공학석사학위논문

A study on development of models estimating the amount of flood waste

수해폐기물 발생량 추정모형 개발에 관한 연구

2020년 2월

서울대학교 대학원

건설환경공학부

박 만 호

A study on development of models estimating the amount of flood waste

by

Man Ho Park

Advisor: Jae Young Kim

A dissertation submitted in partial fulfillment

of the requirement for the degree of

Master of Science

Department of Civil and Environmental Engienering

The Graduate School

SEOUL NATIONAL UNIVERSITY

Feburary 2020

Abstract

A study on development of models estimating the amount of flood waste

Man Ho Park
Civil and Environmental Engineering
The Graduate School
Seoul National University

Flood waste management is important for reducing the damage and secondary environmental pollution caused by delays in disaster recovery. One key issue related to flood waste management concerns estimating the precise quantity of waste to plan recovery strategies and policy. Multilinear regression has been recognized as a viable technique to estimate the amount of waste generation from flood. There are two types of flood waste estimation methods: pre-event predictions using factors related to regional properties and rainfall hazards, and post-event predictions using damage variables due to floods, such

as the number of damaged buildings. However, adapting the framework suggested in other countries did not work in Korean cases. In this study, an advanced flood waste estimation technique was devised using data grouping, Bayesian linear regression, incorporating interaction terms and nonlinear regression using deep-neural network. The aim of this study was 1) to develop the flood waste regression model using aforementioned four frameworks; 2) to evaluate the performance and limitation of each method; and 3) to suggest the most viable strategy by comparison of the empirical validation results of the models established upon the frameworks. Totally 90 flood cases (2008-2017) in South Korea collected from *annual report on disaster* published by National Emergency Management Agency in Korea were the scope of this study.

The data grouping was performed in post-event flood waste estimation following three grouping characteristics: administrative region (AR; equivalent to special city or province), urbanization rate (UR), and disaster type and offshore accessibility (DO). Such data grouping led to flood waste prediction improvement not only by the single stage grouping but also by successive groupings. Data grouping was effective both for identifying groups with similar contexts and for eliminating disparities in the dataset that impede accurate waste prediction. Among the grouping sequences tested, the grouping order resulted in the most improvement in flood waste prediction was UR, AR, and

DO. This grouping order yielded enhanced waste prediction in 74 cases. However, grouping cannot explain every case because 16 cases were not fitted even after three-phase grouping. In addition, it is not clear whether the enhanced model fitness of some groups is associated with the screening effect of grouping or just a coincidence owing to the limitation of statistical analysis.

Conventional attempts to establish flood waste models used deterministic approaches; however, probabilistic methods have never been applied. Considering the large degrees of uncertainty in waste generation from floods, a probabilistic approach can provide a more accurate model compared to models developed by the conventional deterministic approach. One part of this study applied Bayesian inference to develop a flood waste regression model in South Korea. The aims of the study are as follows: (1) to analyze the characteristics of coefficients estimated by the Bayesian approach; (2) evaluate the performance of the prediction model by Bayesian inference; and (3) assess the effectiveness of Bayesian updating in a flood waste estimation. According to the results, the coefficients obtained via Bayesian inference showed a more significant *p*-value compared to those developed through the deterministic approach. Bayesian inference with a null *prior* distribution slightly reduced error compared to that by deterministic regression, specifically for post-event prediction. Bayesian updating did not effectively increase the accuracy of the

model, while iterative updating required a complex calculation process. These results reveal the potential of the Bayesian approach in flood waste estimations, but also showed the error reduction by Bayesian approach is, in fact, limited.

Incorporating the interaction terms comprised of the products of two independent variables resulted in improved modeling with lower root mean square error and higher adjusted r^2. It seems interaction terms compensated over/underestimated contributions of independent variables and also explained combined effect of two variables in waste generation. The observation throughout the field surveying after typhoon Danas in 2019 revealed that damage in single aspect, such as flooded cropland did not always generate waste, and damage in one variables is sometimes linked to other variables, for example, one completely destructed building can physically affect the nearest building and lead to partial damage to the building. Incorporating interaction terms for flood waste modeling is simple approach not needing costly works and significantly enhances the model performance.

For the final task, the use of deep-neural network to estimate the flood waste generation was tried. Totally 22 kinds of parameters were utilized as input variables which were classified by flood damage variables, factors related to regional characteristics and meteorological parameters. Rectified Linear Unit (ReLU) was applied as an activation function in each node and the adjusted

adaptive scaled gradient descend method was utilized as an optimizer. In order to optimize the performance, hyperparameter tunings were carried out in two phases: 1) grid search for discrete type hyperparameters and 2) Bayesian optimization of continuous type hyperparameters. As a results, the optimized hyper parameter set was (No. layer = 2, No. node = 83, Regularization number 1 = $10^{-1.798}$, Regularization number 2 = $10^{-4.113}$, Initial learning rate = $10^{-4.696}$) and the testing set loss was 6,738,730 equivalent to 2,595.91 of RMSE. The RMSE in testing set by optimized deep-neural network was the most small among the tested four framework. Considering the empirical validation, the use of deep-neural network is the best viable option to measure flood waste generation. This may be related to the nature of flood waste generation; flood waste generation is actually not linear but has threshold in some extent which can be modeled by deep-neural network with ReLU activation function.

Keywords: disaster waste management, waste amount estimation, flood waste, regression analysis, statistical model

Student Number: 2018-26321

TABLE OF CONTENTS

ABSTRACT i

TABLE OF CONTENTS vii

LIST OF FIGURES xii

LIST OF TABLES xvii

CHAPTER 1

INTRODUCTION 1

1.1 Background 1

1.2 Research objectives 3

1.3 Dissertation structure 4

References 7

CHAPTER 2

ADVANCED LINEAR REGRESSION IN FLOOD WASTE ESTIMATION BY CONTEXT-BASED DATA GROUPING

2.1 Introduction 9

2.2 Materials and methods 13

2.2.1 Data collection 13

2.2.2 Multilinear regression 14

2.2.3 Regression diagnosis 15

2.2.4 Sequential data grouping 16

2.3 Results and discussion 20

2.3.1 Multilinear regression without grouping 20

2.3.2 Post-event models following sequential data grouping 22

2.3.2.1 Post-event models from the first grouping 23

2.3.2.2 Post-event models from the second grouping 27

2.3.2.3 Post-event models from the third grouping 33

2.3.3 Summary of linear modeling with grouping 38

2.3.4 Regression diagnosis for grouped models 47

2.3.5 Discussion 50

2.3.6 Application procedure of suggested models 52

2.4 Summary 52

References 54

CHAPTER 3

BAYESIAN LINEAR REGRESSION IN ESTIMATING FLOOD

WASTE GENERATION

3.1 Introduction 57

3.2 Materials and methods 61

3.2.1 Data collection 61

3.2.2 Post event model 62

3.2.3 Bayesian regression protocol 63

3.2.4 Statistical examination 64

3.2.5 Application to the other context 65

3.3 Results and discussion 66

3.3.1 Pre-event estimation with Bayesian inference 67

3.3.2 Post-event estimation with Bayesian inference 74

3.3.3 Application of Bayesian framework in other context 82

3.4 Summary 86

References 88

CHAPTER 4

INCORPORATING INTERACTION TERMS IN LINEAR REGRESSION FOR FLOOD WASTE ESTIMATION

4.1 Introduction 91

4.2 Materials and methods 94

4.2.1 Data collection 94

4.2.2 Multivariate linear regression 95

4.2.3 Variable selection 96

4.2.4 Sensitivity analysis 97

4.2.5 Variance inflation factor calculation 98

4.2.6 Field surveying 99

4.3 Results and discussion 100

4.3.1 Linear regression without interaction terms 100

4.3.2 Linear regression with interaction terms 106

4.3.3 Roles and contributions of interaction terms 112

4.3.4 Analysis of the characteristics of flood waste generation 117

4.4 Summary 122

References 124

CHAPTER 5

OPTIMIZING DEEP-NEURAL NETWORK FOR PREDICTING WASTE GENERATION FROM FLOOD

5.1 Introduction 128

5.2 Materials and methods 131

5.2.1 Data acquisition 131

5.2.2 Structure of the neural network 132

5.2.3 Deep-learning environment 133

5.2.4 Hyperparameter tuning 133

5.3 Results and discussion 135

5.3.1 Optimal structure of DNN for flood waste estimation. 135

5.3.2 Optimal set of continuous type hyperparameters inside the DNN made up of optimal structure 138

5.4 Summary 141

References 142

CHAPTER 6 CONCLUSIONS 144

Abstract in Korean 147

LIST OF FIGURES

Figure 1.1. Structure of dissertation 6

Figure 2.1. A map of special cities and provinces in South Korea (Regions underlined with arrow bars are special cities; map by vemaps.com was slightly edited).....18

Figure 2.2. A plot of reported and estimated flood waste by the prediction model developed from 90 flood cases.....21

Figure 2.3. Flood waste estimation plots of the best-fit grouping sequence ..46

Figure 2.4. Residual plots of the regression analysis by the unclassified dataset and the best-fit grouping sequence (Solid line: 95% confidence interval calculated by *t*-distribution with relevant degree of freedom) 48

Figure 3.1. Plots of the pre-event flood waste prediction model: log-scale plots between actual and estimated waste generation via (a) deterministic regression, (b) Bayesian regression with a null *prior*, and (c) Bayesian regression with a *prior* comprised of values obtained from deterministic regression. Residual plots are shown for the estimated results of (d) deterministic regression, (e) Bayesian regression with a null *prior*, and (f) Bayesian regression with a *prior* comprised of

values obtained from deterministic regression 70

Figure 3.2. Changes in statistical indices of the pre-event flood waste prediction with Bayesian updating: (a) correlation coefficient, (b) root mean square error, (c) mean absolute error, (d) median absolute error, and (e) mean absolute percentage error 73

Figure 3.3. Plots of the post-event flood waste prediction models: plots between actual and estimated waste generation via (a) deterministic regression, (b) Bayesian regression with a null *prior*, and (c) Bayesian regression with a *prior* comprised of values obtained from deterministic regression. Residual plots are shown for the estimated results of (d) deterministic regression, (e) Bayesian regression with a null *prior*, and (f) Bayesian regression with a *prior* comprised of values obtained from deterministic regression 78

Figure 3.4. Changes in the statistical indices of post-event flood waste predictions with Bayesian updating: (a) correlation coefficient, (b) root mean square error, (c) mean absolute error, (d) median absolute error, and (e) mean absolute percentage error 81

Figure 4.1. Plots between actual and estimated waste generation by regression model developed by using 90 cases of flood waste generation in South Korea from 2008 to 2017: (a) model without interaction term,

and (b) model with interaction terms 103

Figure 4.2. Plots between actual and estimated waste generation by regression model developed without variable selection process using 90 cases of flood waste generation in South Korea from 2008 to 2017: (a) model without interaction term, and (b) model with interaction terms....104

Figure 4.3. Standardized residual plots of regression models developed by using 90 cases of flood waste generation in South Korea from 2008 to 2017: (a) model without interaction term, and (b) model with interaction terms.....105

Figure 4.4. Standardized residual plots of regression models developed without variable selection process by using 90 cases of flood waste generation in South Korea from 2008 to 2017: (a) model without interaction term, and (b) model with interaction terms 106

Figure 4.5. Response of waste estimation according to sensitivity analysis under the 99% confidential intervals of the number of completely destructed building and the number of flooded buildings: (a) model developed without interaction terms; (b) model developed with interaction terms (white solid lines: gradient curves of output responses; black solid line: a tangent line to one gradient curve) . 113

Figure 4.6. Estimated probability distributions of (a) the number of completely

destructed building and (b) the number of flooded buildings 114

Figure 4.7. Response of waste estimation according to sensitivity analysis in an interval ranged from -10σ to 10σ for the number of completely destructed building and the number of flooded buildings: (a) model developed without interaction terms; (b) model developed with interaction terms (white solid lines: gradient curves of output responses; black solid line: a tangent line to one gradient curve) .115

Figure 4.8. The graphical inspection diagram of calculating the curvature radius of response gradient in a sensitivity plot ranged in (-10σ, 10σ) of input variables (target contour line is assigned to 1,000,000 tons of waste output) 116

Figure 4.9. Pictures from the field surveying in flooded area (a): flooded cropland after the strike of the typhoon Danas in 16th-21st, July, 2019 (34°46'19"N, 127°35'17"E, Yeosu, Jeollanam-do, Republic of Korea) and (b): completely destructed building after the strike of the typhoon Danas in 16th-21st, July, 2019 (35°05'39"N, 129°02'53"E, Busan, Republic of Korea; some part was blocked to protect private information).....120

Figure 5.1. The elemental structure of the neural network used in this study 133

Figure 5.2. The results of ten times replicated grid search with root mean square as target function (ten times averaged) in (a) training set and (b) testing set.....136

Figure 5.3. The uncertainty of ten times replicated grid search with root mean square as target function (ten times averaged) in (a) training set and (b) testing set.....137

Figure 5.4. The results of detailed grid search in regions where the number of hidden layers = 2, and the number of nodes ranged from 78 to 92 138

LIST OF TABLES

Table 2.1. Summary of special cities and provinces in South Korea.....19

Table 2.2. Summary of flood waste modelling after first data grouping 22

Table 2.3. *P*-values of coefficients estimated in regression after first classification.....25

Table 2.4. Summary of flood waste modeling after second data grouping.....29

Table 2.5. *P*-values of coefficients estimated in regression after second classification.....31

Table 2.6. Summary of flood waste modeling after third data grouping 34

Table 2.7. *P*-values of coefficients estimated in regression after third classification.....36

Table 2.8. Summary of flood waste modelling with data grouping.....40

Table 2.9. Calculation sheet of grouping effectiveness index (GEI) 43

Table 2.10. Summary of Shapiro-Wilk test on the residuals according to the grouping 49

Table 3.1. Estimated coefficients of the pre-event flood waste prediction model developed by 90 cases of flood waste generation in South Korea from 2008 to 2017.....68

Table 3.2. Statistical indices of the pre-event flood waste prediction model

developed by 90 cases of flood waste generation in South Korea from 2008 to 2017 72

Table 3.3. Estimated coefficients of the post-event flood waste prediction model developed by 90 cases of flood waste generation in South Korea from 2008 to 2017 76

Table 3.4. Statistical indices of the post-event flood waste prediction model developed from 90 cases of flood waste generation in South Korea from 2008 to 2017 80

Table 3.5. Estimated coefficients of pre-event flood waste prediction model developed by 39 cases of flood waste generation in Taiwan, 2001. 84

Table 3.6. Statistical indices of pre-event flood waste prediction model developed by 39 cases of flood waste generation in Taiwan, 2001. 85

Table 4.1. Estimated coefficients of flood waste regression model without interaction term developed by using 90 cases of flood waste generation in South Korea from 2008 to 2017 101

Table 4.2. Estimated coefficients of flood waste regression model without interaction term developed by using 90 cases of flood waste generation in South Korea from 2008 to 2017 102

Table 4.3. Estimated coefficients of flood waste regression model considering interaction term developed by using 90 cases of flood waste

generation in South Korea from 2008 to 2017 107

Table 4.4. Estimated coefficients of flood waste regression model considering interaction term developed by using 90 cases of flood waste generation in South Korea from 2008 to 2017 109

Table 4.5. Variance inflation factors for parameters in flood waste regression model developed by using 90 cases of flood waste generation in South Korea from 2008 to 2017 118

Table 5.1. The results of Bayesian optimization with root mean square as target function in logarithm-transformed regularization rate in the first layer (Log_RNum1), that in the second layer (Log_RNum2), and initial learning rate (Log_INI_LR) 139

CHAPTER 1

INTRODUCTION

1.1 Background

Imbalance of atmospheric moisture and heat resulted in extreme rainfalls at Asian region annually. The hydrological discrepancy is problem that occur both in time and space dimensions. For an example of Seoul city, average amount of rainfalls in summer season (from June to August) is 892.1 mm, accounting for nearly 61% of annual average precipitation (KMA, 2019). Thermal gap between the warm air mass near the Pacific and the cold air mass near Siberia region causes a monsoon front and a typhoon every year. As a result of contribution of spatiotemporal inequality of rainfall, Gangneung city had experienced 870.5 mm of rainfall in just one day when the typhoon Rusa attacked South Korea (NIDP, 2002). Even worse, the risks of extreme rainfall are expected to increase every year due to climate change (Easterling et al., 2000).

According to the *annual report on disaster* by NEMA (2009-2018), in South Korea, typhoon and heavy rain occupied around 90%

disaster damage in a recent decade. The flood attacks natural environment and urban infrastructure and the destructed facilities and personal possessions transformed to disaster waste. Large amount of flood waste generation is a huge burden to achieve resilient and environmental sound society. Disaster waste matrix delays rescue activity by road blockage and induces secondary pollution problem such as leakage of contaminants to surrounding environment (Dubey et al., 2007; Pilapitiya et al., 2006). In addition, resource recirculation became difficult in disaster time when the waste is not recovered well (Wakabayashi et al., 2017). Therefore, rapid collection and disposal of flood waste is an important step toward safe society and benign environment.

Rapid and precise estimation is concieved to provide important design factors for planning and implementation of disaster waste management (USEPA, 2008). Curent flood waste management is comprised of three phase: collection, temporary storage, and disposal. Each step needs accurate amount of flood waste to calculate required resources, such as labor, vehicles, and site selection. The first trial in flood waste estimation was made by Hirayama and Kawata (2005). They demonstrated successful application of this method using damaged

buildings as input variables. Referring this example and the results of their own investigation, Korea Ministry of Environment also recommends to predict flood waste generation using unit waste generation value per damaged building (MOE, 2017). However, according to the previous investigation using Korean dataset (Park et al., 2019), the framework did not work well in explaning Korean cases.

The largest error in the model may be associated with the limitation of conventional model in reflecting regional characteristics of Korean cases. Every flood event contains heterogeneity originated from regional/meteorological property. Therefore a research need exists in developing country-specific model to estimate flood waste generation. In international view, the domestic trial can be shared as a reference materials to cope with each country's problem. To achieve that, the frameworks never tried in flood waste estimation is needed to be investigated while they exceed the limitation of conventional approaches. The frameworks addressing the case-dependent heterogeneity, stochastic properties of flood waste generation, interactive, and nonlinear patterns in flood waste generation were conceived to be promosing technologies. The detail on each methodology will be given in each section corresponding to the specific objectives.

1.2 Research objectives

The primary objective of this study is to suggest prospective modeling methodologies for estimating the amounts of waste generated from flood. Totally four frameworks were assessed as a potential modeling tools and these includes:

1) Deterministic linear regression with disaster context-based data grouping
2) Bayesian approach in linear regression
3) Incorporating the interaction terms for linear regression
4) Use of deep-neural network for nonlinear regression

Based on the aforementioned frameworks, the specific objectives of this study are:

1) Evaluating the effectiveness of each methodology
2) Drawing out the limitation of each framework
3) Unlocking the secret of flood waste generation patterns behind the current knowledge based on modeling results

1.3 Dissertation Structure

This dissertation is comprised of six chapters as shown in Fig. 1.1. Chapter 1 introduces current status in flood waste estimation research and relevant research questions in this study. Chapter 2 describes the applicability of disaster-context based data grouping to enhance the estimation performance of the multiple linear regression model. Chapter 3 presents the potential and limitation of Bayesian approach in multilinear regression to estimate flood waste generation. Chapter 4 addresses the contribution of interaction terms in flood waste linear model and the interpretation of modeling result to suggest potential disaster waste generation pattern. In chapter 5, the use of deep-neural network as a nonlinear modeling method was investigated. Chapter 6 summarizes all the previous chapters accompanied with conclusion drawn throughout this study.

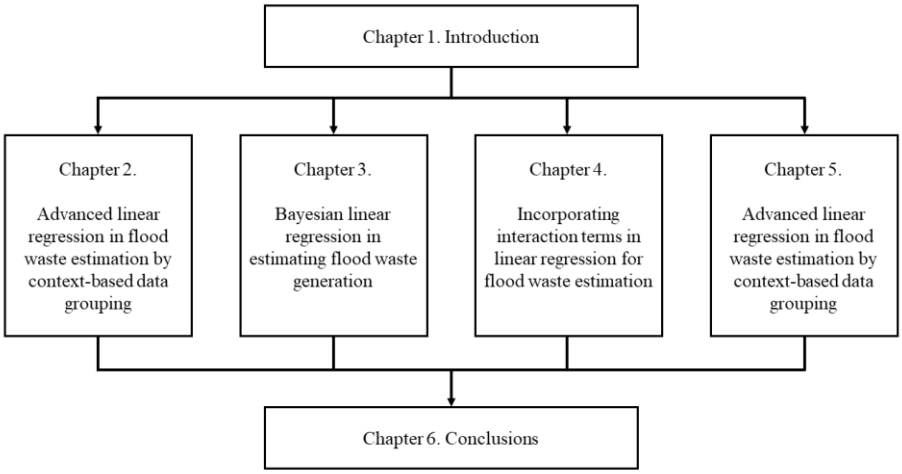

Figure 1.1 Structure of dissertation

References

Dubey, B., Solo-Gabriele, H.M., Townsendt, T.G., 2007. Quantities of arsenic-treated wood in demolition debris generated by Hurricane Katrina. Environ Sci Technol 41, 1533-1536.

Easterling, D.R., Evans, J., Groisman, P.Y., Karl, T.R., Kunkel, K.E., Ambenje, P., 2000. Observed variability and trends in extreme climate events: a brief review. Bulletin of the American Meteorological Society 81, 417-426.

Hirayama, N., Kawata, Y., 2005. Quantity of disaster waste for emergency response of public authirities on flood disaster. Institute of Social Safety Science 7, 325-330.

KMA, 2019. Climate - Regional Characteristics, https://www.weather.go.kr/weather/climate/average_regional.jsp (Access date: 2019. 8.17.)

MOE, 2017. Disaster Waste Safety Management Guidelines.

NEMA, 2009-2018. Annual Report on Disaster.

NIDP, 2002. The Field Survey Report of Damages Caused by the Typhoon RUSA in 2002 - 8.30 ~ 9.1.

Park, M.H., Ju, M., Kim, J.Y., 2019. Limitation of deterministic regression on flood waste quantity estimation in South Korea. Proceedings of 17th International Waste Management and Landfill Symposium,, 387, 381-388.

Pilapitiya, S., Vidanaarachchi, C., Yuen, S., 2006. Effects of the tsunami on waste management in Sri Lanka. Waste Manag 26, 107-109.

USEPA, 2008. Planning for natural disaster debris, Office of Solid Waste and Emergency Response and Office of Solid Waste

Wakabayashi, Y., Peii, T., Tabata, T., Saeki, T., 2017. Life cycle assessment and life cycle costs for pre-disaster waste management systems. Waste Management 68, 688-700.

CHAPTER 2

ADVANCED LINEAR REGRESSION IN FLOOD WASTE ESTIMATION BY CONTEXT-BASED DATA GROUPING

2.1 Introduction

Disaster waste management is a crucial issue in emergency response and effectiveness for recovery stages in terms of both economic feasibility and environmental protection. Disaster waste can block road used by rescuers and otherwise impede recovery activities. Crowley (2017) revealed that preparedness and planning for disaster waste management impacted the effectiveness of recovery and recycling processes. According to the Federal Emergency Management Agency (FEMA, 2012), in recent cases in the United States, the average costs for waste management accounted for approximately 27 percent of disaster management costs. The disaster waste matrix lists hazardous materials related to household or construction and demolition waste (Dubey et al., 2007; Pilapitiya et al., 2006). Without proper management, these hazardous materials can pollute the environment by entering the soil and groundwater. Disaster waste can also threaten public health by creating

breeding sites for vermin and disease vectors (Petersen, 2004).

Lessons learned from previous events show that pre-disaster waste estimations are essential for disaster waste management. Chen et al. (2007) stated that one of the reasons for failure in waste management after Typhoon Nali in 2001 was the failure to estimate possible waste generation. The United States Environmental Protection Agency (USEPA, 2008) stated that identifying and forecasting debris amounts and types are first step to prepare disaster waste management practice, and lessons learned from past events suggest waste estimations are beneficial for both disaster planning and response. Pre-estimation of disaster waste provides information required for site designation and design of temporary staging sites, waste collection routing, and funding from public sectors.

