applied sciences

MDPI

Article

Heatwave Damage Prediction Using Random Forest Model in Korea

Minsoo Park [1], Daekyo Jung [2], Seungsoo Lee [2] and Seunghee Park [1,3,*]

1 School of Civil, Architectural Engineering & Landscape Architecture, Sungkyunkwan University, Suwon 16419, Korea; pms5343@skku.edu
2 Department of Convergence Engineering for Future City, Sungkyunkwan University, Suwon 16419, Korea; jdaekyo@skku.edu (D.J.); skklss@skku.edu (S.L.)
3 Technical Research Center, Smart Inside Co., Ltd., Suwon 16419, Korea
* Correspondence: shparkpc@skku.edu

Received: 31 October 2020; Accepted: 19 November 2020; Published: 20 November 2020

Abstract: Climate change increases the frequency and intensity of heatwaves, causing significant human and material losses every year. Big data, whose volumes are rapidly increasing, are expected to be used for preemptive responses. However, human cognitive abilities are limited, which can lead to ineffective decision making during disaster responses when artificial intelligence-based analysis models are not employed. Existing prediction models have limitations with regard to their validation, and most models focus only on heat-associated deaths. In this study, a random forest model was developed for the weekly prediction of heat-related damages on the basis of four years (2015–2018) of statistical, meteorological, and floating population data from South Korea. The model was evaluated through comparisons with other traditional regression models in terms of mean absolute error, root mean squared error, root mean squared logarithmic error, and coefficient of determination (R^2). In a comparative analysis with observed values, the proposed model showed an R^2 value of 0.804. The results show that the proposed model outperforms existing models. They also show that the floating population variable collected from mobile global positioning systems contributes more to predictions than the aggregate population variable.

Keywords: heatwaves; big data; random forest regression model; machine learning; prediction

1. Introduction

According to the National Center for Environmental Information of the National Oceanic and Atmospheric Administration, the average annual global temperature has reached an all-time high over the past five years (0.75–0.95 °C rise from the average annual temperature in the 20th century) and is continuing to gradually increase. Global warming has considerably changed the climate in recent decades, increasing the probability and intensity of meteorological and climatic disasters [1,2]. The duration and intensity of heatwaves are expected to increase with an increase in the average annual temperature, and deaths from heatwaves are expected to double [3]. The record heatwave in the United Kingdom in 2003, which killed 70,000 people, is expected to become normal summer weather by 2040 [4].

Because heatwaves cause human and physical disasters every year, it is important to minimize disaster damage by establishing timely and preemptive disaster responses. A disaster response is a continuous decision making process conducted on the basis of a variety of information and past experiences that are continuously gathered from a range of locations. Further, disaster response is conducted from the moment a disaster is perceived to have occurred until the time when it ends. In the past, data collection techniques were less effective and provided limited information for use in

contextual judgment and decision making. Consequently, owing to the lack of information available for contextual judgment and decision making, disaster responses were highly dependent on the subjective experiences of decision makers. Furthermore, although data collection technology has developed rapidly with an increase in the information available for decision making, the capability of humans to process and use this information in disaster response is limited, especially in cases that require swift decision making.

The importance of utilizing big data and artificial intelligence (AI)-based analysis for the rapid processing of various types of data has been recognized. Big data refers to large and diverse forms of data that cannot be processed by traditional database systems. Further, big datasets can include signals, images, and documents whose sizes increase exponentially; such data are abundant, owing to the development of sensing and social media-oriented communication technologies within the present Internet of Things environment [5,6]. Big data systems not only utilize a variety of data quickly but are also expected to play a crucial role in analyzing meaningful information. However, early systems only focused on data collection and storage [7]. To produce meaningful results from big data, AI technology as well as simple statistical and visualization functions must be employed for analysis and prediction.

Heatwave definitions vary among different countries [8]; however, heatwaves are generally defined on the basis of the normal weather and temperatures corresponding to the seasons of a region, and they are said to occur when there is a large deviation from the normal climate pattern in a given region. These extreme weather conditions occur locally and extensively, which limits rapid disaster response. In particular, because such extreme weather conditions occur extensively throughout a region, response procedures, such as preparing resources immediately in the event of a disaster, are limited. This indicates the need to develop early warning systems to guide disaster responses. Previous studies have focused on mortality as an endpoint for the analysis of damage caused by heatwaves [9–11], and only few studies have focused on morbidity as an indicator [12]. In addition, most studies have adopted only weather-related parameters as predictor variables of mortality. However, even under the same weather conditions, the damage pattern can vary, and it depends on other variables, such as the vulnerable population. This emphasizes the need to consider various variables as well as weather-related parameters to predict heatwave damage.

In this study, a heat-related health prediction model was developed on the basis of a machine learning algorithm for early warning systems. The purpose of this study is to help decision makers to preemptively respond, reducing human and economic losses. This paper is organized as follows: Section 2 describes the architecture of the random forest (RF) architecture and variables that can represent damage caused by heatwaves obtained from a big data collection site operated in South Korea. Experimental results, including variable evaluation, model optimization, and RF's accuracy evaluation in comparison with the tradition regression models is mention on Section 3. A trained model was applied to the site and visualized—this is also specified in Section 3. Section 4 presents the discussion and conclusion of this study.

