farm is shown in Fig. 1 such a system is expected to harvest crops around thirty times a year and depending on the system configuration. Since its efficiencies have been proven by the test beds, this study focuses on the methodology of selecting an optimal site(s) for the vertical fish farms.

This "Future Farming" in the time of abnormal climate changes uses LED lights instead of sun-
light and sprinkles nutrient–mixed sprays on the roots instead of soaking them in the water. According to the simulations, it is possible to grow foodstuffs that are enough to feed fifty thousand people with the thirty-story farm.

This system allows farmers to grow their products regardless of weather conditions by controlling lights, water, temperature and humidity. It’s often referred to as a “Plant Farm”. For the agricultural sector, the test bed experiments were completed in Japan, US and Europe, where some of their organizations have entered the business stages whereas the livestock sector is actively studying the system with some pharmaceutical companies. Fig. 2 shows sample of Vertical Farm.

The vertical farm, which has been proposed by Professor Dickson Despommier, refers to a method of growing plants, crops or fishes in a building structure (Vertical Farm) [3–7]. This method is being studied as a national project and currently experimented in some test beds in the Republic of Korea (ROK). This research is focusing on a selection method of optimal farm sites for these farms, especially for the aquafarms.

2.2. Reference of Big Data

The reference data used in the study contained the atmospheric data collected throughout a year on an hourly basis. The measured elements were dry–bulb temperatures, atmospheric pressures, cloudiness, snowfalls, rainfalls, wind directions, relative humidity, horizontal solar irradiation, wind speeds, dew–point temperatures, absolute humidity, air densities, and enthalpy.

Using the R-studio, the analysis for atmospheric data of Seoul in 1981 was carried out first, followed by the same processes for other cities. Fig. 3 shows Variables of Seoul’s atmospheric environment data measured by the hour. Selected variables for analyses includes dry–bulb temperature(℃), atmospheric pressure (hPa), cloudiness (1–10), snowfall (cm), rainfall (cm), relative humidity (%), horizontal solar irradiation (W/m²), wind speed (m/s), dew–point temperature (℃), absolute humidity (Kg/Kg), air density (Kg/m³), and enthalpy (KJ/Kg) and these were designated as DryTemp, hPa, CloudAmount, Snow Amount, RainAmount, RelativeHumidity, Horizontal Solar, WindSpeed, DewpointTemp, Absolute Humidity, AirDensity, Enthalpy, respectively. The data was extracted from the 30-year-long (1981–2010) atmospheric environment data of Seoul, Daejeon, Daegu, Busan, Gwangju, Incheon and Ulsan, the cities of the ROK.

3. BASIC BIG DATA ANALYSIS METHOD FOR SELECTION OF SUITABLE VERTICAL FARM SITES

Although the waters of Jeju Island are known as clean areas, the level of heavy–metal contamination is becoming rather serious as they are close to the mainland China. However, it is difficult for the researchers or farmers to exactly point out the causes of contamination as it’s being argued that the pollutants are usually accumulated through the mechanism that has several parameters, or in–flow of heavy–metal–mixed dusts. Thus, we are
not dealing with the heavy-metal accumulation in this study. Unless future vertical farms are not equipped with adequate atmospheric filters, it would be logical to make the best selections for the farm sites by using an atmospheric analysis model to avoid contaminated areas.

3.1. Analysis Using R

Daily variable values can be checked with the R-studio and the average, variance and the average, variance and standard deviation of each variable can be calculated using the mean, var and sd functions. Fig. 4 shows means, variances, and standard deviations of variables of Seoul measured using the R-studio. Observing the histogram of DryTemp, it has a leftward long tail and most of the data is highly concentrated in the temperatures between 20°C and 30°C. For hPa, the data is evenly distributed having a large peak in the middle, where the data is concentrated between 1000 hurdles and 1010hPa. The data in the histogram of the CloudAmount is weighted towards both ends. For SnowAmount and RainAmount, the distribution in respective histograms are highly unbalanced and data is mostly concentrated between 0cm and 1cm. RelativeHumidity a slightly long tail extended towards the left but on the whole, the data is evenly distributed. The histograms of HorizontalSolar,
WindSpeed, AbsoluteHumidity and AirDensity have a similar form in which the long tail is extended toward the right. Their data is mostly distributed between low values. For DewpointTemp, it has a slightly long tail to the left and the data is weighted towards the sections where the temperatures are higher. Finally, the Enthalpy, its form does not assume the form of bell curve but is symmetric as a whole. The same data analysis was performed for other cities as well.