In South Korea, floods are the most prevalent type of natural disaster and have generated a large amount of waste (Jeong and Kim, 2012). Extreme rainfall events are increasing as global climate change alters rainfall patterns (Easterling et al., 2000). Because flood events are increasing, both in frequency and intensity, it has become necessary to develop a proper framework for flood waste management. Relevant regulations or guidelines, however, are currently insufficient due to a lack of historical experience with these types of disasters (Oh and Kang, 2013). A flood waste estimation model should therefore be

developed in order to establish such a framework.

The multiple regression method has been recognized as a viable technique for developing flood waste estimation models in many countries (Chen et al., 2007; Gabrielli et al., 2018; Hirayama and Kawata, 2005). Briefly, this method is a statistical tool finding and explaining the linear relationships between target and input variables. Compared to other nonlinear modeling methods, linear fitting poses ease in modeling and practical operation. Damage variables including factors like a number of damaged housings, length of damaged road, were suggested as independent variables for regression, in the flood waste model for South Korea (Cho, 2018). Since these parameters can be obtained earlier than waste can be quantified, the framework ensures promptness in waste estimation. In case of flood, a GIS or hazard map can provide information on damage variables prior to event occurrence through application of the expected rainfall scenario on equipped map data (Hirayama et al., 2009; Hirayama et al., 2010; USEPA, 2008).

However, despite efforts to improve model, its accuracy is still in question. The largest error in the model is related to the complexity of disaster data. Every flood event contains heterogeneity stemming from disaster context and regional characteristics. A downscaling strategy is therefore a viable option for reducing error related to such context-specific complexities (Jo and Kim,

2016). For example, the Ministry of Environment Japan (MOEJ, 2005) has suggested that application of the regression method may enhance estimation accuracy at the regional scale (equivalent to the town-scale) to enhance estimation accuracy.

In South Korea, the National Emergency Management Agency (NEMA, 2009-2018) had flood waste records for 92 cases of flooding between 2008-2017, and these records were insufficient for developing a town-scale estimation model. In order to substitute the regional separation of the dataset, it was postulated that data grouping by properties other than region would improve the accuracy of waste estimation. In this study, waste data were classified based on various scenarios and the waste estimation model was developed. The specific aims of this study were (1) to examine the effectiveness of data grouping on the improvement of waste prediction; and (2) to suggest the most appropriate grouping sequence. The scope of this study was limited to flood data in South Korea, but the concepts discussed in this study can be applicable to other countries. Insufficient disaster information is a common problem in many countries (Brown et al., 2011), therefore this chapter presents a novel strategy for improving estimations in similar situations at the international level.

2.2 Materials and methods

2.2.1 Data collection

Totally 90 cases of flood waste generation, between 2008 and 2017, were collected from the *Annual Report on Disaster* from 2008 to 2017 (NEMA, 2009-2018). Flood damage data before 2008 were not accessible because records had not been kept. Two cases were excluded from the whole events because their waste generation values appeared to be outliers. The values of the two excluded points were 90,330 and 160,778 t, respectively, while the maximum value from remaining 90 cases was 15,000 t, which was similar to other values that included: 14,000, 11,000, 10,790 t, etc.

Flood damage data, including flood waste generation, number of damaged buildings, area of damaged cropland (ha), length of damaged roads (m), length of damaged rivers (m), and length of damaged small streams (m) were utilized for the following analysis. The number of damaged buildings was classified into three types: totally destructed, half-destructed and just flooded buildings. These parameters were selected as independent variables for regression studies. Parameter selections were based on a previous study in South Korea (Cho, 2018). Damaged buildings have been recognized as significant waste sources. As Chen et al. (2007) suggested, the flooded area can

be used as an independent variable for flood waste estimation. However, records of flooded areas were inaccessible and thus other parameters including damaged cropland, roads, rivers and small streams were used as alternatives. Damaged croplands can be sources of agriculture-derived waste in times of disaster. Damaged road and river infrastructure can generate large quantities of demolition debris. In addition, damaged rivers or streams can transport plants waste and debris from upstream regions.

The author tried to utilize damage variables accessible in *Annual Report on Disaster* because unifying source is important to ensure consistency in data. Obviously, there are other damage variables including damaged cattle sheds, ships, plastic greenhouse, etc. However, these variables were too biased to some outlier cases, thus seemed not conform to the assumption of linear regression. The selected variables in upper paragraph showed relatively even distribution, convincing the robustness of modeling.

2.2.2 Multilinear regression

Multilinear regression was performed to estimate coefficients of post-event linear model using RStudio ver.1.1.442. The structure of the model equation is as follows:

$$Y = \sum_{i=1}^{n} a_i X_i \tag{2.1}$$

where Y (dependent variable) is the flood waste generation in tons, X_i (independent variable) is the flood damage variable, and a_i is the coefficient for X_i. X_1, X_2, X_3, X_4, and X_5 stand for the number of damaged buildings, area of damaged cropland (ha), length of damaged roads (m), length of damaged river (m), and length of damaged small streams (m). The coefficients of coefficients of X_1, X_2, X_3, X_4, and X_5 are a_1, a_2, a_3, a_4, and a_5, respectively. The R script for multilinear regression in nonlinear least squares package (nls) was adjusted slightly to yield positive coefficients, as negative waste generation volumes are physically impossible. When removing some variables shows the better performance in terms of adjusted r^2, we selected the model of which some input variables were excluded.

2.2.3 Regression diagnosis

Residuals analysis was conducted to verify the fit of regression models. The standardized residual was utilized as a representative value and calculated as the following equation:

$$\hat{e}_{st,i} = \frac{\hat{e}_i}{\widehat{sd}(\hat{e}_i)} = \frac{y_i - \hat{y}_i}{\widehat{sd}(Y_i - \hat{Y}_i)} \tag{2.2}$$

where $\hat{e}_{st,i}$ is standardized residual, $\hat{e}_i$ is residual error, $\widehat{sd}(\hat{e}_i)$ is the estimated standard deviation of residual error, y_i is actual waste generation, $\hat{y}_i$ is estimated waste generation, and capital Y is real value which cannot be

obtained. Standardized residuals beyond the 95% confidence intervals were regarded as outliers and the existence of outliers in classified datasets was verified. Residuals computations were carried out using nls package in RStudio ver.1.1.442.

The normality and homoscedasticity of residuals were examined as these two properties are key features of the residuals according to the assumption of linear regression. The normality was examined by Shapiro-Wilk test using shaprio.test function in stats package of R ver.3.6.1. The homoscedasticity of residuals were examined by visual inspection. To be specific, it was interpreted that constant variation assumption was violated when standardized residuals showed biased patterns.

2.2.4 Sequential data grouping

Three data grouping scenarios were postulated to enhance the accuracy of waste prediction. These grouping included administrative region (AR), which referred to the scale of provincial or metropolitan area, urbanization rate (UR) of damaged region, and a combination of the factors of disaster type and offshore accessibility (DO). Disaster waste records were reported in the administrative region where the disaster occurred. Thus using AR for grouping was a convenient method. AR also accounts for regional

characteristics. UR was an important factor for determining municipal solid waste generation characteristics during non-disaster time. Hence, it could also affect disaster waste generation. Disaster type was related to rainfall or wind characteristics, and ocean accessibility was related to tidal effects on flood disaster.

Figure 2.1 shows the locations of metropolitan cities and provinces in South Korea. Each city or province is comprised of many municipalities. Detailed information about each city and province is summarized in Table 2.1. The UR was divided into intervals of 0–30, 30–40, 40–60, 60–70, 70–90, and 90–100%. The intervals of UR were not equal but were intentionally manipulated to ensure consistent sample sizes. The disaster types included the typhoon and heavy rain. It should be noted that these were arbitrary criteria not directly related to disaster context. Thus, this approach was somewhat limited, but using these criteria was necessary due to the lack of current information on disaster characterization.

Figure 2.1. A map of special cities and provinces in South Korea (Regions underlined with arrow bars are special cities; map by vemaps.com was slightly edited)

Table 2.1. Summary of special cities and provinces in South Korea

Name of province or special City	Type	Population	Area (km²)	Density (People/km²)
Busan	Metropolitan city	3,574,340	766	4,666
Daegu	Metropolitan city	2,512,604	884	2,842
Incheon	Metropolitan city	2,953,255	1,062	2,810
Gwangju	Metropolitan city	1,415,953	501	2,824
Daejeon	Metropolitan city	1,442,857	540	2,673
Sejong	Special autonomous city	122,263	465	380
Seoul	Special city	10,464,051	605	17,288
Ulsan	Metropolitan city	1,126,879	1,056	1,030
Chungcheongbuk-do	Province	1,588,633	7,433	213
Chungcheongnam-do	Province	2,064,665	8,204	251
Gangwon-do	Province	1,549,780	20,569	75
Gyeonggi-do	Province	12,239,862	10,171	1,203
Gyeongsangbuk-do	Province	2,739,179	19,030	144
Gyeongsangnam-do	Province	3,374,725	10,532	320
Jeollabuk-do	Province	1,895,882	8,043	236
Jeollanam-do	Province	1,938,136	11,858	163
Jeju Special Self-governing Province	Province	583,284	1,849	315

Model development was performed after data grouping by scenario using the aforementioned methodology. The grouping scenario required that the best fit between estimated and actual waste generations was selected. The basis of decision was not r^2 but adjusted r^2 to compensate the linearity increase due to the number of input variables rather its accuracy. One might raise a question regarding use of adjusted r^2 as a basis, however, obviously adjusted r^2 is one of the most prominent statistics used to describe the model performance in regression analysis field. The data estimated by multilinear regression were filtered from the entire dataset. The remaining data were then analyzed using a similar protocol. This process was repeated until subsequent grouping did not improve fitting.

2.3 Results and discussion

2.3.1 Multilinear regression without grouping

Multilinear regression between flood damage variables and flood waste amounts in 90 cases yielded the waste prediction model as noted in Eq. 2.3:

$$Y = 3.681X_1 + 0.154X_3 + 0.048X_4 + 0.064X_5 \quad (2.3)$$

Figure 2.2 shows the plot of reported and estimated flood waste generations. As shown in Figure 2.2, the model was developed using flood cases in which the amounts of flood waste generation were poorly predicted. The adjusted r^2, and standard error were 0.328 and 2,480 tons, respectively. P-values for the estimated coefficients were 7.34×10^{-7}, 0.228, 0.124, and 0.087 for a_1, a_3, a_4, and a_5. Results for the a_1, P-values were far from significant ($P < 0.05$).

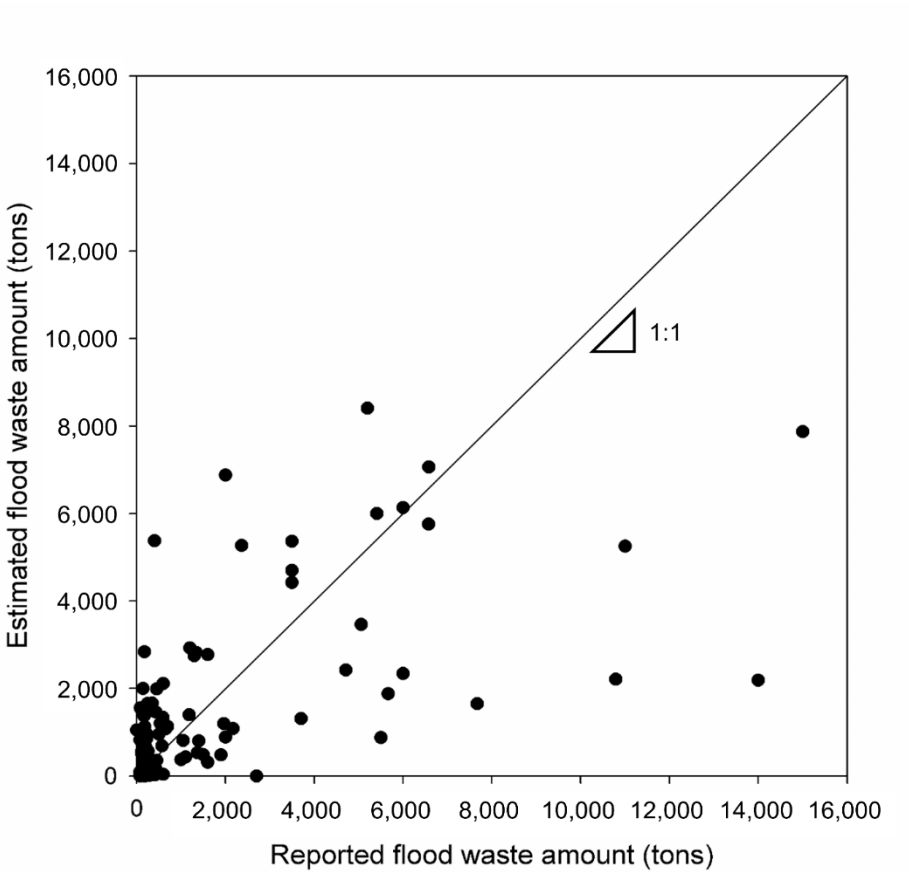

Figure 2.2. A plot of reported and estimated flood waste by the prediction model developed from 90 flood cases

Even though Eq. 2.3 was the best fit case for regression by various parameter selections, its prediction accuracy was insufficient for practical application. Because each flood case had a different disaster context, a single model based on the entire dataset was unable to capture the complex nature of the flood events. The low accuracy of Eq. 2.3 suggests that flood waste prediction model establishment must consider the context of each disaster event.

2.3.2 Post-event models following sequential data grouping

As noted in the Section 2.1, flood waste modeling that considers the disaster context will be beneficial for improving prediction accuracy. Data grouping by relevant disaster characteristics was performed before regression. Three stages of sequential modeling and grouping were conducted. The models showing adjusted r^2 larger than 0.7 were selected in the first and second stages, while those showing adjusted r^2 larger than 0.5 were selected in the final stage. The r^2 value criterion for model selection in each stage was arbitrarily selected, in order to select the most appropriate data by sequential grouping.

2.3.2.1 Post-event models from the first grouping

The first data grouping identified the well-fitted disaster events (Table 2.2). Among the three scenarios, grouping by AR was the most effective for improving prediction accuracy. In total, 31 cases (Gyeonggi-do, Jeollabuk-do, and Ulsan cases) showed enhanced accuracy by regression, with adjusted r^2 larger than 0.837. The administrative region may represent diverse factors related to flood waste generation that include geological/hydrological properties, disaster resilience, and household possessions. In terms of the number of data approving enhanced accuracy, grouping by DO was the most effective scenario. In all, 39 cases (heavy rain in offshore and typhoon in offshore cases) showed a better fit, with adjusted r^2 larger than 0.7. In the case of sorting by UR, only 13 cases (40-60%) were selected by grouping. The selected cases showed enhanced accuracy based on not only adjusted r^2 but also on standard error. Every selected dataset exhibited a standard error lower than that of the model developed with the unclassified data set (2,480 tons). Moreover, *P*-values of estimated coefficients in this section (Table 2.3) were less than 0.05, except for a_5 in the model for Gyeonggi-do. The *P*-value for the coefficient was 0.078. Considering that *P*-values of three out of four parameters were larger than 0.05 in the model developed in Section 2.3.1, the lowered *P*-values indicate the enhanced credibility of the results.

Table 2.2. Summary of flood waste modelling after first data grouping

Grouping scenario	Group property	Cases number	Model equation	Adjusted r^2	Standard error
Administrative region	Chungcheongnam-do	8	$Y = 5.25X_1 + 0.0409X_5$	0.410	314
	Gyeonggi-do*	10	$Y = 7.65X_1 + 0.0449X_5$	0.930	1,192
	Gyeongsangbuk-do	6	$Y = 0.712X_4$	0.462	2,654
	Gyeongsangnam-do	7	$Y = 0.179X_3$	-0.732	771
	Jeollabuk-do*	13	$Y = 1.55X_1 + 0.118X_4$	0.837	827
	Jeollanam-do	30	$Y = 4.17X_1 + 0.214X_4$	0.050	1,414
	Ulsan*	8	$Y = 29.9X_1$	0.930	1,452
Urbanization rate	0-30%	11	$Y = 5.22X_1$	-0.427	240
	30-40%	16	$Y = 0.816X_3$	0.152	2,442
	40-60%*	13	$Y = 38.1X_2$	0.713	1,001
	60-70%	20	$Y = 10.2X_2 + 0.146X_4$	0.659	1,726
	70-90%	13	$Y = 2.15X_1 + 0.136X_5$	0.496	1,452
	90-100%	17	$Y = 1.69X_1 + 1.58X_3 + 0.249X_4$	0.492	3,341
Disaster type and location	Heavy rain in inland	28	$Y = 3.16X_1 + 0.475X_3 + 0.0859X_4$	0.360	2,966
	Heavy rain in offshore*	9	$Y = 1.68X_1 + 6.71X_2$	0.965	167
	Typhoon in inland	23	$Y = 12.5X_1$	0.515	2,172
	Typhoon in offshore*	30	$Y = 8.20X_1 + 29.4X_2$	0.732	1,225

*: selected cases among models developed after first grouping (adjusted $r^2 > 0.7$)

Table 2.3. *P*-values of coefficients estimated in regression after first classification

Classification scenario	Group property	*P*-value				
		a_1	a_2	a_3	a_4	a_5
Administrative region	Chungcheongnam-do	6.61×10^{-3}	-	-	-	8.82×10^{-2}
	Gyeonggi-do*	1.17×10^{-6}	-	-	-	7.77×10^{-2}
	Gyeongsangbuk-do	-	-	8.60×10^{-3}	-	-
	Gyeongsangnam-do	-	-	2.80×10^{-1}	-	-
	Jeollabuk-do*	2.14×10^{-2}	-	-	1.13×10^{-5}	-
	Jeollanam-do	1.04×10^{-1}	-	-	1.82×10^{-2}	-
	Ulsan*	3.30×10^{-6}	-	-		-
Urbanization rate	0-30%	9.28×10^{-3}	-	-		-
	30-40%	-	-	5.46×10^{-3}		-
	40-60%*	-	3.00×10^{-5}	-		-
	60-70%	-	3.05×10^{-1}	-	9.20×10^{-3}	-
	70-90%	3.25×10^{-2}	-	-	-	1.92×10^{-3}
	90-100%	2.12×10^{-1}	-	9.84×10^{-3}	1.69×10^{-1}	-
Disaster type and location	Heavy rain in inland	2.74×10^{-3}	-	6.96×10^{-2}	2.84×10^{-1}	-
	Heavy rain in offshore*	2.19×10^{-6}	1.08×10^{-3}	-	-	-
	Typhoon in inland	4.93×10^{-6}	-	-	-	-
	Typhoon in offshore*	1.70×10^{-4}	1.22×10^{-6}	-	-	-

*: selected cases among models developed after first classification (adjusted $r^2 > 0.7$), and -: unselected independent variable for modeling

It was observed that the number of damaged buildings was selected in most classified cases. Except for two cases (Jeollanam-do, and 90-100% of UR), *P*-values for a_1 were lower than 0.05, implying that damaged buildings were significant contributor to flood waste generation. The relationship between damaged buildings and flood waste has already been identified in previous studies (Cho, 2018; Jeong and Kim, 2012; Kang et al., 2015). Given the importance of damaged buildings in flood waste generation, Korea Ministry of Environment suggested using a 1.7 ton/damaged building value as a flood waste prediction unit (MOE, 2017). According to the results shown in Table 2.2, in most cases a_1 was larger than 1.7. Even when a_1 is similar to 1.7, the contributions of other parameters on waste generation should be incorporated. This is because the complexity of disaster waste generation cannot be accurately conveyed with a single unit of waste generation value. The results and subsequent findings of this study showed not only the effectiveness of grouping but also provide a rationale for using data grouping.

A least 51 cases, more than half of total, remained at low accuracy after the first grouping. Even some of the sorted dataset suffered from diminished accuracy. In cases of where the dataset sorted by AR, data from Gyeongsangnam-do, and Jeollanam-do showed the adjusted r^2 to be less than that of the model developed in Section 2.3.1 (0.328). In cases of sorting by UR,

data from 0-30% and 30-40% had decreased adjusted r^2. Data sorted by DO did not show any prediction aggravation. This indicated that the first grouping was not sufficient to improve the model.

The first grouping proved that it is possible to isolate a dataset comprised of events from similar contexts. However, the ability to do this was limited to a part of the dataset because the criteria used for data sorting in this study were not specific enough to sort disaster cases with homogeneous characteristics. For example, provincial regional grouping explains the regions comprised of sub-regions, and the smaller sub-regions are comprised of smaller unit regions, equivalent to town or village. The sum of such diverse properties within each small component results in heterogeneity in the group. The concept was too broad to explain differences among disaster cases within the group. It would be impossible, however, to apply narrower criteria for data grouping because of the limited number of disaster records. Successive grouping was therefore used to overcome this limitation.

2.3.2.2 Post-event models from the second grouping

After the first grouping, datasets showing better fit were filtered by each grouping scenario, and the remaining cases were studied in this section. Totals of 59, 77, and 51 cases remained after grouping by AR, UR, and DO,

respectively. For further prediction enhancement, remainders were classified by another scenario not used in the first grouping.

Although some of the data had already been sorted, the second grouping led to improvements in prediction in some cases. Table 2.4 summarizes the results of flood waste regression after the second grouping and Table 2.5 summarizes the *P*-values of the obtained coefficients in the models. The cases showing enhanced prediction with adjusted r^2 greater than 0.7 existed in every second grouping scenario. In some cases, the waste prediction was improved after the second grouping although the same criteria in the first grouping did not lead to enhanced waste prediction. For example, the dataset grouped by 90-100% UR in the first grouping did not show model credibility considering its low adjusted r^2 (Table 2.2), however, the dataset grouped by UR after being filtered by AR resulted in higher model accuracy as evidenced by increased adjusted r^2 (Table 2.4). A similar phenomenon was observed in Chungcheongnam-do data after the first grouping by UR. This was because some of the events in the first grouping had different disaster contexts which were filtered after the first grouping by other criteria. This implies that data grouping not only leads to identifying a data cluster with similar contexts, but also facilitates screening out the data that can impede adequate clustering.