2. Methodology

2.1. Test Area

South Korea was selected as the test bed, and its heatwave characteristics were investigated to establish the range and duration of the collected data for model training. The typical weather pattern that causes heatwaves in South Korea is a significant rise in temperature during the daytime, owing to stagnant high atmospheric pressure, which is a widespread occurrence across the country [13]. Although heatwave standards vary by country, a heatwave warning in South Korea is issued when the daily maximum temperature is expected to be above 33 °C for at least 2 consecutive days. Alerts are concentrated mainly from June to August. Heatwave occurrences in South Korea exhibit substantial interannual variability, but recently, they have become more frequent in late May and early September, and their frequency and intensity have increased [14,15]. In particular, record-breaking heatwaves

occurred in 2016 and 2018, causing many casualties [16]. The Korea Disease Control and Prevention Agency (KDCA; formerly the Korea Centers for Disease Control and Prevention) has been operating a nationwide thermal disease monitoring system since 2011 to determine the weekly health damage caused by heatwaves from late May to early September every year. South Korea has 17 administrative districts composed of 8 municipalities and 9 provinces. In accordance with the characteristics of these test beds, we set the range resolution to match the 17 administrative divisions, and the temporal resolution was set to a 1-week period to match the disease monitoring system data from the KDCA.

2.2. Variable Selection

Relevant variables were selected to predict heat-related damage. Heat-related diseases mainly occur in the form of cardiovascular and respiratory diseases and heatstroke [17]; consequently, various epidemiological studies of their occurrence have been conducted worldwide [17–19]. Among the most important characteristics of the damage caused by heatwaves and the corresponding vulnerabilities are the damage patterns of disasters, which cannot be obtained from temperature variables alone [20,21]. Studies have shown that the damage caused by disasters is related to geographic features [22,23], surface relative humidity [24–26], wind speed [27,28], population density [29], economic status [30], and vulnerable occupational groups (laborers, construction, and agricultural workers) [31–34]. On the basis of important characteristics determined in previous studies, we selected the following variables: temperature, humidity, wind speed, number of vulnerable occupational groups, insurance premiums per person, personal income per person, floating population, and registered population of residents (the number of people counted by the administration). The vulnerable population can be inferred from data on insurance premiums, income, and vulnerable occupational groups; further, as the values of these indicators increased, the number of patients with thermal diseases increased. However, both the aggregate and floating populations were used as population variables, and it was expected that the floating population, which reflects real-time information, would be a more useful variable for predictions than the aggregate population.

2.3. Random Forest Regression

RF is an ensemble machine learning method that combines several separately trained models to create a strong learner that can be applied for classification and regression [35]. Such a combination of individual models can reduce overfitting and improve generalization. Therefore, RF has the advantages of high prediction accuracy and algorithm robustness. When training ensemble classifiers, techniques involving the use of different datasets or properties are applied to create different training models. As shown in Figure 1, RF is based on the bootstrap method which is resampling technique that involves random sampling of a dataset with replacement. Then repeats the process k times to obtain several independent and identically distributed training subsets $\{S_{train,1}, S_{train,2}, \ldots, S_{train,k}\}$, which have n samples. Then, m features from the n samples are selected without accepting duplicate samples. Prediction results from different decision trees build each training subset. The most commonly obtained forecast results are selected and determined by the final forecast [35,36]. In conclusion, although some trees created in RF may be exposed to overfitting, overfitting of the RF can be prevented by generating a large number of trees. RF algorithms have been applied to various disaster fields to predict [37,38], forecast, and evaluate risks [39,40].

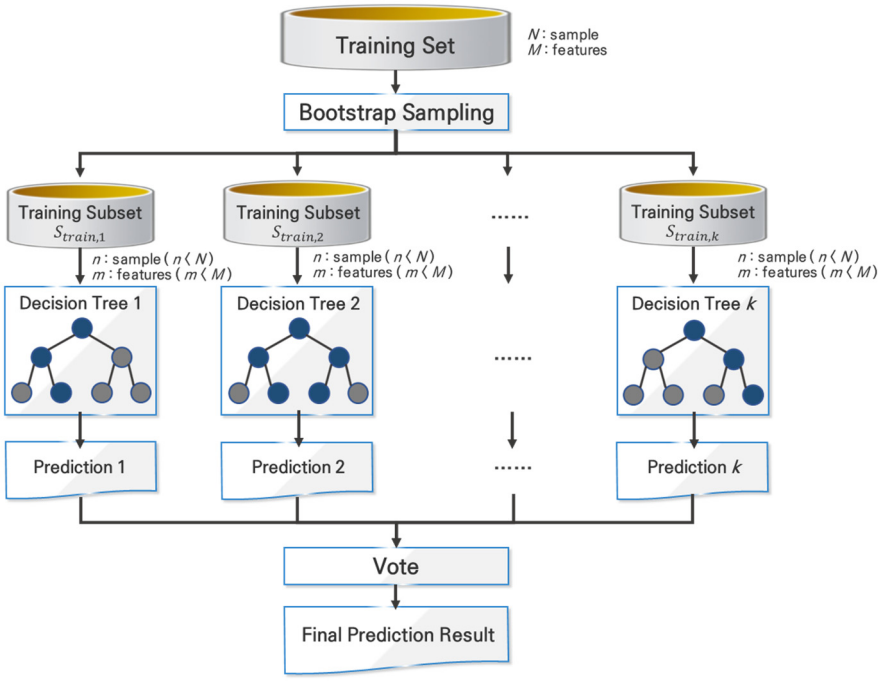


Figure 1. Architecture of a random forest.