Fig. 5 shows DryTemp of analysis. All the data analysis results for other cities were similar to that of Seoul. There was a little difference depending on the value sizes but they had a histogram form where their long tails were mostly extended to the left. In case of the Ulsan city, the data was distributed more evenly than the others. In general, the data of atmospheric pressures was evenly distributed and most of the cities had a symmetric form. Also, the data was mainly concentrated between 1,000hPa and 1,010hPa. Incheon and Ulsan showed rather higher values compared to the others. Fig. 6 shows hPa of analysis, and Fig. 7 shows AirDensity of analysis. The histograms showed a form that has a long tail extended to the right and the data was mainly concentrated at the sections with low values. There were some differences between the cities. Busan exhibited a bell-type histogram with a large peak. The data of Daegu and Gwangju was concentrated at the sections with low values while Ulsan's data was most evenly distributed, showing a bell-type histogram.

To understand the effects of other variables (i.e., cloudiness, snowfall, rainfall, relative humidity and wind speed, etc.), the graphic representations have been created. The graphs allows easier and convenient assessment of relationships between DryTemp and the other variable.

Fig. 8 shows correlations between Seoul's DryTemp and the other variables determined with R-studio plotting. It was conveniently understood that the variables that have the significant values to explain the DryTemp were the atmospheric pressure, DewpointTemp, AbsoluteHumidity, Air-
density and Enthalpy. Especially, the graph indicated that the AirDensity was inversely proportional to the DryTemp quite exactly.

3.2. Multiple Regression Analysis

The multiple regression analysis is to analyze the data to understand the correlations between each variable. Therefore, only the One:One linear
relationship can be determined. However, since the meteorological information contains many variables, the multiple regression analysis is needed to determine the correlations. For the analysis, the prediction equation that has various variables will be used.

Fig. 9 shows all the variable pairs of Seoul represented by the `pairs()` function of R-studio. The multiple regression analyses were performed on the dependent variable (temperature) using the `lm` function. As a result, the model was suitable and the regression coefficient was statistically sig-
significant. The value of $F$ was $1.951e+07$ and the determination coefficient $R^2$ was 1. The value of $R^2$ was higher than the one obtained from the simple regression analysis, which means that the explanatory power against the model is higher. That is, the temperature can be also determined (explained) with other variables.

3.3. Residual

The residual graph that shows the difference between the value of dependent variable estimated by the model and the value obtained from an actual observation has been created. The residuals are the differences in these two dependent variables. With the residual (error), one can understand the uncertainty of information that cannot be explained with the statistical models.

Fig. 10 shows the Q–Q normality plotting implemented with the R–studio. Moreover, such difference can also be interpreted as an error or an uncertainty information that cannot be adequately explained with the statistical models. Since the residuals are the random variables having the random effects, they are useful in testing the appropriateness of distribution assumed by the statistical model. Observing that this residual plot is evenly shaped, one can determined that there are no problems in error terms as well. We were able to con-
clude that the dependent variable (temperature) can be determined (explained) through the multiple regression analyses of other variables.

4. CONCLUSION AND FUTURE WORKS

The meteorological data has been collected for a period of a year for analysis purpose. The results from the analyses were used to establish an R-language-based analytical model to locate the most suitable site for the vertical farm, possible in or near the cities. Such analysis results naturally showed some differences depending on the geographical factors of individual candidate cities. The data obtained through simple/multiple regression and univariate analyses was visualized for easier understanding and comparison of the results by the researchers. This was quite effective in determining some of the data characteristics which would have been unnoticeable. Log-transformation was used also to construct the analytical model, along with a basic research involving categorization of the explanatory variables while modelling the typical characteristics of the data.