Table 2.4. Summary of flood waste modeling after second data grouping

Prior grouping scenario	Second grouping scenario	Group property	Number of dataset	Model equation	Adjusted r^2	Standard error
Administrative region	Urbanization rate	0-30%	10	$Y = 0.0570X_5$	-0.345	240
		30-40%	13	$Y = 0.428X_3 + 0.414X_4$	0.290	2,388
		40-60%	7	$Y = 2.53X_1$	-0.693	429
		60-70%	12	$Y = 10.1X_1 + 0.257X_4$	0.331	1,675
		70-90%	9	$Y = 0.193X_3 + 0.109X_5$	0.383	1,539
		90-100%*	8	$Y = 1.46X_1 + 0.508X_3 + 0.596X_4$	0.753	1,251
	Disaster type and location	Heavy rain in inland	15	$Y = 1.28X_1 + 1.24X_3 + 0.247X_4$	0.513	2,418
		Heavy rain in offshore*	7	$Y = 14.3X_2$	0.938	175
		Typhoon in inland	16	$Y = 2.41X_1 + 10.2X_2$	0.607	709
		Typhoon in offshore	21	$Y = 5.15X_1 + 0.192X_5$	-0.153	1,085
Urbanization rate	Administrative region	Chungcheongnam-do*	7	$Y = 7.33X_1 + 0.000415X_2 + 0.0315X_5$	0.960	85
		Gyeonggi-do*	9	$Y = 7.86X_1$	0.967	853
		Gyeongsangbuk-do	6	$Y = 0.712X_4$	0.462	2,654
		Gyeongsangnam-do	6	$Y = 0.358X_3$	-0.575	766
		Jeollabuk-do*	8	$Y = 1.67X_1 + 33.6X_2$	0.987	264
		Jeollanam-do	25	$Y = 4.71X_1 + 0.208X_4$	0.042	1,541
		Ulsan*	8	$Y = 29.9X_1$	0.930	1,452
		Heavy rain in inland	24	$Y = 3.12X_1 + 0.636X_3$	0.356	3,133

	Disaster type and location	Heavy rain in offshore*	9	$Y = 1.68X_1 + 6.71X_2$	0.965	167
		Typhoon in inland	18	$Y = 13.4X_1$	0.545	2,340
		Typhoon in offshore*	26	$Y = 8.54X_1 + 28.9X_2$	0.730	1,301
Disaster type and location	Administrative region	Gyeonggi-do*	10	$Y = 7.65X_1 + 0.0449X_5$	0.930	1,192
		Jeollabuk-do*	9	$Y = 0.139X_4$	0.879	826
		Jeollanam-do	13	$Y = 0.401X_4$	0.352	1,635
	Urbanization rate	0-30%	7	$Y = 5.95X_1$	-2.180	253
		30-40%	7	$Y = 0.507X_4$	0.218	3,348
		40-60%*	9	$Y = 38.0X_2$	0.707	1,171
		60-70%	12	$Y = 10.7X_1$	0.401	1,975
		70-90%	7	$Y = 0.133X_4 + 0.105X_5$	0.339	1,859
		90-100%	9	$Y = 2.26X_3 + 0.348X_4$	0.509	4,068

*: selected cases among models developed after second grouping (adjusted $r^2 > 0.7$)

Table 2.5. *P*-values of coefficients estimated in regression after second classification

First classification scenario	Second classification scenario	Group property	*P*-value				
			a_1	a_2	a_3	a_4	a_5
Administrative region	Urbanization rate	0-30%	-	-	-	-	9.17×10^{-3}
		30-40%	-	-	3.10×10^{-1}	1.11×10^{-1}	-
		40-60%	9.13×10^{-2}	-	-	-	-
		60-70%	1.94×10^{-1}	-	-	1.03×10^{-1}	-
		70-90%	-	-	3.09×10^{-1}	-	1.28×10^{-1}
		90-100%*	3.40×10^{-2}	-	2.43×10^{-1}	5.36×10^{-2}	-
	Disaster type and location	Heavy rain in inland	1.91×10^{-1}	-	4.28×10^{-2}	2.37×10^{-1}	-
		Heavy rain in offshore*	-	-	-	-	9.17×10^{-3}
		Typhoon in inland	2.70×10^{-1}	1.41×10^{-2}	-	-	-
		Typhoon in offshore	1.52×10^{-2}	-	-	-	1.10×10^{-1}
Urbanization rate	Administrative region	Chungcheongnam-do*	7.35×10^{-5}	2.72×10^{-2}	-	-	4.86×10^{-3}
		Gyeonggi-do*	4.07×10^{-8}	-	-	-	-
		Gyeongsangbuk-do	-	-	-	8.60×10^{-3}	-
		Gyeongsangnam-do	-	-	1.79×10^{-1}	-	-
		Jeollabuk-do*	1.23×10^{-4}	1.83×10^{-6}	-	-	-
		Jeollanam-do	1.54×10^{-1}	-	-	4.21×10^{-2}	-
		Ulsan*	3.30×10^{-6}	-	-	-	-

	Disaster type and location	Heavy rain in inland	5.40×10^{-3}	-	5.13×10^{-3}	-	-
		Heavy rain in offshore	2.19×10^{-6}	1.08×10^{-3}	-	-	-
		Typhoon in inland	2.49×10^{-5}	-	-	-	-
		Typhoon in offshore*	4.00×10^{-4}	8.19×10^{-6}	-	-	-
Disaster type and location	Administrative region	Gyeonggi-do*	1.17×10^{-6}	-	-	-	7.77×10^{-2}
		Jeollabuk-do*	-	-	-	1.19×10^{-5}	-
		Jeollanam-do	-	-	-	4.66×10^{-3}	-
	Urbanization rate	0-30%	9.98×10^{-2}	-	-	-	-
		30-40%	-	-	-	4.90×10^{-2}	-
		40-60%*	-	3.08×10^{-4}	-	-	-
		60-70%	1.60×10^{-3}	-	-	-	-
		70-90%	-	-	-	2.05×10^{-1}	1.64×10^{-1}
		90-100%	-	-	4.75×10^{-3}	2.49×10^{-1}	-

*: selected cases among models developed after second classification (adjusted $r^2 > 0.7$)

-: unselected independent variable for modeling

In the second stage modeling, the number of damaged buildings was the most frequently selected parameter in the model, and the *P*-value for this parameter was almost less than 0.05. These results were consistent with those in Section 2.3.2.1, indicating that damaged buildings comprised the most significant source of disaster waste. Similar to the finding in Section 2.3.2.1, a_1 values were different from the unit waste generation (1.7 ton/building) suggested by Korea Ministry of Environment (MOE, 2017), emphasizing that this value was inadequate for predicting the disaster waste generation.

2.3.2.3 Post-event models from the third grouping

More than 32 cases remained after the second grouping, depending on each grouping sequence. Since only three scenarios were used in this study, the third data grouping was the final sorting process. The modeling results after the third grouping are summarized in Table 2.6 and the *P*-values of the obtained coefficients are summarized in Table 2.7.

Table 2.6. Summary of flood waste modeling after third data grouping

Prior grouping scenarios	Third grouping scenario	Group property	Number of dataset	Model equation	Adjusted r^2	Standard error
1. Administrative region	Disaster type and location	Heavy rain in inland*	11	$Y = 33.2X_1 + 0.897X_3$	0.691	2,048
		Heavy rain in offshore*	7	$Y = 14.3X_2$	0.938	175
2. Urbanization rate		Typhoon in inland*	16	$Y = 2.41X_1 + 10.2X_2$	0.607	709
		Typhoon in offshore	17	$Y = 6.21X_1$	-0.358	963
1. Administrative region	Urbanization rate	0-30%	9	$Y = 0.0434X_5$	-0.963	217
		30-40%	11	$Y = 0.612X_4$	0.218	2,611
2. Disaster type and location		40-60%	7	$Y = 2.53X_1$	-0.693	429
		60-70%	12	$Y = 10.1X_1 + 0.257X_4$	0.331	1,675
		70-90%*	5	$Y = 0.800X_3$	0.586	1,527
		90-100%*	8	$Y = 1.46X_1 + 0.508X_3 + 0.596X_4$	0.753	1,251
1. Urbanization rate	Disaster type and location	Heavy rain in inland*	11	$Y = 1.22X_1 + 1.10X_3 + 0.404X_4$	0.535	2,539
2. Administrative region		Heavy rain in offshore*	6	$Y = 14.3X_2$	0.935	191
		Typhoon in inland*	12	$Y = 0.558X_3$	0.765	609
		Typhoon in offshore	16	$Y = 3.49X_1 + 0.222X_3 + 0.164X_5$	-0.193	1,208
1. Urbanization rate	Administrative region	Gyeonggi-do*	9	$Y = 7.88X_1$	0.967	853
2. Disaster type and location		Jeollabuk-do*	5	$Y = 9.36X_1$	0.992	249
		Jeollanam-do*	10	$Y = 36.0X_1$	0.536	1,574
	Urbanization rate	0-30%	6	$Y = 10.3X_1$	-1.702	234

1. Disaster type and location		30-60%	10	$Y = 0.217X_5$	0.046	3,187
		60-70%	7	$Y = 34.1X_1$	0.469	1,954
2. Administrative region		70-100%	9	$Y = 1.99X_1 + 0.854X_4$	0.031	4,382
1. Disaster type and location	Administrative region	Gyeonggi-do*	9	$Y = 7.88X_1$	0.967	853
		Jeollanam-do*	10	$Y = 36.0X_1$	0.536	1,574
2. Urbanization rate						

*: selected cases among models developed after third grouping (adjusted $r^2 > 0.5$)

Table 2.7. *P*-values of coefficients estimated in regression after third classification

Prior classification scenario	Third classification scenario	Group property	*P*-value				
			a_1	a_2	a_3	a_4	a_5
1. Administrative region	Disaster type and location	Heavy rain in inland*	1.81×10^{-2}	-	8.23×10^{-2}	-	-
2. Urbanization rate		Heavy rain in offshore*	-	2.01×10^{-5}	-	-	-
		Typhoon in inland*	2.70×10^{-1}	1.41×10^{-2}	-	-	-
		Typhoon in offshore	7.49×10^{-3}	-	-	-	-
1. Administrative region	Urbanization rate	0-30%	-	-	-	-	3.78×10^{-2}
2. Disaster type and location		30-40%	-	-	-	5.64×10^{-3}	-
		40-60%	9.13×10^{-2}	-	-	-	-
		60-70%	1.94×10^{-1}	-	-	1.03×10^{-1}	-
		70-90%*	-	-	1.60×10^{-2}	-	-
		90-100%*	3.40×10^{-2}	-	2.43×10^{-1}	5.36×10^{-2}	-
1. Urbanization rate	Disaster type and location	Heavy rain in inland*	2.46×10^{-1}	-	1.06×10^{-1}	1.26×10^{-1}	-
2. Administrative region		Heavy rain in offshore*	-	1.08×10^{-4}	-	-	-
		Typhoon in inland*	-	-	2.51×10^{-6}	-	-
		Typhoon in offshore	2.12×10^{-1}	-	2.04×10^{-1}	-	2.34×10^{-1}
1. Urbanization rate	Administrative region	Gyeonggi-do*	4.07×10^{-8}	-	-	-	-
2. Disaster type and location		Jeollabuk-do*	1.21×10^{-5}	-	-	-	-
		Jeollanam-do*	2.77×10^{-3}	-	-	-	-
1. Disaster type and location	Urbanization rate	0-30%	5.72×10^{-2}	-	-	-	-
2. Administrative region		30-60%	-	-	-	-	6.57×10^{-2}

		60-70%	1.82×10^{-2}	-	-	-	-
		70-100%	2.60×10^{-1}	-	-	5.50×10^{-2}	-
1. Disaster type and location	Administrative region	Gyeonggi-do*	4.07×10^{-8}	-	-	-	-
2. Urbanization rate		Jeollanam-do*	2.77×10^{-3}	-	-	-	-

*: selected cases among models developed after third classification (adjusted $r^2 > 0.7$)

-: unselected independent variable for modeling

The effect of the third grouping on model development was similar to the results obtained in previous sections. The third grouping resulted in the clusters showing better fit, likely for the reasons as discussed in Section 2.3.2.1. This shows that continuous grouping, at least till the third stage, was effective for sorting the dataset within similar contexts. Therefore, it can be concluded that when it is not possible to group data by narrower criteria, sequential grouping by broader characteristics can effectively overcome the limitations of disaster waste data prediction.

2.3.3 Summary of linear modeling with grouping

Overall modeling results are summarized in Table 2.8. Three-stage grouping resulted in advanced flood waste prediction with adjusted r^2 larger than 0.535 and at least 51 of the 90 total cases were enhanced through modeling. The number of selected cases and adjusted r^2 showed data grouping to be a promising tool for enhanced flood waste prediction. There were six grouping sequences in total, and with no established criterion for evaluating effectiveness it was difficult to define which sequence was best for enhancing prediction. For example, grouping in the order of UR, AR, and DO showed the fewest remainders, however, the minimum adjusted r^2 value was lower in that sequence than it was in other grouping sequences. An index incorporating the

number of selected cases and accuracy within an established model should be utilized to evaluate the effectiveness of each sequence.

Table 2.8. Summary of flood waste modelling with data grouping

Grouping sequence	Summary of first grouping	Summary of second grouping	Summary of third grouping	Summary of whole selected cases
1. Administrative region 2. Urbanization rate 3. Disaster type and location	● 31 cases selected ● Adjusted $r^2 > 0.837$	● 8 cases selected ● Adjusted $r^2 > 0.753$	● 34 cases selected ● Adjusted $r^2 > 0.607$	● 73 cases selected ● Adjusted $r^2 > 0.607$
1. Administrative region 2. Disaster type and location 3. Urbanization rate		● 7 cases selected ● Adjusted $r^2 > 0.938$	● 13 cases selected ● Adjusted $r^2 > 0.586$	● 51 cases selected ● Adjusted $r^2 > 0.586$
1. Urbanization rate 2. Administrative region 3. Disaster type and location	● 13 cases selected ● Adjusted $r^2 > 0.713$	● 32 cases selected ● Adjusted $r^2 > 0.930$	● 29 cases selected ● Adjusted $r^2 > 0.535$	● 74 cases selected ● Adjusted $r^2 > 0.535$
1. Urbanization rate 2. Disaster type and location 3. Administrative region		● 35 cases selected ● Adjusted $r^2 > 0.730$	● 24 cases selected ● Adjusted $r^2 > 0.420$	● 74 cases selected ● Adjusted $r^2 > 0.536$
1. Disaster type and location 2. Administrative region	● 39 cases selected ● Adjusted $r^2 > 0.732$	● 19 cases selected ● Adjusted $r^2 > 0.879$	● 0 case selected ● Adjusted $r^2 < 0.469$	● 58 cases selected ● Adjusted $r^2 > 0.732$

3. Urbanization rate			
1. Disaster type and location	● 9 cases selected	● 19 cases selected	● 67 cases selected
2. Urbanization rate	● Adjusted $r^2 > 0.707$	● Adjusted $r^2 > 0.536$	● Adjusted $r^2 > 0.536$
3. Administrative region			

In order to examine which sequence was most effective, grouping effectiveness index (*GEI*) was devised:

$$GEI = \sum number\ \ of\ \ selected\ \ cases\ \ \times adjusted\ \ r^2\ \ in\ \ the\ \ group \quad (5)$$

The index can explain both the number of dataset with enhanced modeling and the model performance in each group and the larger value of index was regarded as better grouping performance. The *GEI* of each grouping sequence is summarized in Table 2.9. According to the *GEI* calculation, the grouping sequence with the order of UR, AR, and DO showed the highest *GEI* (60.703), satisfying the number of selected cases and model accuracy for each group.

Table 2.9. Calculation sheet of grouping effectiveness index (GEI)

First scenario	Group property	No. data	Adjusted r^2	Second scenario	Group property	No. data	Adjusted r^2	Third scenario	Group property	No. data	Adjusted r^2	*GEI*
Administrative region	Gyeonggi-do	10	0.93	Urbanization rate	90-100%	8	0.753	Disaster type and location	Heavy rain in inland	11	0.691	57.524
	Jeollabuk-do	13	0.837						Heavy rain in offshore	7	0.938	
	Ulsan	8	0.93						Typhoon in inland	16	0.607	
Administrative region	Gyeonggi-do	10	0.93	Disaster type and location	Heavy rain in offshore	7	0.938	Urbanization rate	70-90%	5	0.586	43.141
	Jeollabuk-do	13	0.837						90-100%	8	0.753	
	Ulsan	8	0.93									
Urbanization rate	40-60%	13	0.713	Administrative region	Chungcheonnam-do	7	0.96	Disaster type and location	Heavy rain in inland	11	0.535	60.703
					Gyeonggi-do	9	0.967		Heavy rain in offshore	6	0.935	
					Jeollabuk-do	8	0.987		Typhoon in inland	12	0.765	
					Ulsan	8	0.93					

Urbanization rate	40-60%	13	0.705	Disaster type and location	Typhoon in offshore	26	0.73	Administrative region	Gyeonggi-do	9	0.967	47.168
									Jeollabuk-do	5	0.992	
									Jeollanam-do	10	0.536	
Disaster type and location	Heavy rain in offshore	9	0.965	Administrative region	Gyeonggi-do	10	0.93	Urbanization rate	-	-	-	47.856
	Typhoon in offshore	30	0.732		Jeollabuk-do	9	0.879		-	-	-	
Disaster type and location	Heavy rain in offshore	9	0.965	Urbanization rate	40-60%	9	0.707	Administrative region	Gyeonggi-do	9	0.967	51.071
	Typhoon in offshore	30	0.732						Jeollanam-do	10	0.536	

Fig. 2.3 shows flood waste prediction by the selected models with the grouping sequence that had the highest *GEI*. Compared to Fig. 2.2, linear relationships between reported and estimated flood waste generation were far clearer, indicating flood waste prediction improvement. It should be noted, however, that points were biased to end-points in some cases, as in the model developed with flood cases in Jeollabuk-do (Fig. 2.3). Bias to extreme cases aggravates the prediction in other ranges, thus further study is necessary to achieve better estimation in every waste generation range. However, it remains clear that the classified dataset resulted in better predictions than the unclassified dataset (See Fig. 2.2 and Fig. 2.3).

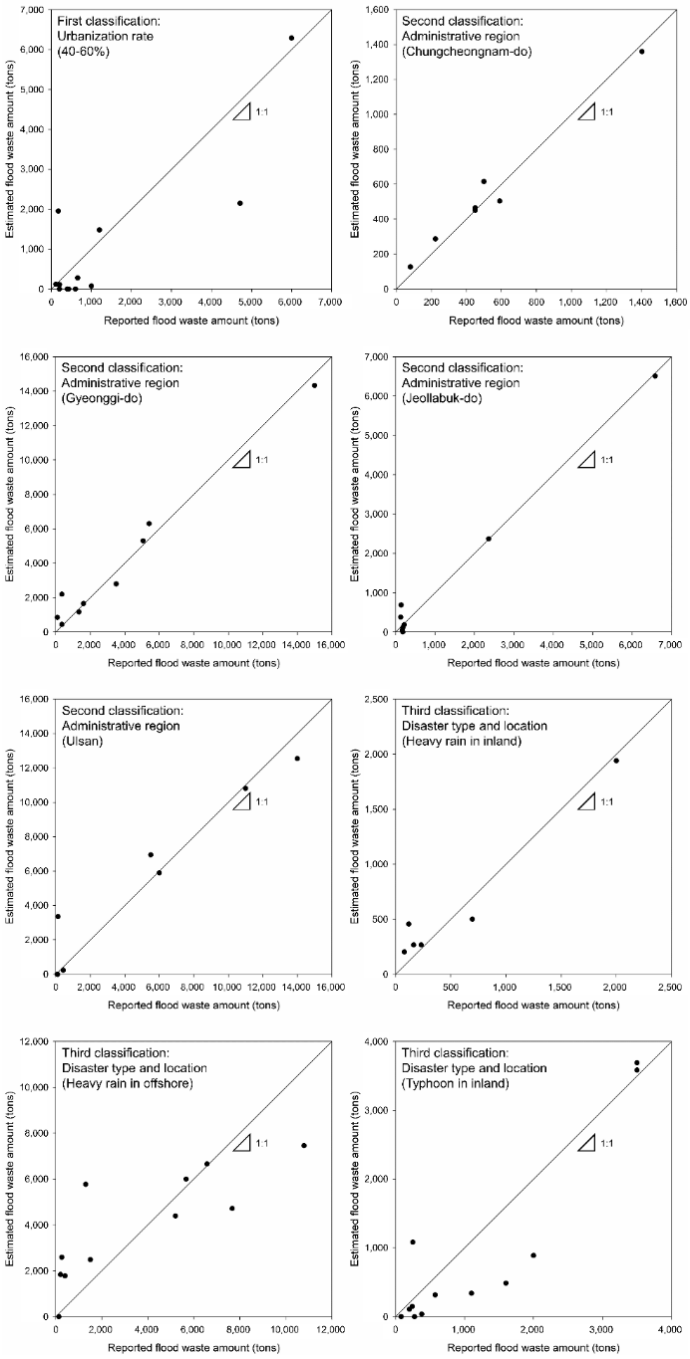

Figure 2.3. Flood waste estimation plots of the best-fit grouping sequence

2.3.4 Regression diagnosis for grouped models

The distribution of residuals is an important tool for examining whether or not a regression model is acceptable. Fig. 2.4 shows the plots of standardized residuals depending on fitted values by the regression model. The regression model using the unclassified dataset showed seven residuals exceeding the 95% confidence interval, indicating poor prediction by that model. Residuals exceeding the confidence interval suggest that the data did not conform to the assumption of multilinear regression (Brown and Mac Berthouex, 2002). In other words, the unclassified dataset cannot be explained by multilinear regression due to their inherent heterogeneity. The residuals of the regression by the whole dataset did not conform to normality assumption according to the low *p*-value in Shapiro-Wilk test ($P < 0.01$; Table 2.10).

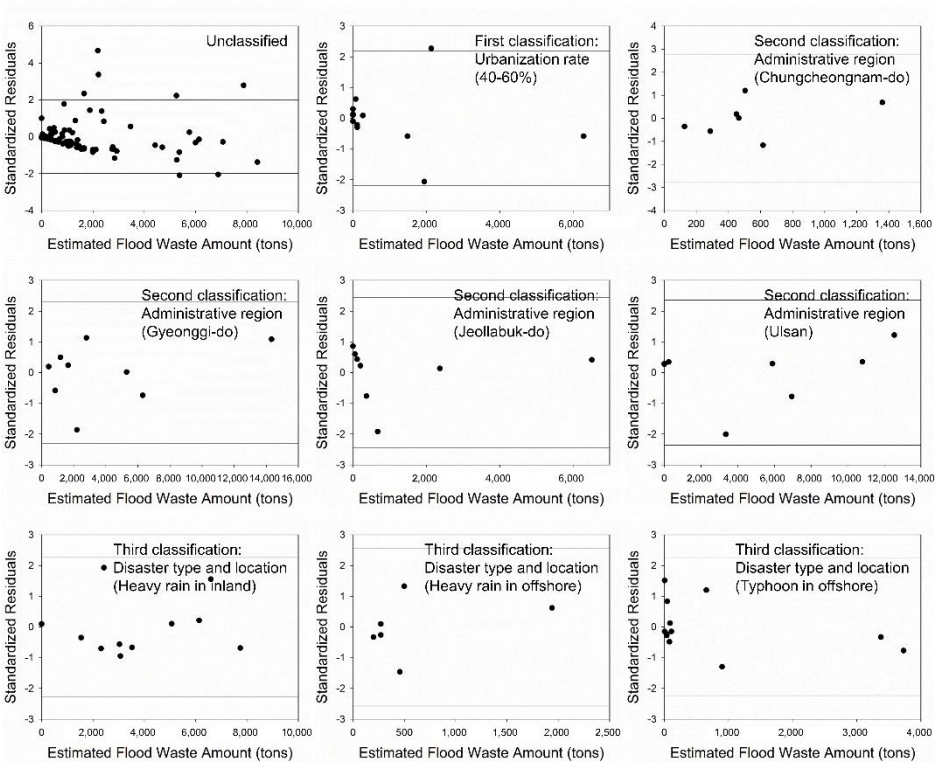

Figure 2.4. Residual plots of the regression analysis by the unclassified dataset and the best-fit grouping sequence (Solid line: 95% confidence interval calculated by *t*-distribution with relevant degree of freedom)

Table 2.10. Summary of Shapiro-Wilk test on the residuals according to the grouping

Grouping property	Grouping level	Weight	*p*-value
Without grouping	Without grouping	0.79152	$5.90 \times 10^{-10*}$
40-60%	1. Urbanization rate	0.87927	0.070
Chungcheongnam-do	2. Administrative region	0.99282	0.997
Gyeonggi-do	2. Administrative region	0.93648	0.545
Jeollabuk-do	2. Administrative region	0.81733	0.044**
Ulsan	2. Administrative region	0.81613	0.042**
Heavy rain in inland	3. Disaster type and location	0.82573	0.021**
Heavy rain in offshore	3. Disaster type and location	0.97898	0.946
Typhoon in inland	3. Disaster type and location	0.92699	0.349

*: cases with *p*-value < 0.01

**: cases with *p*-value < 0.05

After grouping, only one residual exceeded the 95% confidence interval, in case of first grouping by UR, and the lowered number of outliers suggest a better fit with the assumption of multilinear regression. The normality test for residuals showed the residuals passed normality test in most cases under $P = 0.05$ and in every cases with $P = 0.01$ (Table 2.10). Enhanced normality of the residuals supports the argument that grouping resolved heterogeneity in original dataset. In some groups, however, the residuals were not uniformly distributed, but biased to both left and right sides. This indicates the regression model did not fit to homoscedasticity assumption and resulted in a significant deficiency in waste generation prediction for the intermediate ranges as discussed in Section 2.3.3. However, due to the limited amount of accessible data, this issue cannot be currently resolved. Rather, it should be noted as a subject of future study when more disaster records and experiences are available. Or, other frameworks can be utilized later to resolve the limitation in linear regression, such as stochastic approach, nonlinear regression, etc.

2.3.5 Discussion

Because the number of data in each modeling reduced by grouping, one might raise a question on a mechanism of enhancement on model performance with grouping. More specifically, the grouping approach can be

criticized that it is not clear whether the increased linearity in each group was attributed to the consideration on the nature of disaster in the group or just decreased degree of freedom in linear regression. However, this argument was rejected by some cases showing large linearity with enough degree of freedom. For example, the model developed using the cases of flood from typhoon in offshore by the first grouping (Table 2.2) showed high linearity with 0.732 of adjusted r^2 while the degree of freedom (28) was large because the number of cases (30) was far larger than that of input varibles (2). Author admits that it is difficult to find out actual mechanism in enhanced modeling, but it should be noted that the better estimation was not solely explained by reduction of degree of freedom by a coincidence.