A loss function measures the similarity between the values predicted by a model and the correct values. To increase the accuracy of a model, the loss should be reduced as the model is trained. Different loss functions are used depending on the characteristics of the model (classification or regression) and dataset. The representative loss functions for measuring errors in regression models are mean absolute error (MAE) and mean squared error (MSE):

$$MAE = \frac{\sum_{i=1}^{N} |y_i - \hat{y}|}{N} \quad (1)$$

$$MSE = \frac{\sum_{i=1}^{N} (y_i - \hat{y})^2}{N} \quad (2)$$

where N is the total number of data points, y is the real (observed) output value, and $\hat{y}$ is the predicted output value. When determining the MAE, the difference between the observed and predicted values of each data point is summed, and when determining the MSE, the square of the difference between observed and predicted values is summed. Therefore, the MSE is more sensitive to outlier values than the MAE. When the temperature exceeds a certain range, the heatstroke patients with thermal damage is characterized by a rapid increase in the incidence of patients. Consequently, MAE was considered as a loss function in this study to apply the characteristic of the target data.

To evaluate regression models, the proximity of predicted values to the observed data is quantified on the basis of the MAE, root mean squared error (RMSE), root mean squared logarithmic error (RMSLE), and coefficient of determination (R^2), which are mainly used to evaluate accuracy [41,42]. However, the mean deviations of MAE, RMSE, and RMSLE (the lower the value, the higher the accuracy) have different values depending on the scale; therefore, it is difficult to make inferences using the absolute values alone. In contrast, R^2 is a relative value because it is the variance ratio of dependent variables predicted from independent variables; thus, the performance can be intuitively determined. R^2 generally ranges from 0 to 1. Note that if the R^2 value of a model is 0.7 or more, the model is usually considered reasonable [43].

The RF model was established to predict the number of patients with heat-related diseases caused by heatwaves. Socioeconomic, demographic, meteorological, and demographic data were collected

and used as input variables for the model. The Boruta algorithm was used to filter the variables in the RF model [38]; this algorithm uses a Z score calculated by dividing the average loss by its standard deviation. It was implemented using an R package [44] to confirm whether certain variables can be used as predictive model inputs. Typical parameters of the random forest algorithm are ntree and max depth. To select optimal hyperparameters, the minimum loss function value (MAE) was found by increasing the number of decision trees (n-tree) and their maximum depth (max depth). After separating the dataset comprising the selected variables into training and test datasets, we evaluated the model trained using the training dataset by comparing it with other traditional regression models. Finally, the mean decrease in impurity (i.e., Gini importance) was used to extract the variable importance values, i.e., to determine the predicted contribution of each variable's model.

3. Results of Predicting the Number of Heatwave-Related Patients

3.1. Data Collection and Pre-Processing for Model Training

The variables and target data are listed in Table 1 with their data sources and renewal cycles. The variables are categorized as static or dynamic. Further, the abbreviations of the variables are used hereafter in the main text, figures, and tables. The static variables were pre-collected from a government agency that manages big data. They are universally updated quarterly and yearly, making them less volatile when predicting the number of heatwave-related patients in summer. In contrast, the dynamic variables, such as floating population and weather information, change with time. In South Korea, big data regarding the floating population are estimated on the basis of mobile big data collected hourly and monthly by SK Telecom's nationwide mobile communication base stations, and the estimated data are obtained from the Statistical Data Center. They are also estimated using public big data and communication data provided by the Seoul Open Data Plaza. Weather data are collected hourly and were provided by the Korea Meteorological Administration (KMA).

Table 1. Descriptions of variables to predict the number of heatwave-related patients.

Variable Description	Abbreviation	Units	Data Source
Static variables—socioeconomic and demographic data			Korean statistical information service
Per capita income	Income	×$1000	
Insurance premiums per person	Insurance	×$1000	
Resident registration population	RRP	×1	
Number of vulnerable occupational groups (agricultural, manufacturing, and construction workers)	V-groups	×1000	
Dynamic variables—meteorological data			KMA
Maximum temperature of the week	Max Tem	°C	
Minimum temperature of the week	Min Tem	°C	
Mean temperature of the week	Mean Tem	°C	
Median temperature of the week	Median Tem	°C	
Variance temperature of the week	Variance Tem	°C	
Mean humidity of the week	Mean Hum	%	
Mean wind speed of the week	Mean wind speed	m/s	
Dynamic variables—demographic data			Statistical data center
Floating population	FP	×1	

However, the weather data, particularly those collected from sensors, may have missing values due to sensor defects. To address the problem of missing values, we used datasets consisting of columns with no missing values in order to predict the missing values of other datasets. Target data were based on weekly data obtained from the thermal disease monitoring system managed by KDCA (patients with heat-related diseases and deaths caused by heatwaves in emergency rooms nationwide);

data regarding heat-related diseases such as heat stroke, exhaustion, cramps, fainting, and edema were also provided as weekly data. The resolution of the entire dataset was unified through considering the data properties of the features and targets; the temporal resolution was set to 1 week, and the range resolution was set on the basis of the South Korean administrative divisions. The datasets were randomly used for classification—80% were used as learning data and the remaining 20% as test data. Finally, variables were normalized before being inputted into the RF model to avoid creating a model that depends on specific variable units owing to the different ranges of each variable.

All variables were confirmed using the Boruta algorithm. The contribution of each variable to the RF prediction model is shown in Figure 2. The edges of each box represent the quartiles, and the line through each box represents the median. Each bar represents the 1.5 interquartile range of the nearer quartile, and the open circles represent outliers. The blue boxes correspond to the minimal, average, and maximum Z scores of a shadow attribute in the Boruta algorithm. The green boxplots correspond to confirmed important attributes. It was confirmed that all collected data from the Boruta algorithm can be used as variables of the predictive model.