This study will be the basis of future extended study which will include the big data collected through a period of more than 30 years. We expect that the analytical model created based on such a data will be quite accurate when finding the most suitable locations for the vertical farms, as well as for the conventional indoor aquafarms.

APPENDIX

Since Pukyong National University is continuing the researches on raising high-value fishes, we’ve proceeded with the research on a method of finding the optimal farming sites for the farmers. In aquaculture, selecting the optimal locations means finding the farm sites that have the most suitable environment, economic feasibility, and the least harmful elements. As mentioned earlier, the ROK government has much interest in the vertical farm system and constructing the test beds for the research purpose. It is our hope that our proposal will contribute to such projects to provide healthier foods to the public.

The first draft of this paper [1] was presented Oral Session in Spring Conference of Korea Multimedia Society, May 27–28 (2016). I am grateful to 2 anonymous commentators who have contributed to the enhancement of the paper’s completeness with their valuable suggestions at the Conference. I've had a great opportunity to study some introductory courses such as Aquaculture Facility and Environment, and aquaculture technology under an honorary professor of Aquaculture Technology at Jeju University and a former head of the Korea Aquaculture Society, Prof. S. Roh my motive in writing this paper concerning the selection of the optimal farm sites came from the experience of a field trip to Jongdalri, which had taught me that the fishes cannot be and should not be raised with the waters or in the airs that we cannot drink or breathe ourselves.

REFERENCE


Jun-Ho Huh

finished the Cooperative Marine Science and Engineering Program, Texas A&M University at Galveston, United States of America in Aug. 2006.

Received B.S. in Science Degree from Department of Major of Applied Marine Sciences (Currently Faculty of Marine Biomedical Sciences). B.S. in Engineering Degree (Double Major) from Department of Major of Computer Engineering from Jeju National University at Ara, Jeju, Republic of Korea in Aug. 2007. And completion of the Secondary School (Middle and High schools) Teacher Training Curriculum in accordance with Republic of Korea Secondary Education Act.

M.A. in Education Degree from Department of Major of Computer Science Education, Graduate School of Education, Pukyong National University at Daeyeon, Busan, Republic of Korea in Aug. 2012. And completion of the Secondary School (Middle and High schools) Teacher Training Curriculum in accordance with Republic of Korea Secondary Education Act.

Received the Ph.D. in Engineering Degree from Department of Major of Computer Engineering, Graduate School, Pukyong National University at Daeyeon, Busan, Republic of Korea in Feb. 2016.

He received the Best Paper Award from Korea Multimedia Society five times (Nov. 2014, May 2015, Nov. 2015, May 2016, Oct.). And he received Best Paper Award The 10th 2016 International Interdisciplinary Workshop Series from HSST (Aug. 2016). Also he received Best Paper Award The 16th International Conference on Control, Automation and Systems (Oct. 2016), ICROS with IEEE Xplore.

Research Professor of Dankook University at Jukjeon, Yongin, Republic of Korea (July. 2016~ Sep. 2019). Currently he is Assistant Professor of Department of Software, Catholic University of Pusan, Busan, Republic of Korea (Dec. 2016~).

His research interests are Green IT, Smart Grid, Network Security, IoT, Curriculum of Computer.

Kyungryong Seo

received B.S. in Engineering Degree from Department of Major of Electrical Machinery Engineering from Pusan National University, Busan, Republic of Korea in Feb. 1983.

M.S. in Engineering Degree in Electrical Engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea in Feb. 1990.

Received the Ph.D. Engineering Degree in Electrical Engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea in Aug. 1995.

He received the Best Paper Award from Korea Multimedia Society six times (Nov. 2007, Nov. 2014, May. 2015, Nov. 2015, May 2016). And he received Best Paper Award The 10th 2016 International Interdisciplinary Workshop Series from HSST (Aug. 2016).

Currently he is a Full (Tenure) Professor of Computer Engineering Department, Pukyong National University at Daeyeon, Busan, Republic of Korea.

His research interests are High Speed Computer Network, Network Security, High Availability Computing.