As some cases were not included in any group even after the third grouping, there will be some regions where a proper group and model cannot be found. This is an obvious limitation of grouping approach and should be resolved with further investigation. One interesting observation in remaining cases after third-stage grouping was that more than half of the cases (9 out of 16 cases) were recorded in one province (Jeollanam-do) in one event (Typhoon BOLAVEN in 2012). The regions of the events were mostly rural regions. It is not clear but can be suspected that regional characteristics of rural area, such as, inefficiency in regional governance, or unstandardized architecture have

increased inherent uncertainty in generation of flood waste. The uncertainty cannot be addressed by current framework of deterministic regression, and this should be addressed by other methodology such as stochastic regression or nonlinear regression (and this is clearly beyond the scope of this work).

2.3.6 Application procedure of suggested models

In non-disaster times, authorities responsible for disaster waste management are required to define groups for each region. According to Section 2.3.3, three-stage grouping yielded eight groups based on regional characteristics. For example, if a region's UR is included in the 40-60% range, the authority can utilize the model developed in the first grouping. If a region was not included in that group, it must be determined whether the region was included in any second stage grouping. In this way, authorities can define the appropriate group for each region, eventually selecting a proper model equation for flood waste prediction.

2.4 Summary

This study analyzed 90 flood waste records from South Korea to build a flood waste model with proper grouping. According to the results, data grouping ensured the advanced flood waste prediction, not only in single stage

groupings but also in successive data groupings. Since the three scenarios were examined, six sequences were analyzed in this study. Of the six grouping sequences studied, groupings in the following order were found to be most effective in terms of number selected and model accuracy: ‘urbanization rate’, ‘administrative region’, and ‘disaster type and ocean accessibility.’ In some cases, flood waste prediction within intermediate ranges was limited to the lack of currently accessible data. Although this study deals with the disaster data of South Korea, the methodology can be applied to other countries using case-specific grouping properties.

References

Brown, C., Milke, M., Seville, E., 2011. Disaster waste management: A review article. Waste management 31, 1085-1098.

Brown, L.C., Mac Berthouex, P., 2002. Statistics for environmental engineers. CRC press.

Chen, J.-R., Tsai, H.-Y., Hsu, P.-C., Shen, C.-C., 2007. Estimation of waste generation from floods. Waste management 27, 1717-1724.

Cho, Y.H., 2018. Estimation of flood debris generation using GIS-based inundation maps. The Graduate School, Seoul National University.

Crowley, J., 2017. A measurement of the effectiveness and efficiency of pre-disaster debris management plans. Waste Management 62, 262-273.

Dubey, B., Solo-Gabriele, H.M., Townsendt, T.G., 2007. Quantities of arsenic-treated wood in demolition debris generated by Hurricane Katrina. Environ Sci Technol 41, 1533-1536.

Easterling, D.R., Evans, J., Groisman, P.Y., Karl, T.R., Kunkel, K.E., Ambenje, P., 2000. Observed variability and trends in extreme climate events: a brief review. Bulletin of the American Meteorological Society 81, 417-426.

FEMA, 2012. Public assistance debris operations job aid.

Gabrielli, F., Amato, A., Balducci, S., Galluzzi, L.M., Beolchini, F., 2018. Disaster waste management in Italy: Analysis of recent case studies. Waste

management 71, 542-555.

Hirayama, N., Kawata, Y., 2005. Quantity of disaster waste for emergency response of public authirities on flood disaster. Institute of Social Safety Science 7, 325-330.

Hirayama, N., Kawata, Y., Suzuki, S., Harada, K., 2009. Estimation procedure for potential quantity of tsunami debris on tsunami earthquake disasters. Sardinia 2009, Twelfth International Waste Management and Landfill Symposium. S. Margherita di Pula, Cagliari, Italy: CISA Publisher, Italy.

Hirayama, N., Shimaoka, T., Fujiwara, T., Okayama, T., Kawata, Y., 2010. Establishment of disaster debris management based on quantitative estimation using natural hazard maps. Waste Manag Environ 140, 167-178.

Jeong, S., Kim, J., 2012. Comparison of disaster debris guidelines and analysis of flood debris recovery. Journal of Korea Society of Waste Management 29, 497-503.

Jo, J.H., Kim, T., 2016. Optimal Management of Disaster Waste by its Properties Using GIS. Korea Environment Institute.

Kang, E., Ju, M., Jeon, W. S., Kim, J., 2015. Flood Waste Prediction Method using Rainfall and Flooded Buildings in Seoul. Journal of Korea Society of Waste Management 32, 713-719.

MOE, 2017. Disaster Waste Safety Management Guidelines.

MOEJ, 2005. Flood disaster waste management guidelines.

NEMA, 2009-2018. Annual Report on Disaster.

Oh, G.-J., Kang, Y.-Y., 2013. The status of flood wastes treatment and future tasks in South Korea. Journal of Material Cycles and Waste Management 15, 282-289.

Petersen, M., 2004. Restoring waste management following disasters. International conference on post disaster reconstruction.

Pilapitiya, S., Vidanaarachchi, C., Yuen, S., 2006. Effects of the tsunami on waste management in Sri Lanka. Waste Manag 26, 107-109.

USEPA, 2008. Planning for natural disaster debris, Office of Solid Waste and Emergency Response and Office of Solid Waste

CHAPTER 3

BAYESIAN LINEAR REGRESSION IN ESTIMATING FLOOD WASTE GENERATION

3.1 Introduction

In South Korea, floods are the most frequent type of disaster as the country is surrounded by a monsoon climate region. According to the Korea National Emergency Management Agency, flood damage costs accounted for 89% of the country's total disaster damage costs during a recent decade (NEMA, 2009-2018). For example, typhoon Rusa resulted in damages worth 4.3 billion US dollars, which was the most severe economic loss that has occurred in South Korea since the advent of modern weather observations (NIDP, 2002); this figure was equivalent to 0.71% of the gross domestic product in that year (WorldBank, 2019). However, the stated costs only accounted for property damage. If the indirect costs of the disaster were incorporated, the total burden to society would be tremendous.

One important step in recovering from flood damage is disaster waste restoration. Inundation of facilities and private properties within a damaged region generates a large amount of flood waste. For example, Typhoon Chaba

in 2016 generated 26,846 total tons of waste from Ulsan city (Kim, 2018), which was equivalent to 17 days of municipal solid waste generation of the city in a peacetime setting (MOE, 2017). Outside of South Korea, Taipei city experienced a more severe case in 2001. Typhoon Nari generated 190,000 tons of waste, which was equivalent to 95 days of municipal solid waste in a peacetime setting (Chen et al., 2007). Disaster waste leads to blockages of roads; thus, an entire urban transportation system can be paralyzed during floods even when rescue activities are urgently needed. In addition, improperly managed waste matrices trigger secondary impacts on the ambient environment due to leakages of hazardous materials (Dubey et al., 2007; Pilapitiya et al., 2006) or the breeding of vermin and contagious disease vectors (Petersen, 2004). To reduce overall damage, waste clean-up should be rapidly and adequately executed.

A key goal in disaster waste management is successful quantity estimations of disaster waste. Precise and rapid quantification is beneficial during planning for and responses to disasters (USEPA, 2008); this is because the estimated waste quantity is critical design factors in the recovery process, such as the working period, budget, temporary storage site, required labor, and vehicles. In addition, the funding scale cannot be determined without information regarding waste quantity (FEMA, 2007). More accurate

estimations of waste quantity ensure transparent financial support and a quick decision-making ability for the central government, which are vital tools during "the golden hours" of disaster relief.

There have been many previous efforts to develop models estimating waste generation from floods. GIS/hazard map-based approaches have simulated flood damage and waste generation using inundation maps. Representative GIS-based models include a hurricane debris estimating model by the US Army Corps of Engineers and Hazus by FEMA (USEPA, 2008). Another approach is to estimate potential waste generation from existing material stocks (Tabata et al., 2016; Tsurumaki et al., 2015). The most successful framework in estimating flood waste is the multilinear regression approach, which has an advantage of simple operation. In Japan, a flood waste estimating model by regression using flood damage variables (i.e., the number of damaged buildings) as independent variables was introduced with fair credibility (Chen et al., 2007; Hirayama and Kawata, 2005). Chen et al. (2007) showed a successful estimation of flood waste by using a log-linear model comprised of population density, flooded area, and total rainfall. Thanks to the credibility and ease in the model development procedure, the regression method has been recognized as a promising analysis technique.

Until now, all the flood waste estimating models developed using

regressions were based on the least-squares method or the so-called deterministic approach. Each coefficient is estimated as a single point rather than a random variable, and thus the calculation process is simple. During a deterministic regression, however, considerable information entailed within a dataset is lost because of the compression of the probability distribution into a single point. The loss of information throughout modeling would therefore cause a low prediction accuracy. The weakness of the deterministic approach should be considered in flood waste predictions because all disaster data have a large degree of uncertainty and randomness that originates from the event's chaotic nature. Through this perspective, probabilistic regression can thus be considered an alternative to the conventional deterministic approach.

Bayesian inference is an alternative approach of deterministic statistics or the so-called frequentist approach. From the renowned idea of Bayes (1763), this approach considers the true value of a parameter as random variables. The probability distribution of the parameter is estimated by successive corrections of distribution from the *likelihood* of observed events under a designed *prior* belief (Bolstad and Curran, 2016). There were many trials of Bayesian modeling in the environmental engineering field, such as biochemical oxygen demand decay modeling (Borsuk and Stow, 2000), water quality modeling in a river water-body (Liu et al., 2008), and interpreting eco-

toxicity evaluation data (Billoir et al., 2008). However, Bayesian inference has never been applied in flood waste estimations despite the potential of this method to explain the uncertainty of disaster waste generation.

This study aimed to evaluate the applicability of the Bayesian inference method in flood waste estimations via multilinear regression. The detailed objectives of this study are as follows: (1) to analyze the properties of the estimated coefficient of the model; (2) evaluate the performance of Bayesian approaches in flood waste estimations; and (3) assess the effectiveness of Bayesian updating in flood waste estimations. The scope of this study was based on the available flood datasets in South Korea; however, we believe that the findings in this study will provide valuable information to other countries' disaster response regimes.

3.2 Materials and methods

3.2.1. Data collection

In addition to flood damage data collected in Chapter 3, other parameters related to meteorological and regional characteristics were explored. The disaster type whether flood was originated from heavy rain or typhoon, hourly maximum rainfall (HR, mm), total rainfall (TR, mm), and maximum

wind speed (WS, m/s) observed during the rainfall event were selected as variables to indicate the severity of the flood disaster. The meteorological data were obtained from the Korea Meteorological Agency (KMA, 2019b). The province, urbanization rate, closeness to ocean, whole area, gross regional domestic product, population density, area ratio of urbanized region, permeable land cover, impermeable land cover, and wastewater supply rate were utilized to represent the regional characteristics of disaster-receiving area. The data were collected from Korean Statistical Information System (KOSIS, 2018, 2019a, b).

3.2.2 Pre-event model

Upon this section, pre-event model was also developed along with the post-event model suggested in Section 3.2.1. An exponential model was utilized for a pre-disaster estimation of flood waste. This model was used by previous studies (Chen et al., 2007; Gabrielli et al., 2018) to estimate flood waste in Taiwan and Italy, respectively. Briefly, the model is expressed as the product of independent variables with relevant exponents and a constant. In those previous studies, variables were related to inherent regional characteristics (population density) and flood characteristics (flooded area and total rainfall). In this study, other parameters were incorporated to reflect the more detailed characteristics

of the flood-impacted region and flood characteristics. PD, GRDP, and UR were used as variables indicating the socio-economic status of a region. These parameters were associated with stocks in households, which can comprise potential disaster waste and is also related to waste generation in peacetime. The flooded area was excluded from the analysis because of the lack of accessible data as explained in Section 3.2.1. HR, TR, and WS were applied as variables representing flood hazard. The model structure was as follows:

$$y = a\mathrm{PD}^b \times \mathrm{GRDP}^c \times \mathrm{A}^d \times \mathrm{UR}^e \times \mathrm{HR}^f \times \mathrm{TR}^g \times \mathrm{WS}^h \quad (3.1)$$

where y is flood waste generation in tons, a is an estimated coefficient of the model, while b–h are estimated exponents for model parameters. Linear regression is not possible within this format. To develop the exponential model, a logarithmic transformation was performed for the following linear regression. The logarithmic form of the model is as follows:

$$\log y = a + b \log \mathrm{PD} + c \log \mathrm{GRDP} + d \log \mathrm{A} + e \log \mathrm{UR} + f \log \mathrm{HR} + g \log \mathrm{TR} + h \log \mathrm{WS} \quad (3.2)$$

3.2.3 Bayesian regression protocol

The Bayesian regression was performed using a Bayesian inference for the multiple linear regression function (bayes.lm) in the Bolstad package with the same developmental environment of deterministic regression. Two

prior distributions were applied to estimate the *posterior* distributions of coefficients: null *prior* (uniform distribution) representing no *prior* knowledge on distribution, and normal distribution with the mean and variance of coefficients obtained from the deterministic regression corresponding to the *prior* belief via the deterministic approach. Bayesian regression with updating was performed 100 times in total. The updating was performed by a successive estimation of the distribution using the obtained *posterior* distribution as the *prior* distribution of the next stage estimation.

3.2.4 Statistical examination

To evaluate the estimation performance, several statistical indices were examined. The statistical indices used in this study include the coefficient of correlation (r), root mean square error (RMSE, ton), mean absolute error (MAE, ton), median absolute error (MDAE, ton), and mean absolute percentage error (MAPE, %). These indices are calculated as follows:

$$r = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(f_i-\bar{f})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}\sqrt{(f_i-\bar{f})^2}} \tag{3.3}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i-f_i)^2}{n}} \tag{3.4}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|y_i-f_i|}{n} \tag{3.5}$$

$$\mathrm{MDAE} = median(|y_1 - f_1|, \cdots, |y_n - f_n|) \tag{3.6}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - f_i|}{y_i} \times 100 \quad (3.7)$$

where y_i is the measured value, $\bar{y}$ is the mean value of the measured values, f_i is the predicted value, $\bar{f}$ is the mean value of the predicted values, and n is sample size.

The r and RMSE are conventional indices for deterministic regression. Use of these two indices is occasionally inadequate in representing the model error; thus, other indices are complementarily used to evaluate the models' performance in many modeling studies, including regression, geostatistical modeling, and, recently, artificial intelligence-based machine learning processes (Azadi and Karimi-Jashni, 2016; Jahandideh et al., 2009; Jeong et al., 2019). Except for r, a lower error value stands for a higher model accuracy. These indices were calculated using the functions in the metrics package in R 3.6.0 under RStudio 1.2.1335.

3.2.5 Application to the other context

Evaluating the performance of the Bayesian approach in a flood waste estimation using the data from only one country may be insufficient. To fully assess the potential of the Bayesian approach, an application of the approach to the data from other countries is necessary. Analysis of the results obtained by Bayesian regression of datasets in other countries can provide an evaluation of

broader perspectives. In addition, one important aspect of disaster waste research is the spread of management strategies within international contexts (Brown et al., 2011). Applications of Bayesian approaches devised in this study to the published data of another countries may provide a key clue to sharing techniques in the disaster waste management field.

Due to a lack of related studies, accessible disaster waste datasets are limited. The datasets in Taiwan summarized by Chen et al. (2007) were used to investigate the potential of the framework suggested in this study. Chen et al. (2007) analyzed 39 cases of flood waste generation in 2001. The datasets were comprised of 41 cases, but two cases were excluded from this analysis because the authors declared that they were outliers. They suggested that the exponential model be formulated as follows:

$$\log y = a' + b' \log \mathrm{PD} + c' \log \mathrm{FA} + d' \log \mathrm{TR} \tag{3.8}$$

where a' is a coefficient of the model, b'–d' are the coefficients for the parameters, and FA stands for the flooded area (ha). Utilizing identical datasets and the structure of the model reported by the study, multilinear regressions were performed with deterministic and Bayesian approaches.

3.3 Results and discussion

3.3.1 Pre-event estimation with Bayesian inference

Estimated coefficients of the pre-event flood waste prediction model are summarized in Table 3.1. Except for *a*, there was no significant difference in the coefficient values depending on the estimation method. The least squares method and Bayesian inference yielded almost the same values for *b*–*h*. Although the locations of estimated values were similar, the distribution shapes showed significant differences. Bayesian inference yielded smaller standard errors for the coefficients compared to those calculated by deterministic regression. Unlike the deterministic regression, the Bayesian approach calculates a larger number of calculations for a coefficient estimation. The variance of each coefficient apparently became compressed as the Bayesian estimation was performed following repeated calculations depending on the case number. As a result of a smaller standard error, the *p*-value of each coefficient (calculated by *t*-distribution) become smaller and the estimation greatly increased in confidence.

Table 3.1. Estimated coefficients of the pre-event flood waste prediction model developed by 90 cases of flood waste generation in South Korea from 2008 to 2017

		a	*b*	*c*	*d*	*e*	*f*	*g*	*h*
Deterministic	Estimated value	-10.53	0.606	4.78×10^{-2}	0.505	0.297	-0.066	0.380	-0.385
	Standard error	6.62	0.288	0.376	0.269	0.358	0.384	0.348	0.371
	p-value	0.115	3.81×10^{-2}	0.899	6.37×10^{-2}	0.410	0.864	0.279	0.303
Bayesian with null *prior*	Estimated value	6.47	0.606	4.78×10^{-2}	0.505	0.297	-0.066	0.380	-0.385
	Standard error	2.63×10^{-2}	8.28×10^{-2}	0.141	7.23×10^{-2}	0.128	0.148	0.121	0.138
	p-value	2.05×10^{-119}	1.50×10^{-10}	0.736	6.70×10^{-10}	2.33×10^{-2}	0.656	2.40×10^{-3}	6.46×10^{-3}
Bayesian with deterministic *prior*	Estimated value	6.41	0.606	4.78×10^{-2}	0.505	0.297	-6.61×10^{-2}	0.380	-0.385
	Standard error	2.62×10^{-2}	6.10×10^{-2}	0.102	4.84×10^{-2}	8.64×10^{-2}	9.14×10^{-2}	7.51×10^{-2}	0.100
	p-value	3.48×10^{-119}	9.40×10^{-16}	0.639	9.60×10^{-17}	9.30×10^{-4}	0.472	2.55×10^{-6}	2.35×10^{-4}

Figure 3.1 shows the log-scale plots between estimated and actual waste generation and the residual plots for the pre-event prediction models. The graphical difference in plots in the figure depending on the estimation methods was negligible. Despite the difference in coefficient *a* as noted in Table 3.1, the estimated waste generations showed similar results. The contributions of other factors apparently overwhelmed the differences in coefficient *a*. Figure 3.1a-c shows that the linearity between the estimated and actual waste generation results is quietly low, regardless of estimation methods used. Residual analysis (Fig. 3.1d-f) shows that only two points are outliers in the pre-event flood waste prediction in all regression methods. There is no specific pattern in residual plots and the number of outliers is small. The regression results were fairly matched to the assumption of linear regression; however, this match might not have been caused by an accurate estimation but instead may have been caused by a large standard error of the residuals considering the low linearity in Fig 3.1a-c.

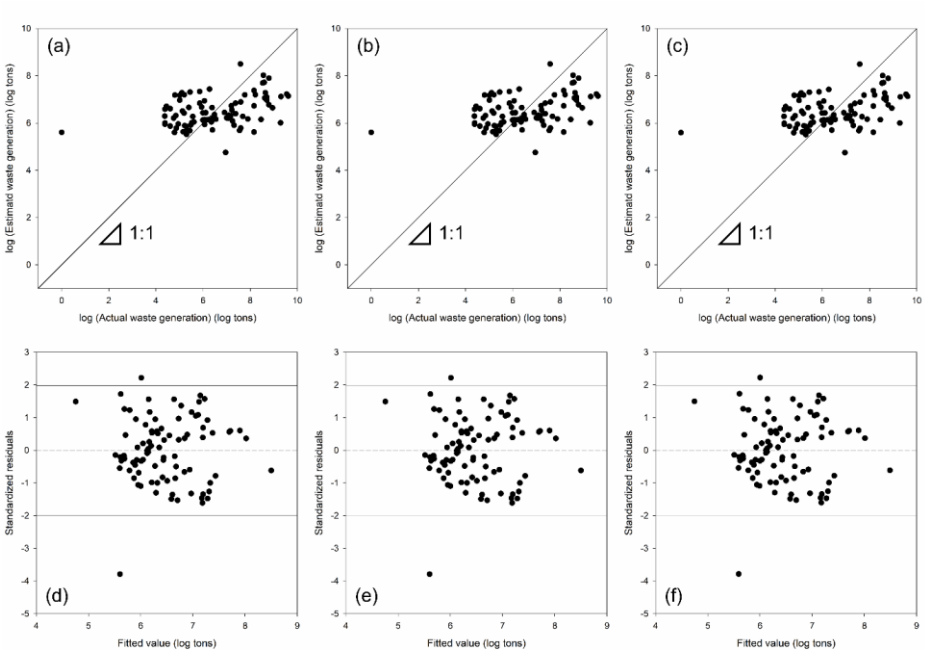

Figure 3.1. Plots of the pre-event flood waste prediction model: log-scale plots between actual and estimated waste generation via (a) deterministic regression, (b) Bayesian regression with a null *prior*, and (c) Bayesian regression with a *prior* comprised of values obtained from deterministic regression. Residual plots are shown for the estimated results of (d) deterministic regression, (e) Bayesian regression with a null *prior*, and (f) Bayesian regression with a *prior* comprised of values obtained from deterministic regression

To more accurately assess the performance of pre-event flood waste prediction models, five statistical indices were calculated, and the results are summarized in Table 3.2. The correlation coefficient is low; thus, the models

were seemingly inappropriate for practical application. All the indices showed similar magnitudes of errors for the three estimation methods. Figure 3.2 summarizes the results of 100 times Bayesian updates. Within the small extent of changes in y-axis values, Bayesian updating did not lead to dramatic changes in the pre-event model's performance regarding five statistical indices tested in this study. Changes in statistical indices were larger when the initial *prior* was comprised of values from the deterministic regression. As the updating process continued, the statistical indices converged to the same value regardless of initial *prior*. It is not clear whether the iteratively calculated model through updating ensures a more accurate performance because changes in the statistical indices showed inconsistent patterns, depending on the type of indices. For RMSE and MAE, the magnitude of error decreased during updating; however, in other terms, the error terms became larger.