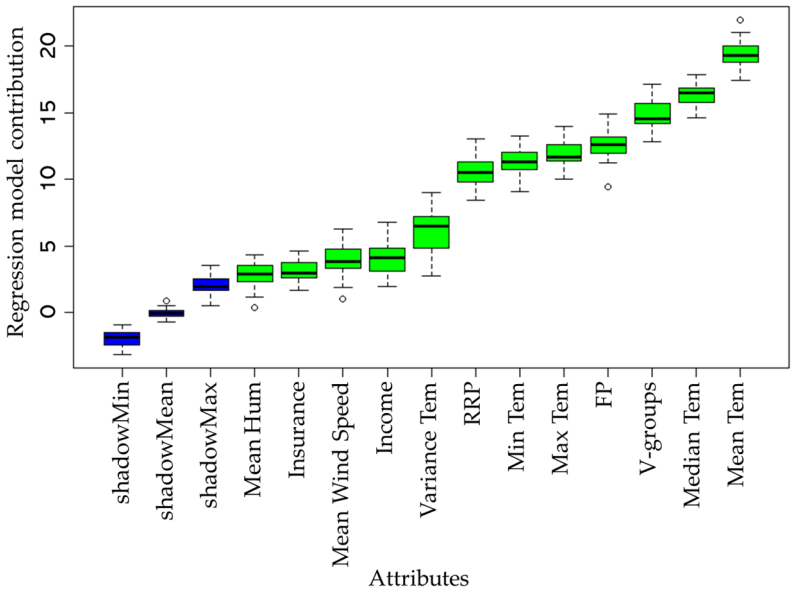
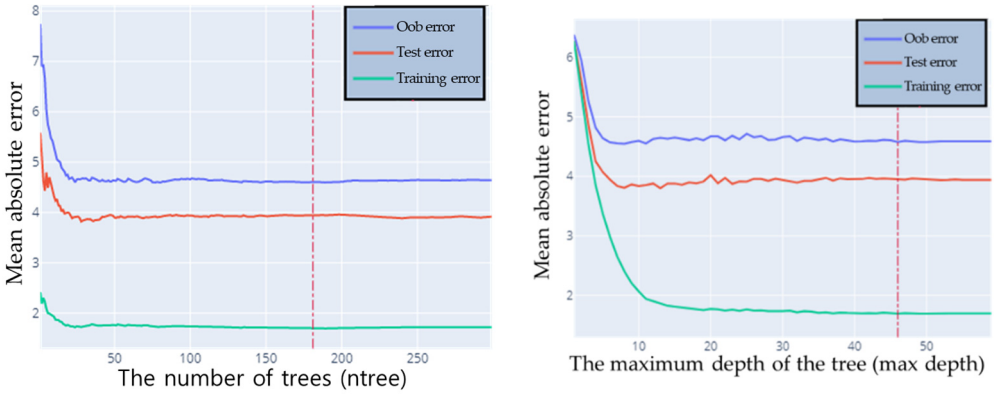


Figure 2. Contribution ranking importance of the 12 independent variables in the random forest (RF)-based variable reduction algorithm from the Boruta package [44] in R.

3.2. Hyper-Parameter Optimization

The experiment was conducted using the Scikit-learn (v.0.22.2) Python package [45] to implement the RF; the hardware platform was an Intel (R) Core (TM) i9-9900k 3.60 GHz CPU with 32 GB of RAM. The out-of-bag (OOB) error is mainly used to measure errors in machine learning models, such as bootstrap aggregation (bagging), which can be substituted for test errors [46]. The lowest MAE was found for the training, OOB, and test errors as n-tree and max depth increased, and is shown in Figure 3. When the number of decision trees was more than 100, all graphs remained almost unchanged, and when the number of decision trees was 181, the lowest OOB error was found (4.59). The training and test errors at this time were 1.67 and 3.94, respectively.



(**a**) (**b**)

Figure 3. Hyperparameter optimization in the RF regression model: (**a**) training curves with respect to number of trees and (**b**) training curves with respect to maximum depth of trees.

On the other hand, when the number of decision trees was 181, the maximum depth of the decision tree remained constant from over 40, and the lowest OOB error (4.58) was found when the depth was 46. The train error and test error at this time were 1.69 and 3.94, respectively. Therefore, the hyperparameters were determined with 181 decision trees and 46 tree depths.

3.3. Model Comparasion

The RF model was trained on the basis of the determined hyperparameters, and test data were applied to the regression model. The linear regression relationship between the predicted data from the model and test data is shown in Figure 4. The black line in the graph represents the regression line. The *x*-axis represents the weekly predicted number of patients with heat-related diseases in a specific region, as predicted by the model, and the *y*-axis indicates the weekly number of real patients with heat-related diseases in the region. The translucent band around the regression line area indicates the size of the confidence interval, which was 95% in this case. The red dotted line indicates when the model accurately reflected reality (slope: 1). The linear fitting slope of this RF model was 1.11. In particular, when high values were predicted, they tended to be underestimated compared to the observed values; however, the models were confirmed to be relatively reasonable.

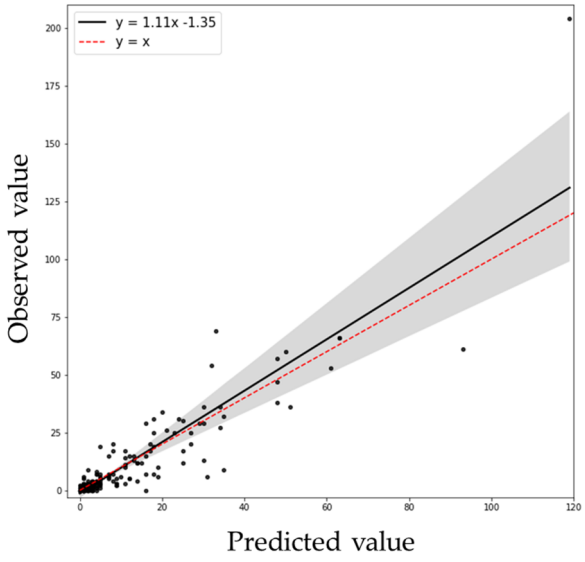


Figure 4. Linear fitting results of test data.