Table 3.2. Statistical indices of the pre-event flood waste prediction model developed by 90 cases of flood waste generation in South Korea from 2008 to 2017

	r	RMSE	MAE	MDAE	MAPE
Deterministic	0.333	3,062.73	1,616.06	639.66	4.651
Bayesian with null *prior*	0.333	3,062.74	1,616.06	639.66	4.651
Bayesian with deterministic *prior*	0.333	3,066.50	1,616.58	635.93	4.603

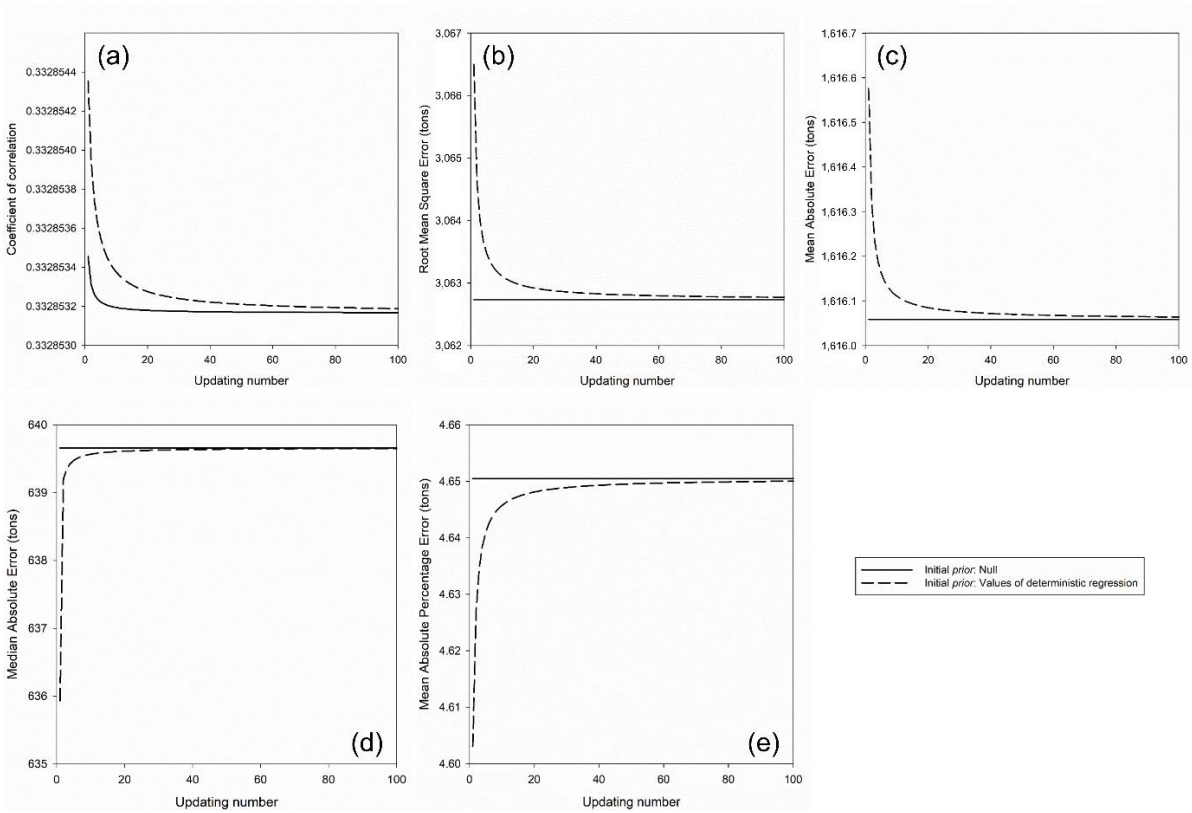

Figure 3.2. Changes in statistical indices of the pre-event flood waste prediction with Bayesian updating: (a) correlation coefficient, (b) root mean square error, (c) mean absolute error, (d) median absolute error, and (e) mean absolute percentage error

This study's analysis showed that neither model was preferable over the other for pre-event flood waste prediction modeling. This result might be associated with the large uncertainty in the pre-event flood waste prediction. Flood waste data contain a large degree of inherent uncertainty due to the complex nature of disaster events themselves. In addition, the parameters used for pre-event estimations are not directly linked to disaster waste generation. The parameters explaining social or geographical characteristics of a region can only address the potential waste generation in the region. Meteorological parameters cannot explain the actual level of flood damage, such as inundation depth or area. This weakness in the accuracy of the pre-event model was already noted by Jo and Kim (2016); however, pre-event modeling can enhance preparedness in a peacetime setting by providing design parameters for social infrastructure to reduce potential flood waste generation. Through acknowledging this limitation, both the deterministic and Bayesian approaches can be applied in peacetime. The accuracy of the pre-event model can be enhanced later if more experience is built through the suggested framework.

3.3.2 Post-event estimation with Bayesian inference

Table 3.3 summarizes the estimated coefficients of the post-event flood waste prediction deterministic and Bayesian regressions models. Unlike

the results in the pre-event estimation, the estimated coefficients of the post-event prediction models varied depending on the estimation methods and even depending on the *prior* used in the Bayesian approaches. The Bayesian approaches yielded more confidential values in terms of the *p*-value of the obtained coefficients. The reason for the larger significance of the coefficients in the Bayesian approach is, in summary, that its smaller standard error as compared to the error from the deterministic regression is associated with the number of calculations as noted in Section 3.3.1.

Table 3.3. Estimated coefficients of the post-event flood waste prediction model developed by 90 cases of flood waste generation in South Korea from 2008 to 2017

		a_1	a_2	a_3	a_4	a_5
Deterministic	Estimated value	3.68	1.34×10^{-4}	0.154	4.85×10^{-2}	6.45×10^{-2}
	Standard error	0.69	3.53×10^{-3}	0.128	3.14×10^{-2}	3.75×10^{-2}
	p-value	8.51×10^{-7}	0.970	0.232	0.126	8.89×10^{-2}
Bayesian with null *prior*	Estimated value	4.89	-7.95×10^{-5}	5.63×10^{-2}	4.29×10^{-2}	5.80×10^{-2}
	Standard error	0.467	1.50×10^{-5}	2.11×10^{-2}	1.17×10^{-3}	1.76×10^{-3}
	p-value	5.54×10^{-17}	8.79×10^{-7}	8.98×10^{-3}	6.49×10^{-54}	3.39×10^{-50}
Bayesian with deterministic *prior*	Estimated value	4.28	3.48×10^{-6}	0.113	4.39×10^{-2}	5.77×10^{-2}
	Standard error	0.234	6.80×10^{-6}	8.47×10^{-3}	4.94×10^{-4}	7.26×10^{-4}
	p-value	3.58×10^{-31}	0.610	1.76×10^{-22}	1.29×10^{-85}	1.50×10^{-81}

Figure 3.3 shows the plots of the post-event flood waste prediction models using different coefficient estimation methods. As shown in Figure 3.3a-c, slight differences in the estimated waste generation were visually identified. Obviously, these differences are related to the differences in the estimated values of the model coefficients. However, a preferable method could not be determined solely based on the plots' patterns. Figure 3.3d-f shows the residual plots of estimated values from tested estimation methods. Compared to the residual plots for the pre-event estimations, a larger number of outliers were observed. This is because of a smaller standard error of the residuals from the post-event regression, and any assertion that the pre-event model showed a superior performance is groundless. In the post-event estimation model, the Bayesian inferences yielded more outliers in the residual plots compared to those generated during the deterministic regression. However, drawing any further interpretations from a visual inspection of Figure 3.3 is difficult.

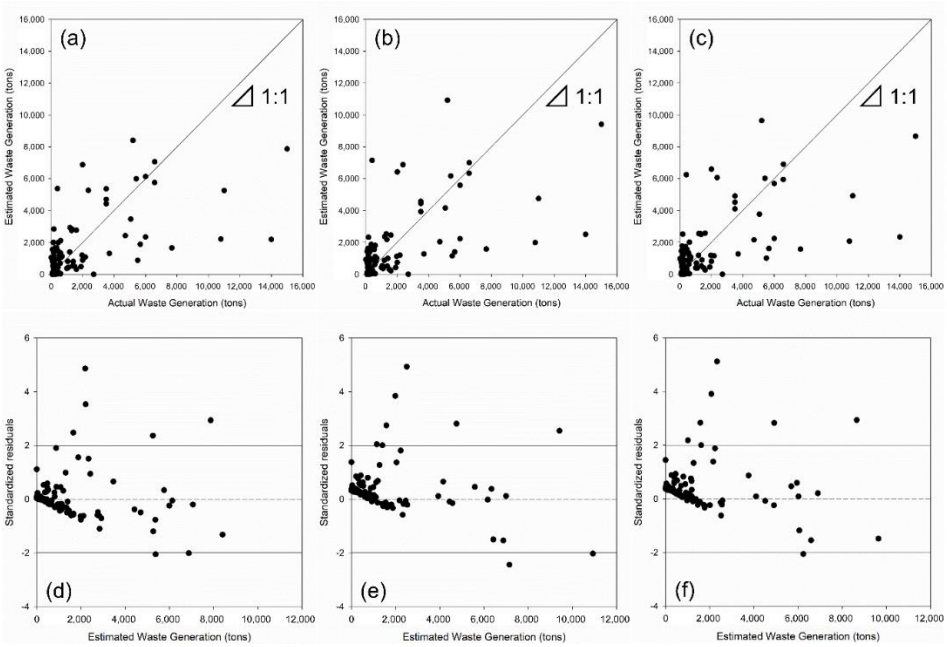

Figure 3.3. Plots of the post-event flood waste prediction models: plots between actual and estimated waste generation via (a) deterministic regression, (b) Bayesian regression with a null *prior*, and (c) Bayesian regression with a *prior* comprised of values obtained from deterministic regression. Residual plots are shown for the estimated results of (d) deterministic regression, (e) Bayesian regression with a null *prior*, and (f) Bayesian regression with a *prior* comprised of values obtained from deterministic regression

Similar to Section 3.3.1, five statistical indices were analyzed to evaluate the models' performances depending on the coefficient estimation methods (Table 3.4). Deterministic regression showed a slightly better

performance in terms of *r* and RMSE. However, there were few differences in these metrics. Considering other indices, Bayesian regression had the best performances and Bayesian regression with a null *prior* showed the smallest errors for MAE, MDAE, and MAPE. The relative differences in MAE, MDAE, and MAPE between deterministic regression and Bayesian regression using null *prior* were larger than those in *r* and RMSE. In summary, Bayesian regression showed slightly lower *r* and RMSE values, but this weakness was compensated for by error reductions in MAE, MDAE, and MAPE. This result demonstrates the potential of the Bayesian inference method for use in post-event flood waste estimation models. However, the improvement by Bayesian inference compared to conventional method was not remarkable, thus it can be criticized that Bayesian regression was a limited alternative for flood waste estimation.

Table 3.4. Statistical indices of the post-event flood waste prediction model developed from 90 cases of flood waste generation in South Korea from 2008 to 2017

	r	RMSE	MAE	MDAE	MAPE
Deterministic	0.600	2,423.86	1,416.48	755.24	13.63
Bayesian with null *prior*	0.592	2,472.76	1,388.58	641.23	12.10
Bayesian with deterministic *prior*	0.599	2,436.95	1,392.37	680.78	12.68

Figure 3.4 summarizes changes in statistical indices of post-event flood waste modeling with Bayesian updating. The changes in y-axis values were larger than those from the pre-event flood waste prediction with iterative updating (Fig. 3.2). Bayesian updating led to deprecations in *r* and RMSE but improvements in MAE, MDAE, and MAPE. Interestingly, a peak point was observed in MDAE (Fig. 3.4d) and the error reduction continued after the peak point. Similar to the pre-event prediction, the variations in indices were larger when the initial *prior* was made up of an estimated value from deterministic regression. In MAE, MDAE, and MAPE, Bayesian regression with a null *prior* always showed a low error magnitude.

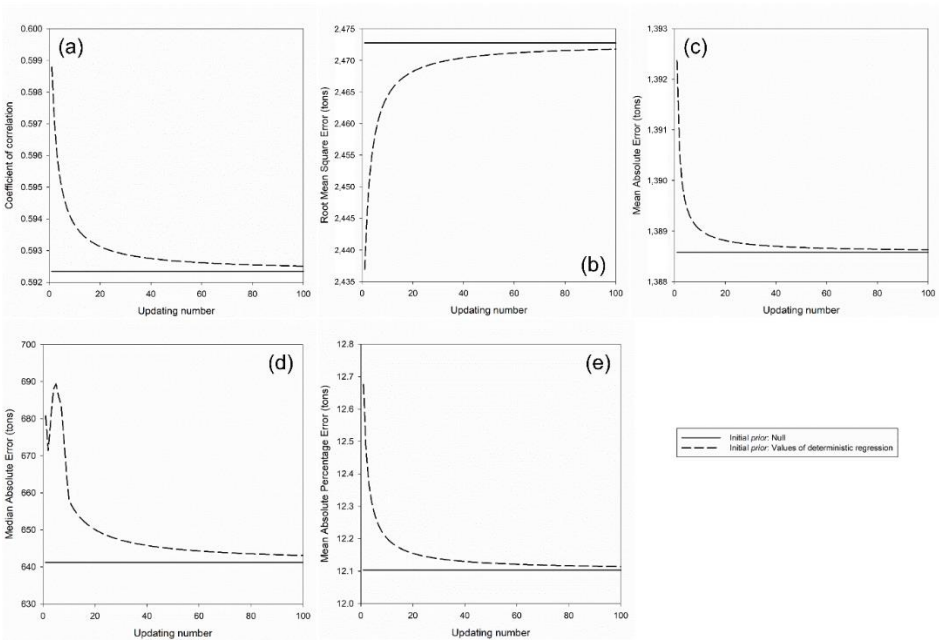

Fig. 3.4. Changes in the statistical indices of post-event flood waste predictions with Bayesian updating: (a) correlation coefficient, (b) root mean square error, (c) mean absolute error, (d) median absolute error, and (e) mean absolute percentage error

3.3.3 Application of Bayesian framework in other context

A pre-event estimation with a logarithm form equation was performed using a flood waste dataset from Taiwan. Different parameters were utilized because the suggested effective parameters were different within the publication provided by Chen et al. (2007). The estimated coefficients of the pre-event flood waste prediction models based on the estimation methods are summarized in Table 3.5. Similar to the results in Section 3.3.1., there was no significant difference in coefficient values except for *a'*, which is equivalent to *a* in Section 3.3.1. In addition, Bayesian regression showed a lower *p*-value for each coefficient, and this result is consistent with that in Section 3.3.1. As shown in Table 3.6, the performances of pre-event estimation models did not show significant differences between estimation methods. Bayesian regression with a *prior* comprised of values obtained from deterministic regression showed reduced error for RMSE, MDAE, and MAPE. For MAE, deterministic regression showed a slightly better performance. The results show similar

patterns as those found in Section 3.3.1. In summary, determining a preferable method for pre-event modeling with a logarithmic form equation is difficult.

Table 3.5. Estimated coefficients of pre-event flood waste prediction model developed by 39 cases of flood waste generation in Taiwan, 2001

		a'	*b'*	*c'*	*d'*
Deterministic	Estimated value	-9.65	0.78	0.54	1.43
	Standard error	1.59	0.23	1.22×10^{-2}	5.59×10^{-2}
	p-value	0.012	0.002	9.18×10^{-5}	0.015
Bayesian with null *prior*	Estimated value	7.21	0.78	0.54	1.43
	Standard error	6.85×10^{-2}	5.52×10^{-2}	1.49	0.31
	p-value	2.20×10^{-45}	4.56×10^{-16}	0.719	5.93×10^{-5}
Bayesian with deterministic *prior*	Estimated value	6.77	0.78	0.54	1.43
	Standard error	6.67×10^{-2}	2.63×10^{-2}	7.11×10^{-3}	0.16
	p-value	8.03×10^{-45}	2.09×10^{-26}	2.10×10^{-40}	7.10×10^{-11}

Table 3.6. Statistical indices of pre-event flood waste prediction model developed by 39 cases of flood waste generation in Taiwan, 2001

	r	RMSE	MAE	MDAE	MAPE
Deterministic	0.845	33,733.44	10,969.62	1,440.43	2.369
Bayesian with null *prior*	0.845	33,733.53	10,969.63	1,440.42	2.369
Bayesian with deterministic *prior*	0.845	37,434.85	11,702.53	1,289.32	1.529

This study had an obvious limitation in that the suggested methods were not applied to post-event estimations of flood waste generation outside of South Korea. This limitation originates from the lack of accessible data recorded by authorities or published in international journals. The spreading of Bayesian inference for post-event modeling in other contexts should be followed by increasing experience and international collaborations in data sharing.

3.4 Summary

In this chapter, the potential of Bayesian approaches for flood waste estimation was evaluated by using datasets from South Korea and Taiwan. The following key findings were drawn from the results: Bayesian inferences yielded slightly more credible values for model coefficients with a lower *p*-value. The differences in the obtained coefficients were more significant in the post-event prediction model compared to those in the pre-event prediction model. Bayesian approaches in the post-event prediction model led to a few error reduction in MAE, MDAE, and MAPE; however, the effect was truly limited in a narrow extent. Use of a null *prior* for Bayesian regression showed the most accurate performance in the post-event estimation. Bayesian updating showed the most dramatic changes in statistical indices when the initial *prior*

was comprised of the values obtained from deterministic regression. However, Bayesian regression without updating showed increasingly accurate results when the initial *prior* was a null *prior*; thus, iterative updating was not necessary to determine an ideal model. Consequently, Bayesian regression with a null *prior* without updating can show slightly more accurate post-event estimation of flood waste generation. Given the small enhancement of modeling by Bayesian inference, an approach beyond the parametric estimation method (e.g., deterministic versus stochastic) is needed for pre-estimation, such as a change in model structure to nonlinear regression.

References

Bayes, T., 1763. LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. Philosophical transactions of the Royal Society of London, 370-418.

Billoir, E., Delignette-Muller, M.L., Péry, A.R., Charles, S., 2008. A Bayesian approach to analyzing ecotoxicological data. Environ Sci Technol 42, 8978-8984.

Bolstad, W.M., Curran, J.M., 2016. Introduction to Bayesian statistics. John Wiley & Sons.

Borsuk, M.E., Stow, C.A., 2000. Bayesian parameter estimation in a mixed-order model of BOD decay. Water Research 34, 1830-1836.

Chen, J.-R., Tsai, H.-Y., Hsu, P.-C., Shen, C.-C., 2007. Estimation of waste generation from floods. Waste management 27, 1717-1724.

Dubey, B., Solo-Gabriele, H.M., Townsendt, T.G., 2007. Quantities of arsenic-treated wood in demolition debris generated by Hurricane Katrina. Environ Sci Technol 41, 1533-1536.

FEMA, 2007. Debris Management Guide. Public Assistance.

Hirayama, N., Kawata, Y., 2005. Quantity of disaster waste for emergency response of public authirities on flood disaster. Institute of Social Safety

Science 7, 325-330.

Kim, H.J., 2018. Efficient measures for the treatment of flood waste in Ulsan metropolitan area. Research Report of Ulsan Development Institute.

Liu, Y., Yang, P., Hu, C., Guo, H., 2008. Water quality modeling for load reduction under uncertainty: a Bayesian approach. Water Research 42, 3305-3314.

MOE, 2017. National waste generation and disposal status (2016).

NEMA, 2009-2018. Annual Report on Disaster.

NIDP, 2002. The Field Survey Report of Damages Caused by the Typhoon RUSA in 2002 - 8.30 ~ 9.1.

Petersen, M., 2004. Restoring waste management following disasters. International conference on post disaster reconstruction.

Pilapitiya, S., Vidanaarachchi, C., Yuen, S., 2006. Effects of the tsunami on waste management in Sri Lanka. Waste Manag 26, 107-109.

Tabata, T., Zhang, O., Yamanaka, Y., Tsai, P., 2016. Estimating potential disaster waste generation for pre-disaster waste management. Clean Technologies and Environmental Policy 18, 1735-1744.

Tsurumaki, M., Yamamoto, Y., Yoshida, N., 2015. A basic study on prediction method of disaster waste based on regional material stock. Journal of Japan Society of Civil Engineers, Ser. G (Environmental Research) 71.

USEPA, 2008. Planning for natural disaster debris, Office of Solid Waste and Emergency Response and Office of Solid Waste

WorldBank, 2019. GDP (current US$), Data - World Bank Open Data - World Bank Group, https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=KR (Access date: 2019.7. 19.)

CHAPTER 4

INCORPORATING INTERACTION TERMS IN LINEAR REGRESSION FOR FLOOD WASTE ESTIMATION

4.1 Introduction

Flood is a disaster that outbreaks when extreme rainfall exceeds the bearing capacity of the urban system. The flood damage is accompanied by various types of losses, and flood waste is one important issue among them. The generation of flood waste usually exceeds the amount that a city can afford, left in the urban area without proper management. Stocked waste blocks the route for recovery activity and induces secondary pollution problems such as leakage of contaminants to surrounding environment (Dubey et al., 2007; Pilapitiya et al., 2006). A slight delay in rescue activity can provoke tremendous additional damage in disaster time. Therefore, rapid collection and disposal of flood waste is a critical path in disaster recovery.

Estimating the quantity of flood waste is the first step to facilitate flood waste disposal. Rapid and precise estimation provide important design factors

for planning and implementation of the whole process of waste recovery (USEPA, 2008). The well-known method of predicting the flood waste amount is a multivariate linear regression of flood damage variables. In Japan, Hirayama and Kawata (2005) have demonstrated the successful application of this method using damaged buildings as input variables. In detail, they classified the damaged buildings upon the six types of damage and estimated the coefficients for these six input variables. The more sophisticated model was suggested by US EPA in the web-based model engine named incident waste decision support tool (I-WASTE) (Thorneloe et al., 2007; USEPA, 2013). This model accepted the more various input variables such as the number of damaged hospitals, elementary schools, etc.

The problem is, these methods developed in the countries cannot be applied in the other country because of difference in disaster characteristics and difficulty in data collection correspondent to the models. Referring to the examples in other countries, and the results of own investigation, many trials have been made to build a relevant model to estimate flood waste quantity. Korea Ministry of Environment recommends predicting flood waste generation using 1.7 ton/damaged buildings as a unit waste generation value per damaged building (MOE, 2017). However, many domestic studies in South Korea revealed that use of the unit value cannot ensure the flood waste estimation in

acceptable ranges (Cho, 2018; Jo and Kim, 2016; Kang et al., 2015; Park et al., 2018). In addition, previous studies by our research group showed that the model developed by conventional multilinear regression from own dataset is limited in its accuracy (Park et al., 2019a, b; Park and Kim, 2019). Furthermore, the use of detailed input variables as suggested in EPA method was impossible in Korea due to lack of records matching to the input variables.

Incorporating interaction terms in multivariate linear regression may be effective to enhance the model performance when input variables share co-linearity (Cortina, 1993; Jaccard et al., 1990). The idea of the interaction term is simple; just adding the term comprised of the product of two independent variables. The interaction term can adjust the over/underestimated contribution of parameters and elucidate the synergetic effect of two variables on the output response. Despite the potential of consideration of the interaction term, there has been no research regarding the estimation of flood waste generation using the interaction term.

This study aims to propose a novel approach in developing a regression model to estimate the quantity of flood waste adopting interaction terms. The first specific objective is to evaluate the effectiveness of incorporating interaction terms to enhance the accuracy of the flood waste regression model. After the credibility of the model including interaction terms

was investigated, the next objective is to assess the role and contribution of interaction terms in modeling and this will be supported by the sensitivity analysis. Synthesizing the results obtained from those analyses, the final goal is to suggest comprehensive knowledge of the patterns in flood waste generation behind the modeling itself. In order to find out the convincing rationale, both statistical examination and field surveying were carried out and the detailed protocols were addressed in the methodology section. The scope of this study is bound to South Korea cases of which flood waste data were accessible.

4.2 Materials and methods

4.2.1 Data collection

The data used in this chapter were similar with those in Chpater 2. The only difference was that in this chapter damaged buildings were classified depending on damage level. Briefily, flood damage data include the number of damaged buildings, area of damaged cropland (C) in ha, length of damaged road (R), damaged river (R'), and small stream (S) in m. The use of these parameters was attempted in previous flood waste modeling researches in the country (Cho, 2018; Park et al., 2018). The damaged buildings were further

classified by three subsets; completely destructed buildings (CB), partially destructed buildings (PB), and flooded buildings (FB) following the idea in Japanese case (Hirayama and Kawata, 2005).

4.2.2 Multivariate linear regression

Multilinear regression was performed using an lm function in the stats package of RStudio 3.6.2. Regressions were performed in two-step: 1) regression without interaction terms, and 2) regression with interaction terms. The structure of the former model is as follows:

$$Y = I + \sum_{i=1}^{n} a_i X_i \tag{4.1}$$

where Y (output variable) is the flood waste generation in tons, I is the intercept, X_i (input variable) is the flood damage variable, and a_i is the coefficient for X_i. The structure of the latter model is as follows:

$$Y = I + \sum_{i=1}^{n} a_i X_i + a_{ij} X_i X_j \tag{4.2}$$

where a_{ij} is the coefficients for interaction terms between X_i and X_j. The use of intercept term aims to respond to unexpected waste generation exterior from the input variables.

After regression analysis, statistics of estimated coefficients and estimated waste generations were assessed. In order to confirm the significance of independent variables p-value of each coefficient was estimated under the

assumption of Student's *t*-distribution. The credibility of predicted waste generations was evaluated using the coefficient of determination (r^2), adjusted r^2, and root mean square error (RMSE). The aim of using the adjusted r^2 is to distinguish whether the increase in r^2 is an effect of improved model accuracy or accidental increase by increasing the number of variables.

In addition to model credibility evaluation, the residual analysis was carried out to examine the fitness of the model for assumption in linear regression. Since deterministic linear regression postulates identical, independently distributed (*i.i.d*) normal distribution of error, residuals should fit the normal distribution and homoscedasticity.(Rencher and Christensen, 2012) The *i.i.d* patterns of residuals were examined by investigating standardized residuals, which are calculated as follows:

$$\hat{e}_{st,i} = \frac{\hat{e}_i}{standard\ error} = \frac{y_i - \hat{y}_i}{standard\ error} \tag{4.3}$$

where $\hat{e}_{st,i}$ is a standardized residual, $\hat{e}_i$ is a residual error, y_i is an actual waste generation, and $\hat{y}_i$ is estimated waste generation for i^{th} observation, respectively. Standardized residuals beyond the 95% confidence intervals were regarded as outliers that did not follow the assumption of normal distribution (Rencher and Christensen, 2012). The threshold of 95% interval value was estimated under the *t*-distribution with the degree of freedom defined by numbers of events and input variables. In addition, normality and

homoscedasticity of the residuals were examined by Shapiro-Wilk normality test (Shaphiro and Wilk, 1965) and score test for non-constant error variance (Kabacoff, 2011), respectively. The former and latter test was conducted by shaprio.test function in stats package and ncv.test function in the car package of R ver.3.6.2.

4.2.3 Variable selection

In order to develop a more efficient model, the variable selection was processed. It is beneficial to use a model with a simpler structure if the model can show acceptable performance without the introduction of other parameters. We adopted a mixed-stepwise variable selection method as this method is conceived to compensate for the limitation of both forward selection and backward elimination procedures. The step function in the stats package of R ver.3.6.2 was applied to carry out the computation. The function selected variables depending on the response of Akaike's information criteria.

4.2.4 Sensitivity analysis

Sensitivity analysis can provide information on how the output of the multi-dimensional model responds to a combination of input variables. The sensitivity analysis was conducted for two types of input variables because the

analysis for a single variable will generate just simple linear response regardless of the inclusion of interaction terms, and the test for more than two variables cannot be depicted in a way human can understand because the response will be represented in more than three-dimension. Two types of the most significant input parameters were selected according to the results of linear regression, and the significance was determined by the *p*-value of the estimated coefficients for the variables. The ranges of sensitivity analysis were determined by 95% confidence interval in the probability distribution of observed frequency of input variables. The probability distribution was estimated by the kernel estimation method which was computed by density function in the stats package of R ver.3.6.2. The designed range was divided by 20 segments (21 points in one axis) for both two-axis for the variables and the response of the model was computed in 21 × 21 grid. While the input value of two variables moved in the grid, other input variables maintained at their mean values, respectively.

4.2.5 Variation inflation factor calculation

The variance inflation factor (VIF) for each parameter was calculated to explain inter-dependency between parameters. The calculation is as follows:

$$VIF_i = \frac{1}{1-r_i^2} \quad (4)$$

where VIF_i stands for VIF of independent variable X_i, and r_i^2 is r^2 value for multilinear regression to fit X_i using the input variables excluding X_i. For example, VIF for CB is calculated by r_{CB}^2, computed from multilinear regression to fit CB using other variables including PB, FB, C, R, R' and S. It should be noted that waste generation data were not incorporated in calculation of VIF, as the factor shows only the inter-relationship between independent variables not including target variable. Generally, VIF closer to 1.0 indicates no inter-dependency of independent variable X_i and others, and VIF greater than 10.0 indicates a strong collinearity between X_i and others (Neter et al., 1996).

4.2.6 Field surveying

In order to figure out the nature of flood waste generation and the contribution of flood damage variables, the authors carried out a field survey on damaged regions stricken by the typhoon Danas which was active in 16th-21st, July 2019. After the typhoon attacked southern regions in South Korea, the authors collected news releases regarding the disaster information focusing on disaster damages analyzed as input variables in this study. In South Korea, every urgent news about disaster events was gathered on a website of the National Disaster Safety Portal, which is managed by the Ministry of the Interior and Safety (MOIS) of South Korea (MOIS, 2019). The well-organized

information system facilitated the planning for the field survey, for example, when and where to visit.

In this field survey, the authors visited rural area in Yeosu city, Jeollanam-do (34°46'19”N, 127°35'17’’E) where inundation of cropland was reported (YonhapNewsAgency, 2019b), and a residential town in Busan city (35°05'39"N, 129°02'53"E) where complete destruction of a house was reported (YonhapNewsAgency, 2019a). An unmanned aerial vehicle (Phantom 4 Pro, SZ DJI Technology Co., Ltd, People's Republic of China) was utilized to take a picture from a location and angle hard to access.

4.3 Results and discussion

4.3.1 Linear regression without interaction terms

Table 4.1 summarizes estimated coefficients of the flood waste regression model when the interaction term was not considered. It seems CB and FB were the most significant variables to estimate flood waste generation considering low *p*-value. Many of insignificant input variables in the model built without variable selection process (Table 4.2) were excluded by variable selection and that excluded were C, R’, and S. The coefficient for CB in this study is much larger than the unit waste generation per CB in Japan cases, which was 12.9 ton/building.(Hirayama and Kawata, 2005) The coefficient

value for PB was negative value and this might be because the negative fraction of PB attenuated the overestimation of waste generation by other factors.

Table 4.1. Estimated coefficients of flood waste regression model without interaction term developed by using 90 cases of flood waste generation in South Korea from 2008 to 2017

	Intercept	CB	PB	FB	R
Coefficient	616.97	149.97	-15.75	3.13	0.14
Standard Error	351.27	49.04	7.67	0.71	0.14
t-statistics	1.76	3.06	-2.05	4.38	1.01
p-value	0.08	2.98×10^{-3}	0.04	3.27×10^{-4}	0.02

Table 4.2. Estimated coefficients of flood waste regression model without interaction term developed by using 90 cases of flood waste generation in South Korea from 2008 to 2017

	Intercept	CB	PB	FB	C	R	R'	S
Coefficient	488.79	125.85	-10.38	3.13	0.00	0.14	0.03	0.03
Standard Error	377.00	52.81	8.81	0.71	0.00	0.14	0.03	0.04
t-statistics	1.30	2.38	-1.18	4.38	-0.26	1.01	0.93	0.75
p-value	0.20	0.02	0.24	0.00	0.80	0.32	0.35	0.46

Figure 4.1a shows a plot of actual and estimated waste generation by the model without the interaction term. The modeling result by the model without variable selection is depicted in Figure 4.2. As seen in Figure 1a, there is a large degree of discrepancy between actual and estimated waste generation. Especially, the model was not able to predict waste generation larger than 10 kton. According to the statistical analysis, RMSE of the model was 2,355.4 ton revealing a large error in estimation. The r^2 and adjusted r^2 values were 0.39 and 0.36, respectively. Even though there has been no model to predict flood waste in South Korea, the application of this model seems not appropriate considering the poor performance of the model.

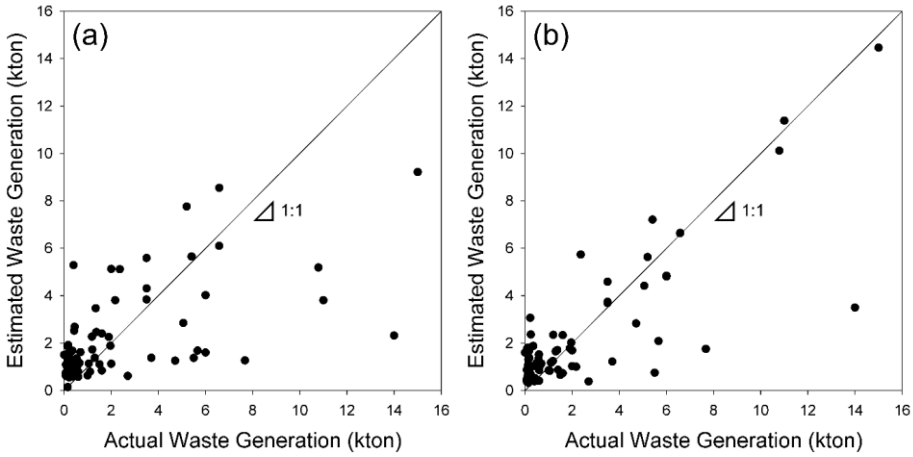

Figure 4.1. Plots between actual and estimated waste generation by regression model developed by using 90 cases of flood waste generation in South Korea

from 2008 to 2017: (a) model without interaction term, and (b) model with interaction terms

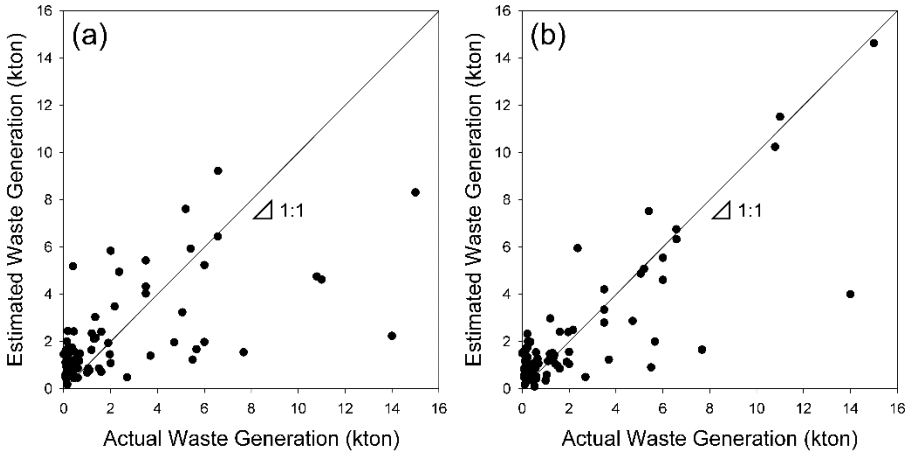

Figure 4.2. Plots between actual and estimated waste generation by regression model developed without variable selection process using 90 cases of flood waste generation in South Korea from 2008 to 2017: (a) model without interaction term, and (b) model with interaction terms

As seen in Figure 4.3a, there was no specific pattern in residual plots of the model without interaction term, thus seemingly the model quietly fitted to the assumption of linear regression. The patterns of residuals were not far different in the model developed without variable selection process (Figure 4.4a). Outliers exceeding confidence interval, however, were observed. The

number of outliers showing standardized residuals larger than 1.99 (95% confidential interval criteria calculated by 85 degrees of freedom) five and that lower than -1.99 was one. Obviously, the number of outliers is larger than the expected number of outliers under normal distribution, which is 5% of 90 samples, thus, 4.5 counts. However, it is not clear that the reason for the occurrence of many outliers is whether due to the limitation of the model formula or due to inevitable random error.

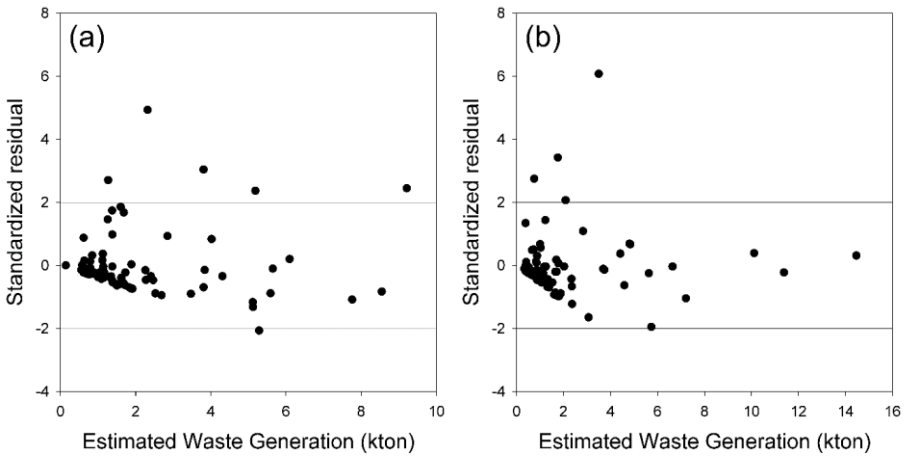

Figure 4.3. Standardized residual plots of regression models developed by using 90 cases of flood waste generation in South Korea from 2008 to 2017: (a) model without interaction term, and (b) model with interaction terms

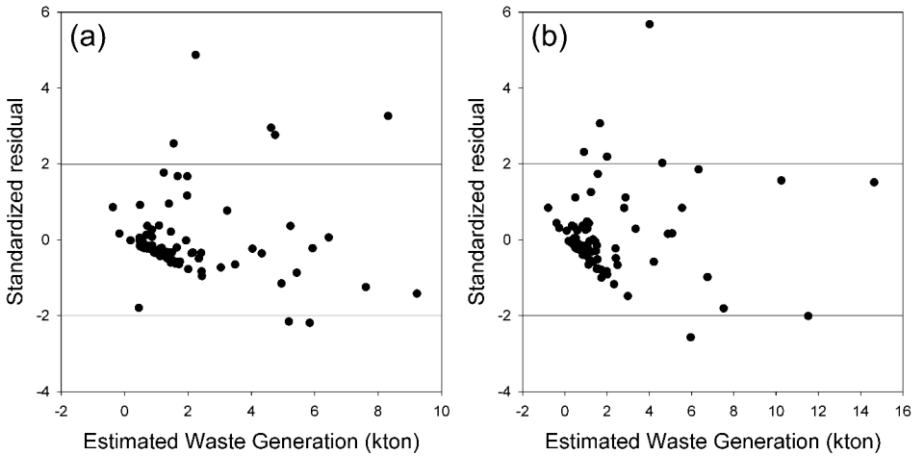

Figure 4.4. Standardized residual plots of regression models developed without variable selection process by using 90 cases of flood waste generation in South Korea from 2008 to 2017: (a) model without interaction term, and (b) model with interaction terms

The pattern of the residuals was further investigated by normality and homoscedasticity examination. As a result, *p*-values for both the Shapiro-Wilk normality test and score test for non-constant error variance were far less than 0.05, indicating the modeling result did not conform to the assumption of linear regression. In detail, *p*-values for the Shapiro-Wilk test and score test for non-constant error variance were 9.374×10^{-11}, and 4.435×10^{-6}.

4.3.2 Linear regression with interaction terms

Estimated coefficients for the model with the interaction terms were summarized in Table 4.3. The model coeffiencints without variable selection are summarized in Table 4.4. Not every coefficient was significant in terms of the *p*-value. Among the single parameter terms, R' showed the lowest *p*-value and this result is quite different from that in the aforementioned section. Interestingly, many of interaction terms showed lower *p*-value than 0.05 (CB × FB, CB × C, FB × R, FB × S, C × R, C × R', and R × R'), implying incorporating interaction terms for regression has a positive impact on flood waste modeling.

Table 4.3. Estimated coefficients of flood waste regression model considering interaction term developed by using 90 cases of flood waste generation in South Korea from 2008 to 2017

	Intercept	CB	HB	FB	C	R	R'
Coefficient	381.60	106.50	-2.70	1.01	-2.16	0.05	0.19
Standard Error	368.10	67.23	8.19	0.97	2.64	0.15	0.09
t-statistics	1.04	1.58	-0.33	1.05	-0.82	0.36	2.12
p-value	0.30	0.12	0.74	0.30	0.42	0.72	0.04
	S	CB × FB	CB × C	HB × FB	FB × C	FB × R	FB × R'
Coefficient	-3.83×10^{-3}	-1.07	2.35	-0.22	0.05	3.62×10^{-3}	4.79×10^{-4}
Standard Error	0.08	0.28	0.69	0.16	0.03	6.67×10^{-4}	2.47×10^{-4}
t-statistics	-0.05	-3.87	3.43	-1.35	1.64	5.43	1.93
p-value	0.96	2.38×10^{-4}	1.00×10^{-3}	0.18	0.11	7.43×10^{-7}	0.06
	FB × S	C × R	C × R'	R × R'	R × S		

Coefficient	-9.30×10^{-4}	-6.33×10^{-3}	1.88×10^{-3}	-1.06×10^{-4}	4.63×10^{-5}
Standard Error	4.37×10^{-4}	2.34×10^{-3}	3.99×10^{-4}	2.37×10^{-5}	2.73×10^{-5}
t-statistics	-2.13	-2.71	4.71	-4.46	1.70
p-value	3.66×10^{-2}	8.55×10^{-3}	1.19×10^{-5}	2.97×10^{-5}	0.09

Table 4.4 Estimated coefficients of flood waste regression model considering interaction term developed by using 90 cases of flood waste generation in South Korea from 2008 to 2017

	Intercept	CB	PB	FB	C	R	R'	S	CB × PB	CB × FB
Coefficient	497.33	11.82	-2.48	1.11	-3.80	0.22	0.16	-0.02	0.83	-1.17
Standard Error	443.05	135.89	19.33	1.04	4.63	0.24	0.12	0.11	1.17	0.59
t-statistics	1.12	0.09	-0.13	1.07	-0.82	0.89	1.31	-0.19	0.71	-2.00
p-value	0.27	0.93	0.90	0.29	0.42	0.38	0.20	0.85	0.48	0.05
	CB × C	CB × R	CB × R'	CB × S	PB × FB	PB × C	PB × R	PB × R'	PB × S	FB × C
Coefficient	2.93	0.02	0.00	0.00	-0.22	-0.22	0.00	0.01	-0.01	0.09
Standard Error	1.90	0.06	0.01	0.03	0.18	2.08	0.00	0.01	0.01	0.09
t-statistics	1.54	0.37	-0.35	0.04	-1.17	-0.10	-0.79	0.38	-0.76	0.98
p-value	0.13	0.71	0.72	0.97	0.25	0.92	0.43	0.71	0.45	0.33
	FB × R	FB × R'	FB × S	C × R	C × R'	C × S	R × R'	R × S	R' × S	
Coefficient	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Standard Error	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
t-statistics	3.16	1.68	-1.59	-1.22	2.82	-0.56	-3.73	1.29	0.81	
p-value	0.00	0.10	0.12	0.23	0.01	0.57	0.00	0.20	0.42	

Figure 4.1b shows enhanced linearity in a plot between actual and estimated waste generation compared to that of Figure 4.1a. This result is similar in the models developed without variable selection process (see Figure 4.2a and b). The model with the interaction terms succeeded to estimate the cases of huge amounts of waste generation (>10 kton), which was impossible for the conventional model without the interaction term. The increase in linearity was confirmed by higher r^2 and lowered RMSE. The r^2 value was 0.67, which is significantly larger than that in the aforementioned section. One can criticize that the increase in r^2 might be not due to improved model accuracy but due to an increased number of input variables (Theil, 1961). In order to thoroughly investigate the performance of the model, adjusted r^2 was also calculated and the value was 0.59, which is larger than both r^2 and adjusted r^2 of the model without the interaction term. In addition, the RMSE value was 1,719.0, which is lower than that of the model without the interaction term.

The residual plot of the model with interaction terms is shown in Figure 4.3b. The number of outliers in the model is slightly lower than that of the model without the interaction term. Specifically, the number of outliers exceeding 2.00 times of standard error (95% confidential interval criteria calculated by 71 degrees of freedom) was four and those lower than -2.00 times of standard error was zero, thus the total number of outliers was smaller than

expected outliers, again 4.5 counts. However, it should be noted that a point showing the largest discrepancy closed to 6.0 times of standard error was observed in the model. This result is contradictory to the patterns of residuals in the model without variable selection (Figure 4.4b), as the model showed larger number of outliers when variable selection was not conducted. Considering the large error point in the residual plot, therefore, caution is needed for the practical application of the model with interaction terms. Despite the enhanced credibility of the model in a viewpoint of deterministic statistics value, including RMSE, r^2, and adjusted r^2, the model can entail some degree of uncertainty. In order to avoid the problem due to ambiguity, consideration of safety factor is needed for practical application.

According to the statistical examination on the residuals, *p*-values for normality and homoscedasticity test were 5.339×10^{-12}, and 0.040. The *p*-values were still lower than 0.05, thus it can be argued that the linear model, in fact, did not fit the dataset in terms of the model assumption whether it includes interaction terms or not. This observation suggests that the linear form model contains inherent limitations in explaining the phenomena of flood waste generation, despite a partial adjustment to consider nonlinear interaction between two variables. The limitation of the model illuminate the next journey to the perfectly nonlinear modeling (and this subject is clearly beyond the scope

of this chapter). Even though the limitation of the revised linear model was acknowledged, it is doubtless clear that incorporating the interaction terms led to advanced flood waste prediction in the scope of linear modeling.

4.3.3 Roles and contributions of interaction terms

The CB and FB were utilized as two representative variables for sensitivity analysis because these two showed relatively small *p*-values than others in the model without the interaction term (Table 4.1) and the interaction terms including those were far significant in the model developed with interaction terms (Table 4.3). A similar trend was observed in the models established without the variable selection process as seen in Table 4.2 and 4.4. This result is far different from the analysis in Japanese cases conducted by Hirayama and Kawata (2005). The Japanese cases showed the parameters correspondent to PB (damaged building more than half of its original structure) in this study were more significant rather than CB or FB. Of course, the utilized input variables in Japanese cases did not match perfectly with our study, however, the distinct difference in waste generation patterns between two countries suggested the needs of regional individual studies.

The response of estimated waste generation in the area made up of a combination of 95% confidence intervals of the two variables is depicted in

Figure 4.5. The 95% confidential interval of each variable was determined from the estimated probability distribution of observed cases and the distribution is shown in Figure 4.6. As shown in Figure 4.5a, the estimated range by the model without interaction term is limited in < 100,000-ton flood waste generation. Figure 4.5b shows that the model with interaction terms can explain a larger amount of waste generation than that by the aforementioned model. This is correspondent to the modeling results in Figure 4.1, as the model with interaction terms succeeded to estimate waste generation more than 10,000 tons. In other words, the interaction term contributed to explaining unexpected large waste generation.

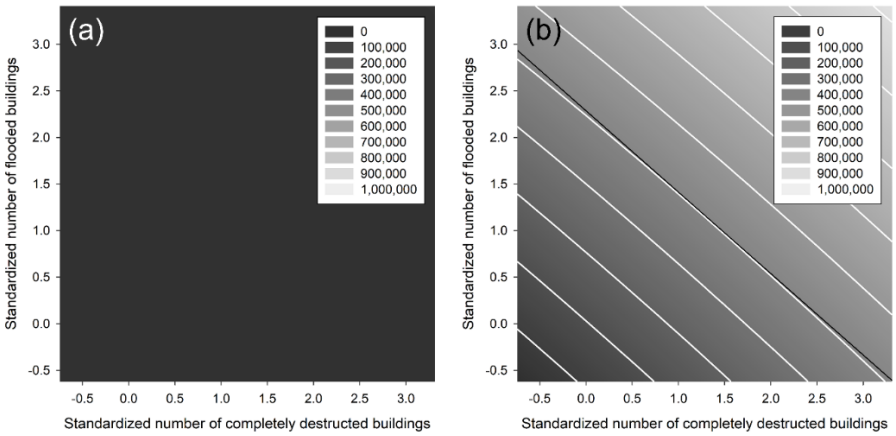

Figure 4.5. Response of waste estimation according to sensitivity analysis under the 99% confidential intervals of the number of completely destructed

building and the number of flooded buildings: (a) model developed without interaction terms; (b) model developed with interaction terms (white solid lines: gradient curves of output responses; black solid line: a tangent line to one gradient curve)

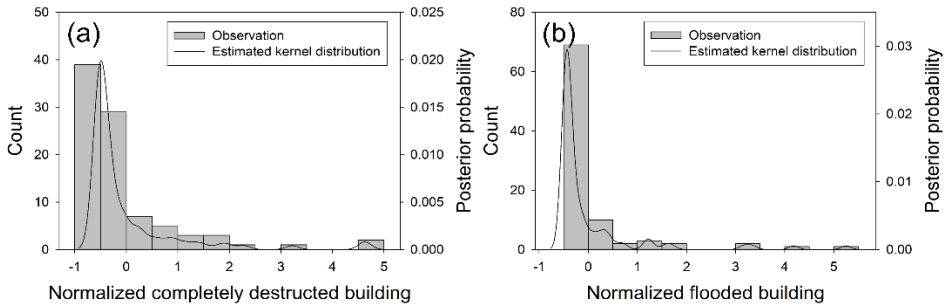

Figure 4.6. Estimated probability distributions of (a) the number of completely destructed building and (b) the number of flooded buildings

In Figure 4.5b, a contour line of output response did not correspond to the tangent line of the curve. This observation revealed that the response by the interaction-included model was obviously nonlinear. In order to examine the response patterns in detail, the response of the model was expanded to the range of (-10σ, 10σ) in Figure 4.7. The response by the model without interaction term was, of course, linear regardless of changes in the range of input variables (Figure 4.7a). In contrast, the nonlinear response by the model with interaction

became more apparent in expanded ranges (Figure 4.7b). However, the curvature of sensitivity plot for the model with the interaction terms was limited in 0.015/σ, which is equivalent to 65.6σ of curvature radius according to the graphical inspection in the sensitivity plot (Figure 4.8). It seems, therefore, the model with the interaction terms can only explain infinitesimal fraction of nonlinear response in flood waste generation.