To compare the accuracy of the regression model more quantitatively, Table 2 compares its results with those of other regression models. In particular, the RF model is compared with linear regression, decision tree, and support vector machine (SVM) models. All models were trained using the same training set, and all the trained models were evaluated by the same test set. However, some of the values predicted by the SVM model were negative; because the values must be greater than or equal to zero, we treated all negative values as zeros. As shown in Table 2, the best values for all the considered metrics, including MAE (3.816), RMSE (8.655), RMSLE (0.645), and R^2 (0.803), were obtained for the RF model. This means that the RF is more accurate than other models for making predictions, and the R^2 value of 0.803 proves that this model is reasonable.

Table 2. Comparisons of performance evaluation.

Method	MAE	RMSE	RMSLE	R^2
Logistic regression	5.301	12.460	0.855	0.593
SVM	5.184	8.800	0.956	0.797
Decision tree	5.524	13.384	0.803	0.531
Random forest	**3.816**	**8.655**	**0.645**	**0.804**

The bold is the best result among other methods.

3.4. Feature Importance

Figure 5 shows the estimated variable importance rankings corresponding to the model. The weekly mean temperature variable, which had a value of 0.440, contributed the most in this model, followed by the vulnerable occupational groups (0.129), weekly median temperature (0.102), floating population (0.098), and weekly max temperature (0.085) variables. These five variables can be considered the main variables for prediction, whereas the rest are less important. Interestingly, the variable importance rankings proved that the floating population variable, which changes with time, had a greater effect on prediction than the population of registered residents. However, regional economic indicators had less impact on diseases related to heatwaves, as observed from the low values for income (0.020) and insurance (0.013).

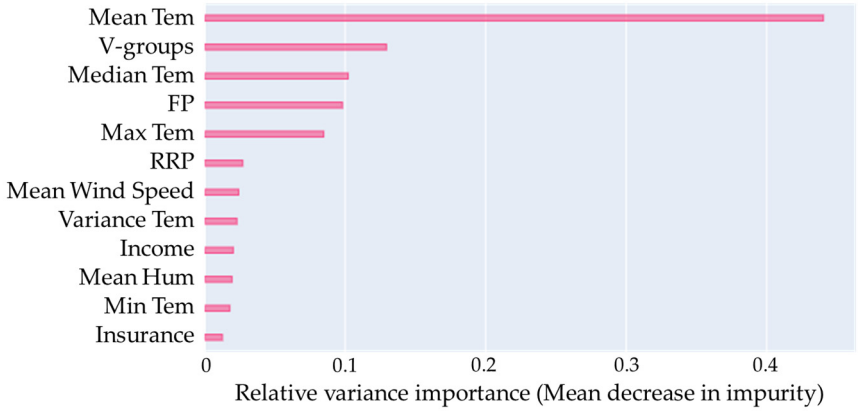


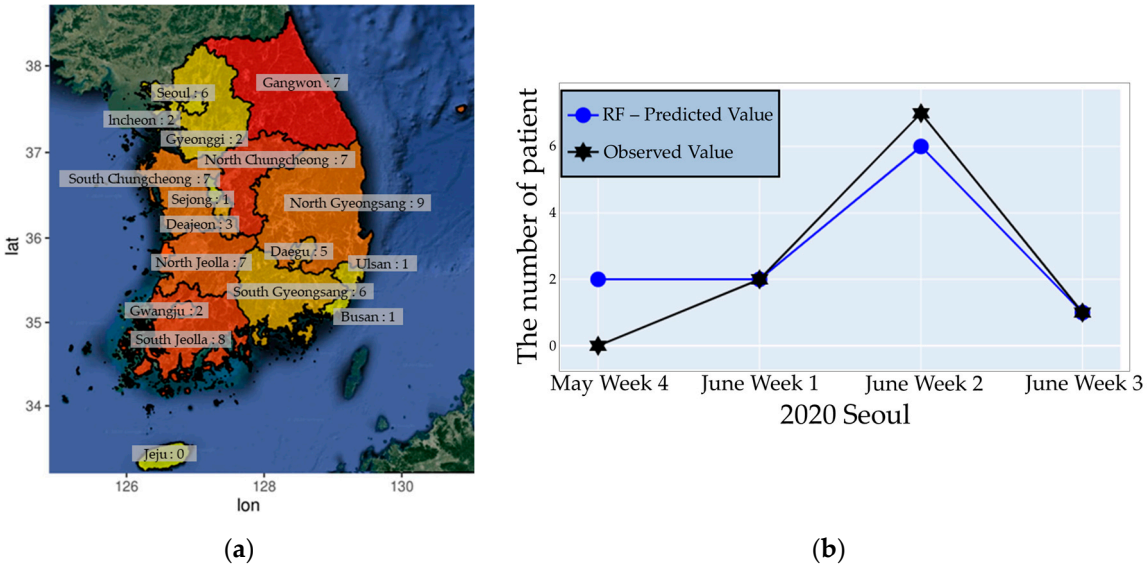
Figure 5. Variance importance in the RF model.