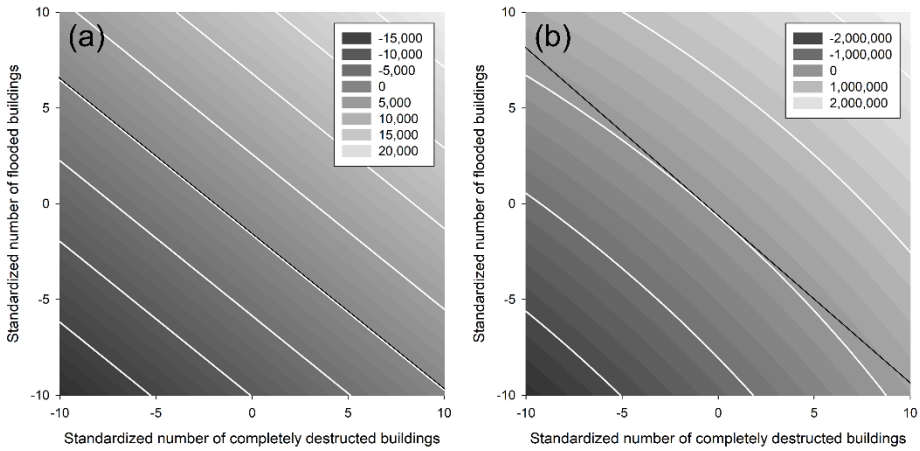

Figure 4.7. Response of waste estimation according to sensitivity analysis in an interval ranged from -10σ to 10σ for the number of completely destructed building and the number of flooded buildings: (a) model developed without interaction terms; (b) model developed with interaction terms (white solid lines: gradient curves of output responses; black solid line: a tangent line to one gradient curve)

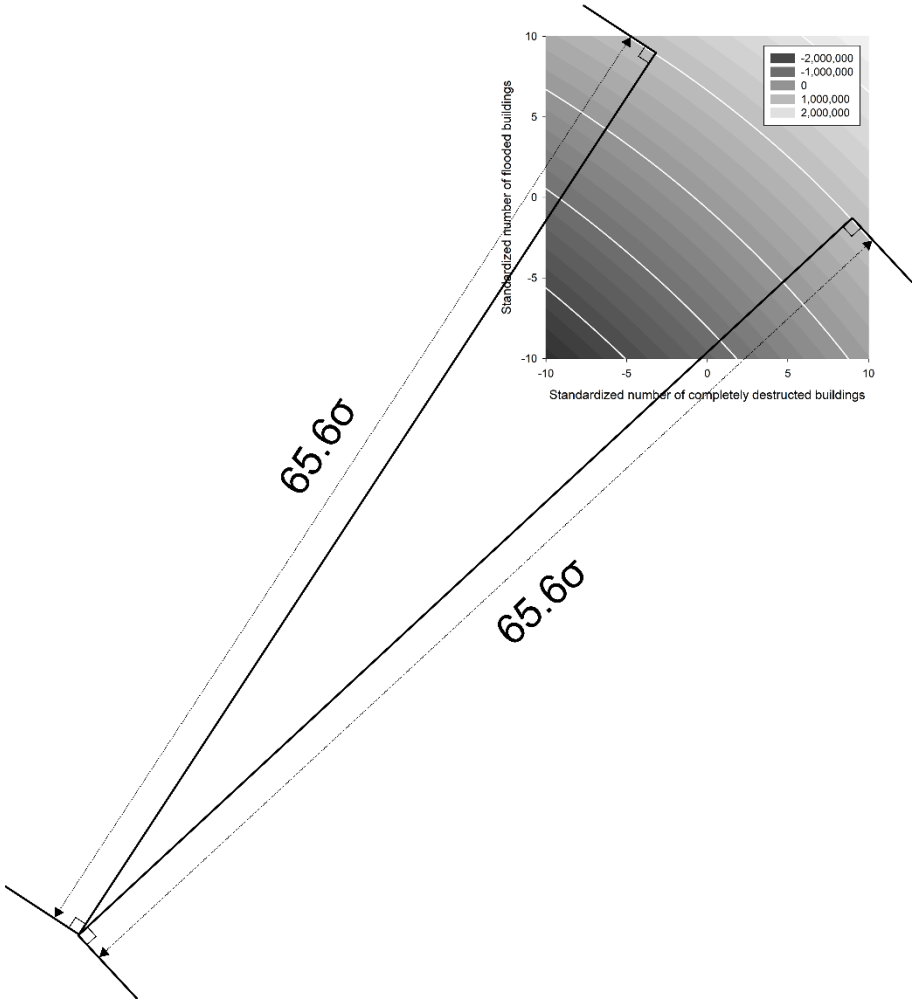

Figure 4.8. The graphical inspection diagram of calculating the curvature radius of response gradient in a sensitivity plot ranged in (-10σ, 10σ) of input variables (target contour line is assigned to 1,000,000 tons of waste output)

To summarize the finding in this section, it seemed interaction terms

is a key to explain the severe waste generations in disaster situation by elucidating nonlinear response of waste generation to multiple waste sources. The nonlinear response by the model, however, was bounded in very small extent and this may be related to the restricted performance of the model. The authors admitted that the analysis in this section was bounded to the response of double variables, and the effects combined by more than triple parameters were not investigated owing to the limitation of the examination framework.

4.3.4 Analysis of the characteristics of flood waste generation

The results of the previous section revealed the positive effect of including interaction terms for flood waste modeling. The next key question is to elucidate the mechanism of enhanced modeling by incorporating the interaction terms. Obviously, it is difficult to explain the causal-effect relationship by empirical statistics analysis since many effects are involved in all statistical data unlike the data acquired by lab-scale physics or chemistry experiments (Bolstad and Curran, 2016). Therefore, it should be noted that the description hereby might be limited in a narrow scope and the authors are trying to explain the mechanism from the perspective of a statistical tool which can be utilized.

Within the context aforementioned, VIF is perceived as a factor to

interpret multiple collinearities between input variables, which is corresponded to the interaction between input variables. Table 4.5 summarizes VIF for input parameters used in this study. The values of VIF for independent variables are always larger than 1.0 and less than 2.0. Since VIF values are larger than 1.0, it can be inferred that there are somehow inter-relationship between input variables. It is a quite natural phenomenon because input variables are flood damage records, which are affected by flood events itself, after all.

Table 4.5. Variance inflation factors for parameters in flood waste regression model developed by using 90 cases of flood waste generation in South Korea from 2008 to 2017

	CB	PB	FB	C	R	R'	S
r^2	0.41	0.45	0.11	0.01	0.43	0.39	0.40
VIF	1.71	1.83	1.12	1.01	1.74	1.65	1.67

One interesting finding is that the inter-relationship between flood damage data is not very strong as VIF values are not far from 1.0, even though flood damage data share one causal event: flood itself. This is probably because flood damage characteristics are not determined only by disaster scenarios but also affected by regional characteristics of the flood-receiving area. Flood damage can be varied depending on the design level of infrastructures enduring

natural disaster and regional stocks when the disaster event occurred. Consideration to the regional characteristics, of course, is recognized as a beneficial option to interpret disaster and foster preparedness to the disaster (Asari et al., 2013; Jo and Kim, 2016; MOEJ, 2018).

Usually, the variable with VIF larger than 10.0 is recommended to be eliminated from the regression to prevent adverse impact on accurate modeling (Neter et al., 1996). If the collinearity between input variables is suspected to be strong, it will be better to remove the variables because other variables can represent the effect of the variable on target data. The problem is, there is no shared guideline when collinearity of variables are not too significant, in other words, VIF of the variable is on the range of 1.0-10.0. As noted in the upper paragraph, the inter-relationship between flood damage variables definitely exists but is not strong due to the combined effect of flood severity and potential damage inherited by flood receiving area. Incorporating the interaction terms in regression is one possible alternative to resolve the problem of intermediate correlation between input variables. Waste generation from the flood is not solely defined by damage in a single aspect and interaction term can represent the complex contribution to the waste generation.

Unlocking the secret of flood waste generation is difficult when a researcher relies on quantitative research alone, and some aspects of disaster

events should be investigated by qualitative research. In this study, pictures and experiences from field surveying were utilized to interpret the interaction effect between flood damage variables on flood waste generation. For example, waste generation is not directly linked to the flooded area of cropland. Figure 4.9a shows an example image of flooded cropland after the strike of typhoon Danas on the rural area in Yeosu city, Jeollanam-do, South Korea. After a field survey followed by the typhoon striking, the authors found that flooded cropland did not generate flood waste, and according to the interview with the residents, they did not regard immersed crops as waste but just worried reduced growth of the crops. If the neighboring river was flooded and debris from upper streams overflowed to the croplands, the area and crops can be contaminated, and eventually, the flood waste from cropland can be generated. This logic is supported by the significance of the interaction term C × R' as shown in its low *p*-value (Table 4.3).

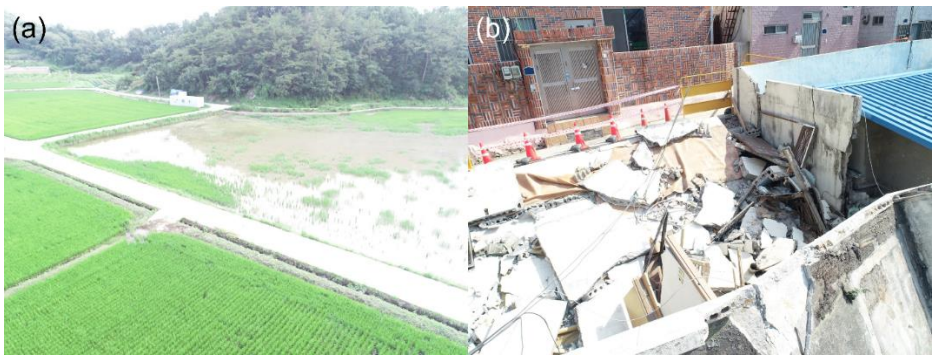

Figure 4.9. Pictures from the field surveying in flooded area (a): flooded cropland after the strike of the typhoon Danas in 16th-21st, July, 2019 (34°46'19"N, 127°35'17"E, Yeosu, Jeollanam-do, Republic of Korea) and (b): completely destructed building after the strike of the typhoon Danas in 16th-21st, July, 2019 (35°05'39"N, 129°02'53"E, Busan, Republic of Korea; some part was blocked to protect private information)

One more example showing an interaction between two variables to flood waste generation was found throughout field survey for example site of a completely destructed building after the typhoon Danas in Busan city in South Korea (35°05'39"N, 129°02'53"E). As seen in Figure 4.9b, the complete destruction of one building accompanied partial damage to the near building. In this case, a part of the wall of the right-sided house was partially destructed. Both complete and partial destruction of the building contributed generation of demolition waste. This example image clearly shows CB and PB shared flood waste generation in some cases. The combined effect on waste generation by CB and PB is represented by a large VIF for these parameters among seven input variables used in this study and might be compensated by the negative value of the coefficient for PB. Since the contribution of CB already was allocated for waste generation by PB also, the overestimated fraction by PB

should have been removed for an appropriate quantification. However, eliminating the contribution by PB is not enough because PB is also a potential waste generation and sometimes PB alone can occur without CB. Therefore compensation for interaction between CB and PB is necessary, and it was achieved by introducing the positive value of the coefficient for the interaction term CB × PB (Table 4.3).

4.4 Summary

Until now, there have been many trials to assist disaster waste management based on early-stage quantity estimation and it was recognized that a key issue in flood waste modeling is to find out significant input variables (Chen et al., 2007; Hirayama and Kawata, 2005). In many countries, however, selecting a key parameter for flood waste regression is hardly attainable due to lack of disaster records and experiences (Brown et al., 2011; Milke, 2011). It is also not easy to collect data on new damage variables as disaster records are scattered across different ministries without being managed in a unified fashion (Jo and Kim, 2016). Above all, choosing a novel input variable for modeling relies on researcher's intuition and requires a tremendous burden in trial and error (Farrar and Glauber, 1967).

Incorporating the interaction terms for regression does not require the aforementioned tasks and effectively improve the model performance. The idea and method of adding the terms are, in essence, incomparably simpler than other options, but still provide a guarantee of the performance. However, as noted in the previous section, still a large degree of error can occur, thus, a safety factor should be devised to prevent problems due to over/underestimation of waste generation.

Another key finding from this study is the nonlinear response of flood waste generation to the harmony of several interactive variables. This perspective is seemed truly obvious but never recognized because linear modeling was conceived as only one available option. Now we are ready to take a step forward the truth of flood waste generation sealed by limited knowledge and this perception will illuminate the possibility of more resilient disaster waste management from the quantity estimation to sound resource recirculation. For example, nonlinear, interactive modeling of flood waste generation will be a prospective research topic.

References

Asari, M., Sakai, S.-i., Yoshioka, T., Tojo, Y., Tasaki, T., Takigami, H., Watanabe, K., 2013. Strategy for separation and treatment of disaster waste: a manual for earthquake and tsunami disaster waste management in Japan. Journal of Material Cycles and Waste Management 15, 290-299.

Bolstad, W.M., Curran, J.M., 2016. Introduction to Bayesian statistics. John Wiley & Sons.

Brown, C., Milke, M., Seville, E., 2011. Disaster waste management: A review article. Waste management 31, 1085-1098.

Chen, J.-R., Tsai, H.-Y., Hsu, P.-C., Shen, C.-C., 2007. Estimation of waste generation from floods. Waste management 27, 1717-1724.

Cho, Y.H., 2018. Estimation of flood debris generation using GIS-based inundation maps. The Graduate School, Seoul National University.

Cortina, J.M., 1993. Interaction, nonlinearity, and multicollinearity: Implications for multiple regression. Journal of Management 19, 915-922.

Dubey, B., Solo-Gabriele, H.M., Townsendt, T.G., 2007. Quantities of arsenic-treated wood in demolition debris generated by Hurricane Katrina. Environ Sci Technol 41, 1533-1536.

Farrar, D.E., Glauber, R.R., 1967. Multicollinearity in regression analysis: the problem revisited. The Review of Economic and Statistics, 92-107.

Hirayama, N., Kawata, Y., 2005. Quantity of disaster waste for emergency response of public authirities on flood disaster. Institute of Social Safety Science 7, 325-330.

Jaccard, J., Wan, C.K., Turrisi, R., 1990. The detection and interpretation of interaction effects between continuous variables in multiple regression. Multivariate behavioral research 25, 467-478.

Jo, J.H., Kim, T., 2016. Optimal Management of Disaster Waste by its Properties Using GIS. Korea Environment Institute.

Kabacoff, R.I., 2011. R IN ACTION.

Kang, E., Ju, M., Jeon, W.S., Kim, J.Y., 2015. Flood Waste Prediction Method using Rainfall and Flooded Buildings in Seoul. Journal of Korea Society of Waste Management 32, 713-719.

Milke, M., 2011. Disaster waste management research needs. Waste management 31.

MOE, 2017. Disaster Waste Safety Management Guidelines.

MOEJ, 2018. Disaster Waste Management Guideline for Asia and the Pacific.

MOIS, 2019. Safety News, National Disaster Safety Portal, http://www.safekorea.go.kr/idsiSFK/neo/sfk/cs/sfc/dis/disasterNewsList.jsp?emgPage=Y&menuSeq=619 (Access date: 2019.7. 21)

Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W., 1996. Applied

linear statistical models. Irwin Chicago.

Park, M.H., Ju, M., Kim, J.Y., 2019a. Data classification to enhance accuracy of flood debris prediction. Proceedings of the 4th Symposium of the Asian Regional Branch of International Waste Working Group, 35-36.

Park, M.H., Ju, M., Kim, J.Y., 2019b. Limitation of deterministic regression on flood waste quantity estimation in South Korea. Proceedings of 17th International Waste Management and Landfill Symposium,, 387, 381-388.

Park, M.H., Kim, H., Ju, M., Jong Kim, H., Young Kim, J., 2018. Development of Regional Flood Debris Estimation Model Utilizing Data of Disaster Annual Report: Case Study on Ulsan City. Journal of Korea Society of Waste Management 35, 777-784.

Park, M.H., Kim, J.Y., 2019. Post-Flood Waste Estimation by Bayesian Regression: A case study on South Korea. Proceedings of the Annual Conference of Japan Society of Material Cycles and Waste Management, 535-536.

Pilapitiya, S., Vidanaarachchi, C., Yuen, S., 2006. Effects of the tsunami on waste management in Sri Lanka. Waste Manag 26, 107-109.

Rencher, A.C., Christensen, W.F., 2012. Chapter 10, Multivariate regression–Section 10.1, Introduction. Methods of Multivariate Analysis, Wiley Series in Probability and Statistics 709, 19.

Shaphiro, S., Wilk, M., 1965. An analysis of variance test for normality. Biometrika 52, 591-611.

Theil, H., 1961. Economic forecasts and policy. North-Holland, Amsterdam. The Netherlands.

Thorneloe, S., Lemieux, P., Rodgers, M., Christman, R., Nickel, K., 2007. Decision support tool for the management of debris from homeland security incidents, XI International Waste Management & Landfill Symposium Sardinia.

USEPA, 2008. Planning for natural disaster debris, Office of Solid Waste and Emergency Response and Office of Solid Waste

USEPA, 2013. Incident Waste Decision Support Tool (I-WASTE), v.6.3.

YonhapNewsAgency, 2019a. An Old House in Yeongdo, Busan collapsed suddenly, Yonhap News Agency.

YonhapNewsAgency, 2019b. "A water plant block the drain..."Watershed Farmers in Yeosu Somyeon Reclamation, Yonhap News Agency.

CHAPTER 5

OPTIMIZING DEEP-NEURAL NETWORK FOR PREDICTING WASTE GENERATION FROM FLOOD

5.1 Introduction

Knowing the size of work is the primary step for successful implementation. The process of disaster waste management is similar in the context; the amount of disaster waste generation should defined before planning and action. Authorities can allocate the required work and budget for collection and disposal activity if the amounts of waste were reported (FEMA, 2007; USEPA, 2008). It is difficult, however, to obtain the accurate volume of waste generation in the disaster site where the unstructured debris was stacked and contaminated water from unidentified sources was abundant. Simultaneously it is not necessary to talk about the scene where dozens of emergency calls rang, to the point where even rescuers were completely out of control to do something.

The difficulty in figuring out the total volume of waste results in waste matrix left unmanaged in sites. According to the (Brown et al., 2010), the waste

remained at the site blocks the road thus hinder both rescuing and recovery activities. Besides the direct blockage into road access, the waste dummy can be potential threat to environment because it can contain unknown harmful substances from household pharmaceutical products (Dubey et al., 2007; Pilapitiya et al., 2006), construction materials (Jang and Townsend, 2001a; Jang and Townsend, 2001b), etc. All mentioned concerns are clear evidence supporting necessity of rapid and proper disaster waste management.

In order to rapidly measure disaster waste recovery, many studies devoted to suggest framework for predicting the amount of waste. The earliest publication in disaster waste modeling area was made by (Hirayama and Kawata, 2005) who proposed linear regression of the number of completely destructed buildings, half-destructed buildings and just flooded buildings to estimate flood waste generation. (Chen et al., 2007) showed successful linear regression with lognormal transformation of population density, the area of flooded region, and total rainfall amount to estimate flood waste generation within an accuracy of $r^2 = 0.541$. The reason these two Asian studies focused on the flood is the huge impacts of flood in Asia, historically. This is same in South Korea, evidenced by the proportion of flood damage in the whole disaster records in a recent decade; up to 89% of damage was originated from rainfall-related cases (NEMA, 2009-2018). In this context, the scope of this study also

focused on flood.

However, there was no further progresses in flood waste regression study. In addition, their framework to establish model did not work in different dataset of South Korea in our previous investigation (Park et al., 2019a, b; Park and Kim, 2019). We were willing to suggest an advanced way of disaster waste modeling along with the recent blooming of data science. Among many methodologies, we applied deep-learning for flood waste prediction with the first trial in disaster waste management area.

Following the successes in modeling many environmental processes, many researchers have been truly interested in the application of DNN on their research questions. However, there is one important engineering problem to establish optimal structure of DNN, which has been never regarded important until now: to design the structure of DNN. The combination of hyperparameters defining the internal structure of DNN significantly affects the performance of model. The hyperparameters includes discrete types, such as the number of hidden layers, nodes in the layer, learning algorithms and continuous types, such as initial learning rate, and regularization rate. In this regard, the first objective of this study is to show the impact of hyperparameters set on performance of flood waste model. The second is to find out the optimal structure of DNN by hyperparameter tuning for the model. Although this study

concentrated on the subject of flood waste generation, authors believe that the framework shown here can be transferred for other environmental modeling studies.

5.2 Materials and methods

5.2.1 Data acquisition

The number of collected events was 90, including all the cases from 2008 to 2017 in South Korea (NEMA, 2009-2018). In order to reflect the interactive effects of multiple factors on flood waste generation, diverse variables were collected from annual report on disaster, data equipped by Korea Meteorological Administration, and Korean Statistical Information Service. The variables include factors associated with flood damage, rainfall aspect, and regional characteristics. The flood damage variables include number of totally destructed buildings, partial destructed buildings, flooded buildings, area of damaged cropland, length of damaged road, length of damaged river and length of damaged small stream. The rainfall aspect-related variables were disaster type (typhoon or heavy rain), maximum hourly rainfall, maximum daily rainfall, total rainfall, and wind velocity. The regional characteristics of flood-receiving area were province, urbanization rate, closeness to ocean, whole area, gross

regional domestic product, population density, area ratio of urbanized region, permeable land cover, impermeable land cover, and wastewater supply rate.

5.2.2 Structure of the neural network

The structure of the neural network used in this study is depicted in Figure 5.1. Input layer received input variables and those variables passed into activation layers with correspondence to nodes in the layer. The number of nodes was identical in every activation layer. Since each node receives all contributions of the nodes at preceded layer, the model can simulate interactive effects of multiple variables. After passing a number of layers, all the values were combined into output throughout regression layer expressed by linear function. The model compares the estimated value in output layer with real observation to recalibrate the model weight and bias in each nodes by back-propagation.

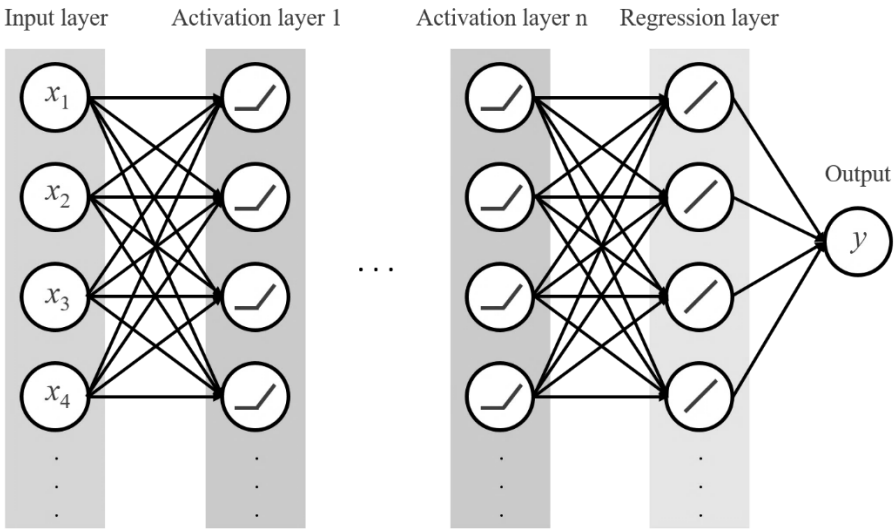

Figure 5.1. The elemental structure of the neural network used in this study

5.2.3 Deep-learning environment

All the computations such as data pre-processing, learning and validation were carried out in Spyder environment using python 3.1. The package mainly used for establishing and learning artificial neural network was Keras in Tensorflow. As an optimizer in back-propagation, adjusted adaptive scaled gradient descend method (so-called adadelta) was utilized. An activation function used in this study was rectified linear unit (ReLU) to reflect nonlinear response of flood waste generation with a certain threshold.

5.2.4 Hyperparameter tuning

Performance of neural network depends on the structure and design of hyperparameters. To achieve optimal neural network in estimating flood waste generation, hyperparameter tuning was performed in two schemes: 1) discrete type hyperparameters optimization; and 2) continuous type hyperparameter optimization. The discrete type hyperparameter included number of hidden layers and number of nodes in each layers. The optimal combination of the number of hidden layers and nodes was explored by grid search with ten times replicated cross-validation. The averaged root mean square error (RMSE) was used for evaluation index, and its uncertainty was calculated by relative standard deviation of tem times replication divided by its average. For cross-validation, in total 80% of dataset were used as training set and remainders were testing set. Initially, the examined number of hidden layers ranged from one to five. The evaluated range of number of nodes was from 11 to 110, with an interval of 11. The interval of the number of nodes were chosen to be half of the number of input variables. After the first examination, the suspected hotspot region was further investigated with detailed inspection interval. Finishing the discrete type hyperparameter tuning, the continuous type hyperparameter tuning followed. The continuous hyperparameters included initial learning rate, and regularization number in each hidden layers, respectively. The continuous type hyperparameter tuning was conducted by Bayesian optimization supported

by bayesian-optimization library in python. For Bayesian optimization, initially four points were randomly selected and 44 points were further explored with exploration rate 0.01.

5.3 Results and discussion

5.3.1 Optimal structure of DNN for flood waste estimation

Figure 5.2 shows the results of ten times replicated grid search to estimate flood waste generation. As seen in Figure 5.2a, RMSE of estimation decreased when both the number of hidden layers and nodes increased. This is associated with the principle of neural network, in other words, more complicated weight optimization through more stacked activation layer shows strong performance to explain complex patterns of flood waste generation. The RMSE in training set decreased even less than 500, which was relatively far lower than that achieved in conventional linear regression in the previous chapters. The lowered RMSE in training set shows DNN with ReLU function is truly appropriate to explain the results of flood waste generations. As proposed in Chapter 4, flood waste generation is suspected to be interactive and nonlinear to input variables with a certain level of threshold. The ReLU function is fitted to explain the suspected characteristics of flood waste

generation.

The results of grid search in testing set was much different with those in training set (Figure 5.2b). Optimal set of hidden layers and nodes seems exist within the range comprised of two layers and nodes from 78 to 92. This may be because of over-training of the dataset. Over-stacked layers and nodes can inhibit the fair explanation of dataset exterior from the training set. In order to eliminating the limitation owing to over-fitting, it is required to be selected proper set of hyperparameters. In this region, the uncertainties in both training and testing set were below 50% (Figure 5.3a and b), convincing the robustness of DNN model for estimating flood waste generation.

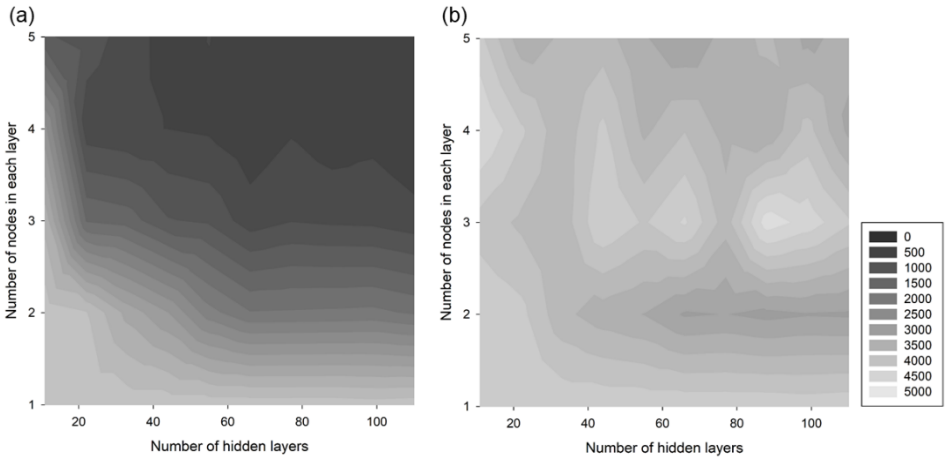

Figure 5.2. The results of ten times replicated grid search with root mean square as target function (ten times averaged) in (a) training set and (b) testing set

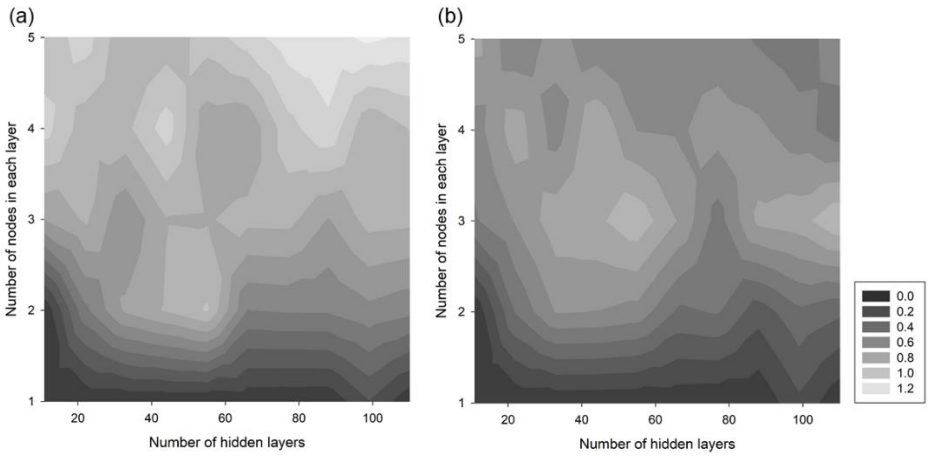

Figure 5.3. The uncertainty of ten times replicated grid search with root mean square as target function (ten times averaged) in (a) training set and (b) testing set

The detailed grid search was performed in a region of number of nodes ranged from 78 to 92, when the number of hidden layers was two. The results shows the testing set RMSE in the region were below 2,000 tons, and that of the training set closed to 3,000 tons (Figure 5.4). The optimal combination of discrete hyperparameter was observed when the number of nodes was 83 and RMSE in the condition was 2,877 tons, similar level of error in conventional regression with whole dataset. It can be argued that DNN performed better than conventional linear regression as the error in testing set was lower and that in training was even similar to the error in conventional method.

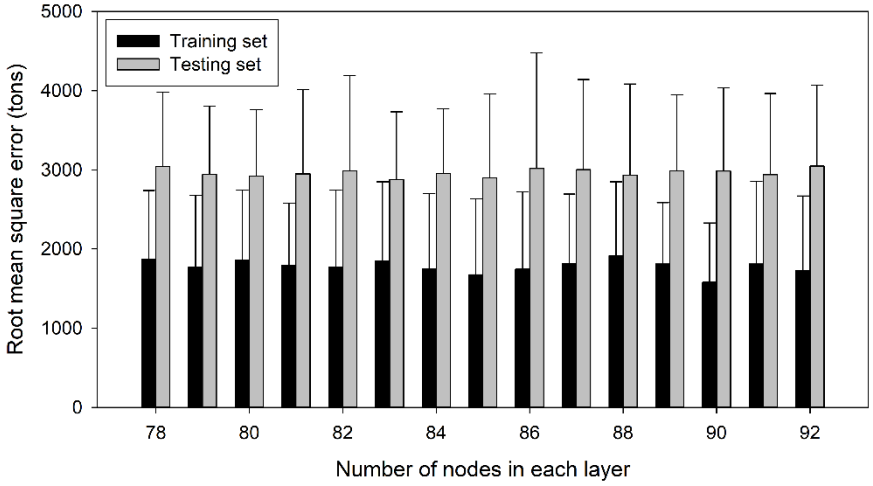

Figure 5.4. The results of detailed grid search in regions where the number of hidden layers = 2, and the number of nodes ranged from 78 to 92

5.3.2 Optimal set of continuous type hyperparameters inside the DNN made up of optimal structure

Bayesian optimization was performed with three variables, the regularization number in the first, second layer, and the initial learning rate. This was because the number of hidden layers was setted to two in the previous section. The results of Bayesian optimization when the number of hidden layers was two, and that of nodes was 83 is summarized in Table 5.1. The optimal point was found in second exploration after initial random search in four points. In the condition, the RMSE of testing set was 2,596 tons, which was smaller

than the RMSE observed before Bayesian optimization. This results show that the continuous hyperparameter optimization is required to maximize the performance of DNN in estimating flood waste generation.

Table 5.1. The results of Bayesian optimization with root mean square as target function in logarithm-transformed regularization rate in the first layer (Log_RNum1), that in the second layer (Log_RNum2), and initial learning rate (Log_INI_LR)

Log_RNum1	Log_RNum2	Log_INI_LR	RMSE	Remark
-2.805	-2.139	-2.589	2,955	Initial Point
-2.820	-3.305	-2.416	2,853	Initial Point
-3.250	-1.433	-1.145	3,181	Initial Point
-3.466	-1.833	-2.884	3,004	Initial Point
-1.390	-1.330	-1.950	2,864	
-1.798	-4.113	-4.696	2,596	Optimum
-2.708	-3.968	-3.573	3,240	
-1.295	-3.712	-1.396	3,083	
-1.062	-1.629	-1.036	2,883	
-4.965	-1.832	-1.037	3,087	
-1.986	-3.605	-1.325	3,073	
-2.346	-2.211	-3.997	2,923	
-3.298	-4.699	-1.137	3,018	
-2.630	-4.912	-4.861	2,922	
-2.234	-2.302	-4.019	2,830	
-2.681	-1.710	-3.018	2,926	

-4.851	-3.714	-3.083	2,922
-3.111	-1.125	-3.484	2,833
-4.008	-2.538	-1.732	2,850
-4.122	-4.144	-4.249	3,656
-3.796	-1.077	-2.745	2,783
-3.892	-1.104	-3.433	2,781
-4.730	-2.327	-2.158	2,737
-4.389	-1.550	-2.714	2,922
-2.728	-1.298	-4.716	2,784
-4.651	-4.919	-1.670	2,817
-1.887	-1.520	-1.086	3,000
-1.803	-3.154	-1.878	3,387
-2.439	-2.765	-3.182	3,059
-4.099	-3.548	-2.161	2,795
-1.165	-1.030	-2.832	3,311
-4.843	-4.921	-4.122	2,875
-1.723	-1.665	-1.614	3,039
-1.037	-2.643	-1.357	2,839
-4.153	-1.063	-4.674	3,009
-4.063	-1.110	-4.650	2,747
-4.042	-4.306	-2.427	3,020
-4.129	-3.868	-2.924	2,846
-1.738	-1.394	-3.817	4,017
-2.268	-3.095	-1.063	3,006
-3.117	-4.903	-1.676	2,871
-3.690	-2.438	-1.915	2,869
-4.495	-1.577	-2.061	3,421
-2.893	-2.749	-3.894	3,045
-2.743	-3.479	-3.603	2,644

-1.793	-3.149	-2.856	2,934
-2.397	-1.599	-3.145	3,181
-3.749	-1.058	-3.088	2,755

5.4 Summary

In this chapter, the potential of DNN in estimating flood waste generation. The results showed the DNN can estimate flood waste generation with fair extent of credibility. This was mainly because the model structure can explain well the nature of flood waste generations, in other words, interactive, nonlinear response of input variables with a certain degree of threshold, which can be captured by ReLU activation functions. The performance of DNN was optimized by two-phase hyperparameter optimization. The optimal set of discrete hyperparameters was two layers with 83 nodes in each layer. The optimal combination of continuous hyperparameters was $10^{-1.798}$ of regularization number in the first layer, $10^{-4.113}$ of regularization number in the second layer, and $10^{-4.696}$ of initial learning rate. The RMSE in testing set within the hyperparameter sets was 2,596 tons.

References

Brown, C., Milke, M., Seville, E., Giovinazzi, S., 2010. Disaster Waste Management on the Road to Recovery: L'Aquila Earthquake case study.

Chen, J.-R., Tsai, H.-Y., Hsu, P.-C., Shen, C.-C., 2007. Estimation of waste generation from floods. Waste management 27, 1717-1724.

Dubey, B., Solo-Gabriele, H.M., Townsendt, T.G., 2007. Quantities of arsenic-treated wood in demolition debris generated by Hurricane Katrina. Environ Sci Technol 41, 1533-1536.

FEMA, 2007. Debris Management Guide. Public Assistance.

Hirayama, N., Kawata, Y., 2005. Quantity of disaster waste for emergency response of public authirities on flood disaster. Institute of Social Safety Science 7, 325-330.

Jang, Y.-C., Townsend, T., 2001a. Sulfate leaching from recovered construction and demolition debris fines. Advances in Environmental Research 5, 203-217.

Jang, Y.-C., Townsend, T.G., 2001b. Occurrence of organic pollutants in recovered soil fines from construction and demolition waste. Waste Management 21, 703-715.

NEMA, 2009-2018. Annual Report on Disaster.

Park, M.H., Ju, M., Kim, J.Y., 2019a. Data classification to enhance accuracy

of flood debris prediction. Proceedings of the 4th Symposium of the Asian Regional Branch of International Waste Working Group, 35-36.

Park, M.H., Ju, M., Kim, J.Y., 2019b. Limitation of deterministic regression on flood waste quantity estimation in South Korea. Proceedings of 17th International Waste Management and Landfill Symposium,, 387, 381-388.

Park, M.H., Kim, J.Y., 2019. Post-Flood Waste Estimation by Bayesian Regression: A case study on South Korea. Proceedings of the Annual Conference of Japan Society of Material Cycles and Waste Management, 535-536.

Pilapitiya, S., Vidanaarachchi, C., Yuen, S., 2006. Effects of the tsunami on waste management in Sri Lanka. Waste Manag 26, 107-109.

USEPA, 2008. Planning for natural disaster debris, Office of Solid Waste and Emergency Response and Office of Solid Waste

CHAPTER 6

CONCLUSIONS

The principal objective of this study is to suggest frameworks to estimate flood waste genration in South Korea. Four methodologies were applied and detailed conclusions for each meathod are given at the end of each chapter. The general conclusions regarding the three specific objectives proposed at the chapter 1 are summarized here.

1) First of all, context-based grouping approach resulted in better estimation of flood waste generation depending on groups' properties. By screening out flood waste cases not fitting to the other cases, the framework found some groups with a harmony inside them. Bayesian linear regression showed slight improvement on estimation accuracy. Linear regression with interaction terms showed large increase in flood waste estimation performance supported by high adjusted r^2 value. Both statistical and qualitative inspection of flood waste cases showed interactive modeling is adequate to explain flood waste generation. Deep-neural network succeded to estimate flood waste generation with

little error in training set and fair level of error in testing set by explaining nonlinear waste generation characteristics.

2) The grouping approach failed to explain all the cases, in other words, some cases remained unsolved after three-phase successive grouping. Bayesian approach did not clearly improve the model performance and this may be originated in the limitation of linear modeling. The linear model with interaction terms did not fit to the assumption of the linear regression, despite the enhanced accuracy of the model. The validity of testing set in deep-neural network was still remain in question. Considering that the ReLU activation explains response of waste generation with the certain level of threshold, further investigation using the dataset including no generation of flood waste with recorded flood damage would be prospective study topic.

3) Until now, multivariate linear regression was unique possible method to estimate flood waste generation. The results in this dissertation, however, shows the flood waste generation patterns did not fit to assumption of linear model, and use of interactive, nonlinear modeling is more promising. This is related to secret of flood waste generation

which is obviously closed to nonlinear and interactive responses from multiple waste sources. Better understanding on flood waste generation characteristics will open a door toward resilient society by correct modeling and enhanced countermeasure to flood waste recovery. The next to step for improved modeling was suggested in the upper paragraph. Possibly, in the case of countermeasure to flood waste management, prevention of flood waste generation can be made by focusing on a part of flood waste generation source to eliminate the linked interaction with the source and others. This will be promising solution with efficient implementation on only the focused sector.

초 록

수해폐기물 발생량 추정모형 개발에 관한 연구

서울대학교 대학원

건설환경공학부

박만호

수해폐기물 관리는 재해복구 지연을 막고 2차 환경오염을 줄이기 위한 기본적인 재해 복구 절차이다. 수해폐기물을 적절한 수집하고, 적절한 규모의 임시 적환장을 설계하고 운영하여 안전하게 최종 처분할 방법을 계획하고 이행하려면 수해 현장에서 신속하고 정확하게 폐기물의 발생량을 추정할 필요가 있다. 다중선형회귀분석은 수해폐기물 발생량을 추정하는 기술로 제안되었으며 해외에서는 현장에서 적용 가능한 기술로 받아들여지고 있다. 다중선형회귀분석을 통한 수해폐기물 추정방법은 지역특성과 강우위험과 관련된 요소를 이용한 사전

예측과 건물 파손건수 등 홍수 피해변수를 이용한 사후 예측 두 종류가 있다. 그러나 다른 나라에서 제안된 방법론은 안타깝게도 한국의 수해폐기물 발생 양상을 잘 설명할 수 없었다. 본 연구에서는 데이터 그룹화, 베이지안 선형 회귀 분석, 상호작용항을 고려한 선형회귀분석 및 심층 신경망을 이용한 비선형 회귀분석으로 총 네 가지의 진보된 홍수 폐기물 추정 기법이 고안되었다. 본 연구의 목적은 1) 전술한 4개의 방법론을 이용하여 홍수 폐기물 회귀 모델을 개발하고, 2) 각 방법의 성능과 한계를 평가하며, 3) 각 방식으로 개발된 모델의 경험적 검증 결과를 비교하여 가장 실현 가능한 전략을 제안하는 것이었다. 모형의 경험적 검증을 위해서 전체 사례의 20%는 모형 개발에 활용하지 않고 성능 평가에 적용하는 교차검증방법을 적용하였다. 본 연구의 범위는 소방방재청이 발간하는 재해연보에서 수집한 국내 홍수 피해 총 90건(2008~2017년)을 대상으로 한다.

사례별 분류는 행정구역(특별시/도에 상당), 도시화율, 재해 유형 및 연안 접근성의 세 가지 분류 특성에 따라 사후 수해폐기물 추정 모형 개발에 적용되었다. 사례별 분류를 통해 수해폐기물 예측 모형의 정확성이 높은 몇 개의 사례집단을 찾아낼 수 있었다. 분류 이후 정확성이 개선되지 않은 사례들은 다른 기준으로 연속적인 분류를 통해 분류되었고 연속적인 분류를 통해 다시 한 번

설명력이 높아지는 사례집단을 찾아낼 수 있었다. 사례별 분류는 재해 양상이 유사한 집단을 식별하는 동시에 사례집단 내에서 정확한 폐기물 예측을 방해하는 불규칙한 사례를 제거하는 데 효과적이었다. 시험한 분류 순서 중, 수해폐기물 예측을 가장 많이 개선한 사례분류 순서는 도시화율, 행정구역, 재해 유형 및 연안 접근성이다. 이 집단화 순서는 74건에서 향상된 폐기물 예측을 산출했다. 그러나 3단계의 사례별 분류 후에도 16건은 여전히 잘 설명되지 않아 사례 분류가 모든 경우를 설명할 수는 없음을 확인하였다. 또한, 이 방법론은 일부 집단의 향상된 모델 적합성이 분류를 통한 선별 효과와 관련이 있는지 아니면 단순한 우연인지 명확하지 않다는 통계적 분석의 한계를 내포하고 있다.

수해폐기물 모델 개발 과정은 전통적으로 결정론적 접근법을 사용해왔으며 확률론적 방법은 적용된 적이 없다. 홍수로 인한 폐기물 생성이 불확실성이 큰 확률과정임을 고려하면, 확률론적 접근방식은 기존의 결정론적 접근방식에 의해 개발된 모델에 비해 더 정확한 모델을 제공할 수 있을 것으로 기대하였다. 이 연구의 한 부분은 한국의 수해폐기물 선형회귀 모델을 개발하기 위해 베이지안 추론을 적용한 것이다. 이 연구의 목적은 (1) 베이지안 방법론에 의해 추정된 선형 모형의 계수의 특성을 분석하는 것, (2) 베이지안 추정에 의한 예측 모델의 성능을

평가하는 것, (3) 베이지안 규칙에 따라 계수의 분포를 반복적으로 업데이트하는 것의 효과를 평가하는 것이다. 그 결과에 따르면, 베이지안 추론을 통해 얻은 계수는 결정론적 접근법을 통해 개발된 계수에 비해 더 작은 유의확률을 보였다. 균등사전분포를 사용한 베이지안 추론은 결정론적 회귀에 비해 눈에 띄는 만큼은 아니지만 모형의 오차를 감소시켰으며 오차의 감소는 사후예측모형에 대해 더욱 두드러지게 나타났다. 베이지안 업데이트는 모델의 정확성을 효과적으로 높이지 못했으며 반복 업데이트는 복잡한 계산 과정을 필요로 하므로 업데이트는 굳이 필요가 없을 것으로 보인다. 이러한 결과는 수해폐기물 추정에서 베이지안 방법론의 가능성을 보여주지만, 베이지안 접근방식에 의한 정확성 개선은 사실상 제한적이라는 것을 보여주었다.

두 개의 입력변수의 곱으로 구성된 상호작용 항을 도입하면 모형의 평균 제곱 오차를 낮추고 수정된 결정계수를 높일 수 있었다. 상호작용 항은 두 입력변수에 의한 기여도를 과대/과소하게 추정한 경우를 보정하여 폐기물 생성에서 두 변수의 결합 효과를 설명하는 것처럼 보인다. 변수간 상호관계가 있을 수 있다는 통계분석 결과를 2019년 태풍 다나스 이후 현장조사에서 관찰한 사례로 재확인하였다. 홍수로 인한 피해 농경지가 발생했지만 폐기물을 발생시키지 않은 사례와, 완파된 건물이 가까이 있는

건물에 물리적인 영향을 미친 사례를 관찰하여 수해폐기물 발생은 완전히 선형적으로 설명되지 않고 변수 간 상호영향을 고려할 필요가 있음을 확인하였다. 수해폐기물 모형 개발을 위해 상호작용 조건을 도입하는 것은 새로운 변수를 추가로 확보하여 모형을 개발하는 것에 비교하면 비용이 들지도 않는데 모델 성능을 크게 향상시키는 단순한 접근법이다.

마지막으로 수해폐기물 발생량을 추정하기 위한 심층신경망의 이용을 시도했다. 홍수 피해 변수, 지역 특성 및 기상 매개변수로 분류되는 총 22개의 입력 변수를 신경망 학습에 활용하였다. 각 노드에서 정류한 선형 모형(ReLU)을 활성화 함수로 적용하고, 수정된 적응형 스케일링 경사 하강방법을 학습 알고리즘으로 활용했다. 신경망의 성능을 최적화하기 위해 1) 이산형 하이퍼파라미터에 대한 그리드 검색과 2) 연속형 하이퍼파라미터의 베이지안 최적화라는 두 단계로 하이퍼파라미터 튜닝을 수행하였다. 그 결과 최적화된 하이퍼파라미터 조건은 (계층 수 =2, 계층 별 노드 수 =83, 첫 번째 계층의 규제의 정도 $=10^{-1.798}$, 두 번째 계층의 규제의 정도 = $10^{-4.113}$, 초기 학습비율 = $10^{-4.696}$)이었으며, 이때 시험 집단의 제곱근평균오차는 2,595.91톤이었다. 최적화된 심층신경망은 교차검증을 통해 평가한 시험 집단에 대한 오차가 네 가지 방법론 중에서 가장 작았다. 교차검증결과를 고려할

때, 심층신경망의 이용은 수해폐기물 발생량을 추정하기 위한 최선의 선택지로 생각된다. 이는 수해폐기물 생성의 특성과 관련이 있을 수 있다. 수해폐기물 발생량은 실제로는 입력변수에 대해 선형적이지 않으며 특정 문턱을 넘어서면 발생이 시작될 개연성이 있으며, ReLU 활성화 기능을 갖춘 심층신경망이 이를 가장 적절히 모사하는 것으로 생각된다.

주요어: 수해폐기물, 재해폐기물 관리, 통계적 모형, 폐기물 발생량 예측, 회귀분석

학 번: 2018-26321