3.5. Model Application and Visualization

To apply the validated model, we predetermined the dynamic variables from predictions. Because the time series of variables and result values were the same, the predicted variable values must be used for prediction. With regard to static variables, we employed the latest data as inputs among the information that is updated periodically, which is the same as in model learning. The values of the dynamic population were replaced with dynamic variables using weather forecast data provided by

KMA on a weekly basis and a time series forecasting library, called Prophet [47], which is provided by Facebook.

The performance results and visualization of the model are shown in Figure 6. From the end of May, which was when heatwave management began, the substituted variables were inputted into the model for 4 weeks, and then the predicted values were obtained and compared with the observed values obtained from KDCA on the weekend. The forecasted and observed values for Seoul were compared for 4 weeks, and in the second week of June, the predicted values for each administrative district of South Korea were numerically quantified to visualize the high-risk areas and provide information to heatwave disaster response decision makers. The high-risk areas are shown in dark colors, whereas the lower risk areas are shown in lighter colors. Considering objectivity by region, we used the number of patients and the predicted floating population ratios to calculate the risk. Four weeks of data were applied to real-world situations, resulting in an R^2 value of 0.70.



(**a**) (**b**)

Figure 6. (**a**) Visualization of predicted number of heatwave-related patients in the second week of June 2020 in South Korea. (**b**) Predicted and observed data for number of heatwave-related patients in Seoul over a month.

4. Discussion and Conclusions

Heatwave damage prediction has been investigated in the United States, Europe, and Asia [9,48–50]. However, existing predictions are limited to practical notification systems owing to unrepresentative data and insufficient data accuracy [51]. According to previous research, this problem is due to the use of heatwave mortality alone as the endpoint of damage. Because the mortality rate of heatwaves is exceedingly small compared to that of the general population, it is more effective to predict risk by morbidity, which is relatively higher in proportion than heatwave mortality.

"Temperatures exceeding 33 °C" is the only available criterion for identifying the danger of heatwaves in South Korea, which allows the government to raise risk awareness by alerting the public. However, the damage to the population (deaths and sickness) caused by heatwaves varies even at the same temperature. Therefore, using only temperature data cannot determine the level of damage to peoples' health. On the basis of epidemiological investigations performed in previous research, we selected relevant variables and evaluated them by the Boruta algorithm. Then, a random forest-based heatwave damage prediction method was proposed, and its performance was compared with other traditional models. Previous studies considering demographic information have mainly used data with static characteristics, such as monthly statistical information. More accurate predictions were achieved by matching the exposure to heatwaves in a specific area to the population in that area

using dynamic population data updated more frequently, allowing this variable to contribute more to the prediction.

In the evaluation of the importance of variables, the average temperature variable and the number of occupational groups that are considered to be vulnerable to heat waves were highly evaluated (the average temperature for a week is the sum of the week, that is, the accumulated temperature). This result supports the importance of predicting the cumulative temperature in advance and responding in advance in order to minimize heat damage. In addition, it provides grounds that preemptive responses from the government, such as operation of sprinkler trucks and installation of shade curtains, should be made in areas with many vulnerable occupations. The learning process performed to build the machine learning model used independent variables based on previously recorded data. However, it is difficult to apply big data to a real-time environment owing to limitations such as irregularity in the frequency of data. Furthermore, when the time series of the dependent and independent variables are configured identically in training, the conditions for predictions are not effectively established in practical systems. Therefore, a method that employs predicted variable values was proposed. Although the accuracy of predicting future patients by applying predicted data is lower than the test accuracy during validation, the R^2 value of 0.70 supports the fact that this model provides reasonable information. Governments can use the methods developed in this study to provide disaster response decision makers with a reasonable basis for prioritizing an administrative area to provide a preemptive response and disaster support.

Nonetheless, this study has several limitations. First, the temporal resolution of the predicted values is relatively coarse-grained; thus, it is impossible to provide daily predictions for heatwaves, which are expected to occur every day. Most studies on heatwaves in South Korea have provided weekly information [11,49]. As a result, it was inferred that the resolution of target values (heat patients) is tailored to the minimum information time unit of data source. Secondly, during the experiment, data obtained across the country were input into one learning algorithm, and regional differences between administrative districts were not considered. South Korea is a relatively small country; although the regional environmental difference is relatively smaller than that in larger countries, the differences in geography and weather between its eastern and western regions are substantial. Therefore, developing individual algorithms for each region can improve model performance. This problem can be solved because the model can be trained for each region if sufficient datasets are available. Since 2018, South Korea has been managing vulnerable populations by operating heatwave shelters. The prediction model established in this study will contribute to future studies to select regions at risk of heatwaves and provide decision makers with a basis for installing heat shelters using high regional resolutions and estimating the cumulative number of patients relative to the population.

Author Contributions: Idea development and original draft writing, M.P.; project administration, D.J.; draft review and editing, S.L.; supervision and funding acquisition, S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a grant (2019-MOIS31-011) from the Fundamental Technology Development Program for Extreme Disaster Response funded by the Ministry of Interior and Safety, Korea, and supported by the Korea Ministry of Land, Infrastructure and Transport (MOLIT) as an Innovative Talent Education Program for Smart City.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Djalante, R. Key assessments from the IPCC special report on global warming of 1.5 °C and the implications for the Sendai framework for disaster risk reduction. *Prog. Disaster Sci.* **2019**, *1*, 100001. [CrossRef]
2. Peduzzi, P. The disaster risk, global change, and sustainability nexus. *Sustainability* **2019**, *11*, 957. [CrossRef]
3. Schär, C.; Vidale, P.L.; Lüthi, D.; Frei, C.; Häberli, C.; Liniger, M.A.; Appenzeller, C. The role of increasing temperature variability in European summer heatwaves. *Nature* **2004**, *427*, 332–336. [CrossRef] [PubMed]
4. Public Health England. *Heatwave Plan for England*; Public Health England: London, UK, 2019.

5. Lim, J.; Bok, K.; Yoo, J. Design and implementation of a realtime public transport route guidance system using big data analysis. *J. Korea Cont. Assoc.* **2019**, *19*, 460–468.
6. Choi, S.H.; Park, Y.J.; Shim, J.H. Strengthening of disaster management ability through big data utilization. *J. Korean Soc. Civ. Eng.* **2015**, *63*, 21–28.
7. Lee, J.H.; Baek, S.H.; Lee, S.J.; Bae, H.Y. The method for Real-time complex event detection of unstructured big data. *Korea Spat. Inf. Soc.* **2012**, *20*, 99–109.
8. Meehl, G.A. More intense, more frequent, and longer lasting heat waves in the 21st century. *Science* **2004**, *305*, 994–997. [CrossRef]
9. Green, H.K.; Andrews, N.J.; Bickler, G.; Pebody, R.G. Rapid estimation of excess mortality: Nowcasting during the heatwave alert in England and Wales in June 2011. *J. Epidemiol. Comm. Health* **2012**, *66*, 866–868. [CrossRef] [PubMed]
10. Anderson, G.B.; Oleson, K.W.; Jones, B.; Peng, R.D. Classifying heatwaves: Developing health-based models to predict high-mortality versus moderate united states heatwaves. *Clim. Chang.* **2018**, *146*, 439–453. [CrossRef]
11. Kim, D.W.; Deo, R.C.; Park, S.J.; Lee, J.S.; Lee, W.S. Weekly heat wave death prediction model using zero-inflated regression approach. *Theor. Appl. Climatol.* **2019**, *137*, 823–838. [CrossRef]
12. Williams, S.; Nitschke, M.; Weinstein, P.; Pisaniello, D.L.; Parton, K.A.; Bi, P. The impact of summer temperatures and heatwaves on mortality and morbidity in Perth, Australia 1994–2008. *Environ. Int.* **2012**, *40*, 33–38. [CrossRef] [PubMed]
13. Lee, W.S.; Lee, M.I. Interannual variability of heat waves in Korea and their connection with large-scale atmospheric circulation patterns. *Int. J. Climatol.* **2016**, *36*, 4815–4830. [CrossRef]
14. Suh, M.S.; Oh, S.G.; Lee, Y.S.; Ahn, J.B.; Cha, D.H.; Lee, D.K.; Hong, S.Y.; Min, S.K.; Park, S.C.; Kang, H.S.; et al. Projections of high resolution climate changes for Korea using multiple-regional climate models based on four RCP scenarios. Part 1: Surface air temperature. *Asia Pac. J. Atmos. Sci.* **2016**, *52*, 151–169. [CrossRef]
15. Min, K.H.; Chung, C.H.; Bae, J.H.; Cha, D.H. Synoptic characteristics of extreme heatwaves over the Korean peninsula based on era interim reanalysis data. *Int. J. Climatol.* **2020**, *40*, 3179–3195. [CrossRef]
16. Lee, H.D.; Min, K.H.; Bae, J.H.; Cha, D.H. Characteristics and comparison of 2016 and 2018 heat wave in Korea. *Atmosphere* **2020**, *30*, 1–15.
17. Reid, C.E.; O'Neill, M.S.; Gronlund, C.J.; Brines, S.J.; Brown, D.G.; Diez-Roux, A.V.; Schwartz, J. Mapping community determinants of heat vulnerability. *Environ. Health Perspect.* **2009**, *117*, 1730–1736. [CrossRef]
18. Huisman, M.; Kunst, A.E.; Mackenbach, J.P. Socioeconomic inequalities in morbidity among the elderly: A European view. *Soc. Sci. Med.* **2003**, *57*, 861–873. [CrossRef]
19. Basu, R. High ambient temperature and mortality: A review of epidemiologic studies from 2001 to 2008. *Environ. Health* **2009**, *8*, 40. [CrossRef]
20. Vose, R.S.; Applequist, S.; Bourassa, M.A.; Pryor, S.C.; Barthelmie, R.J.; Blanton, B.; Bromirski, P.D.; Brooks, H.E.; Degaetano, A.T.; Dole, R.M.; et al. Monitoring and understanding changes in extremes: Extratropical storms, winds, and waves. *Bull. Am. Meteorol. Soc.* **2014**, *95*, 377–386. [CrossRef]
21. Zubov, D.; Barbosa, H.A.; Duane, G.S. A nonanticipative analog method for long-term forecasting of air temperature extremes. *arXiv* **2015**, arXiv:1507.03283.
22. Gershunov, A.; Johnston, Z.; Margolis, H.G.; Guirguis, K. The California heat wave 2006 with impacts on statewide medical emergency. *Geogr. Res. Forum* **2011**, *31*, 53–69.
23. Guirguis, K.; Gershunov, A.; Tardy, A.; Basu, R. The impact of recent heat waves on human health in California. *J. Appl. Meteor. Climatol.* **2014**, *53*, 3–19. [CrossRef]
24. Basu, R.; Samet, J.M. Relation between elevated ambient temperature and mortality: A review of the epidemiologic evidence. *Epidemiol. Rev.* **2002**, *24*, 190–202. [CrossRef] [PubMed]
25. Kovats, R.S.; Hajat, S. Heat stress and public health: A critical review. *Annu. Rev. Public Health* **2008**, *29*, 41–55. [CrossRef] [PubMed]
26. Chen, X.; Li, N.; Liu, J.; Zhang, Z.; Liu, Y. Global heat wave hazard considering humidity effects during the 21st century. *Int. J. Environ. Res. Public Health* **2019**, *16*, 1513. [CrossRef] [PubMed]
27. Lemonsu, A.; Viguié, V.; Daniel, M.; Masson, V. Vulnerability to heat waves: Impact of urban expansion scenarios on urban heat island and heat stress in Paris (France). *Urban Clim.* **2015**, *14*, 86–605. [CrossRef]
28. Li, D.; Sun, T.; Liu, M.; Wang, L.; Gao, Z. Changes in wind speed under enhance urban heat islands in the Beijing metropolitan area. *J. Appl. Meteorol. Climatol.* **2016**, *55*, 2369–2375. [CrossRef]

29. Vescovi, L.; Rebetez, M.; Rong, F. Assessing public health risk due to extremely high temperature events: Climate and social parameters. *Clim. Res.* **2005**, *30*, 71–78. [CrossRef]
30. Kim, Y.; Joh, S. A vulnerability study of the low-income elderly in the context of high temperature and mortality in Seoul, Korea. *Sci. Total Environ.* **2006**, *371*, 82–88. [CrossRef]
31. Hajat, S.; Kovats, R.S.; Lachowycz, K. Heat-related and cold-related deaths in England and Wales: Who is at risk? *Occup. Environ. Med.* **2007**, *64*, 93–100. [CrossRef]
32. Bonauto, D.; Anderson, R.; Rauser, E.; Burke, B. Occupational heat illness in Washington state, 1995–2005. *Am. J. Ind. Med.* **2007**, *50*, 940–950. [CrossRef] [PubMed]
33. Spector, J.T.; Bonauto, D.K.; Sheppard, L.; Busch-Isaksen, T.; Calkins, M.; Adams, D. A case-crossover study of heat exposure and injury risk in outdoor agricultural workers. *PLoS ONE* **2016**, *11*, e0164498. [CrossRef] [PubMed]
34. Heo, S.; Lee, E.; Kwon, B.Y.; Lee, S.; Jo, K.H.; Kim, J. Long-term changes in the heat–mortality relationship according to heterogeneous regional climate: A time-series study in Korea. *BMJ* **2016**, *6*, 1–10. [CrossRef]
35. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
36. Yao, Z.; Xu, X.; Yu, H. Floor heating customer prediction model based on random forest. In Proceedings of the 17th International Conference on Computer and Information Science, Singapore, 6–8 June 2018; pp. 573–578.
37. Dang, V.H.; Dieu, T.B.; Tran, X.L.; Hoang, N.D. Enhancing the accuracy of rainfall-induced landslide prediction along mountain roads with a GIS-based random forest classifier. *Bull. Eng. Geol. Environ.* **2019**, *78*, 2835–2849. [CrossRef]
38. Wang, Y.; Song, Q.; Du, Y.; Wang, J.; Zhou, J.; Du, Z.; Li, T. A random forest model to predict heatstroke occurrence for heatwave in China. *Sci. Total Environ.* **2019**, *650*, 3048–3053. [CrossRef]
39. Wang, Z.; Lai, C.; Chen, X.; Yang, B.; Zhao, S.; Bai, X. Flood hazard risk assessment model based on random forest. *J. Hydrol.* **2015**, *527*, 1130–1141. [CrossRef]
40. Deng, M.; Chen, J.; Huang, J.; Niu, W. Agricultural drought risk evaluation based on an optimized comprehensive index system. *Sustainability* **2018**, *10*, 3465. [CrossRef]
41. Alexander, D.L.J.; Tropsha, A.; Winkler, D.A. Beware of R2: Simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *J. Chem. Inf. Model.* **2015**, *55*, 1316–1322. [CrossRef]
42. Wang, W.; Lu, Y. Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model. In Proceedings of the IOP Conference Series: Materials Science and Engineering, Kazimierz Dolny, Poland, 21–23 November 2019.
43. Zikmund, W.G.; Babin, B.J.; Carr, J.C.; Adhikari, A.; Griffin, M. *Business Research Methods: A South Asian Perspective*, 8th ed.; Cengage Learning: Boston, MA, USA, 2013.
44. Kursa, M.B.; Rudnicki, W.R. Feature selection with the Boruta package. *J. Stat. Softw.* **2010**, *36*, 1–13. [CrossRef]
45. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
46. Breiman, L. *Out-of-Bag Estimation*; Citeseer: Berkeley, CA, USA, 1996.
47. Taylor, S.J.; Letham, B. Forecasting at scale. *Am. Stat.* **2018**, *72*, 37–45. [CrossRef]
48. Wu, Z.; Lin, H.; Li, J.; Jiang, Z.; Ma, T. Heat wave frequency variability over North America: Two distinct leading modes. *J. Geophys. Res. Atmos.* **2012**, *117*. [CrossRef]
49. Zhang, K.; Chen, Y.H.; Schwartz, J.D.; Rood, R.B.; O'Neill, M.S. Using forecast and observed weather data to assess performance of forecast products in identifying heat waves and estimating heat wave effects on mortality. *Environ. Health Perspect.* **2014**, *122*, 912–918. [CrossRef] [PubMed]
50. Lee, H.J.; Lee, W.S.; Yoo, J.H. Assessment of medium-range ensemble forecasts of heat waves. *Atmos. Sci. Lett.* **2016**, *17*, 19–25. [CrossRef]
51. Qi, X.; Yang, J. Extended-range prediction of a heat wave event over the Yangtze river valley: Role of intraseasonal signals. *Atmos. Ocean. Sci. Lett.* **2019**, *12*, 451–457. [CrossRef]