Back-propagation in the neural network of visibility estimation model based on Himawari_8 Satellite during forest fire smoke periods on Sumatera and Borneo Island, Indonesia

H Imanto¹,², Hartono² and M A Marfai²

¹Meteorological, Climatological and Geophysical Agency Jakarta, Indonesia
²Faculty of Geography, Universitas Gadjah Mada, Yogyakarta, Indonesia

arismarfai@ugm.ac.id

Abstract. Smoke from forest and land fires may significantly impair horizontal visibility, which affects a wide range of aspects of life, including human health and transportation. Satellite and its remote sensing technology can monitor a target area spatially. Visibility, one of the proxies for smoke quantifiers, has been proposed as the product of a satellite-based model that can benefit human life. This study used back-propagation in neural network (BPNN), a machine learning technology, to develop a visibility estimation model based on The Himawari-8 satellite using several combinations of BPNN tuning. It also compared the estimated visibility estimation with METAR data, as well as root mean square error (RMSE) and R2 correlation to check its accuracy. In this case, visibility was classified into three, namely class 1 visibility (below 1,600 m), class 2 (between 1,600 and 3,000 m), and class 3 (more than 3,000 m). The results showed that the highest accuracy of the visibility estimation model was obtained from the combination of input bands no. 2, 4, 5, 11, 13, 14, 15, with R2 correlation of 0.703.

1. Introduction
The 2015 forest and land fires emitted a more substantial amount of carbon than the ones in 1997 [1]. Severe air pollution occurred during a super (very strong) El Niño in 2015. During the three months of dry conditions associated with El Niño, biomass that was dry enough to be burned turned into large fires and released dangerous air pollution (Hayakasa & Sepriando, 2018). It deteriorated horizontal visibility throughout the airports on Sumatera and Borneo Island, Indonesia, and inevitably disturbed the aviation operation [2]. Not only did it affect transportation, but its presence in the air also degraded air quality and human health [3].

Classification tree analysis (CTA) is a method for classification or prediction that involves recursively partitioning multi-dimensional data until each group is as homogenous as possible [4]. It has been widely used in various fields of study [5][6][7] because of its low computational cost, good accuracy, and highly interpretable tree structures [8]. CTA can also serve as a filtering process in the selection of smoke pixels.

Visibility, the farthest horizontal distance at which selected object can be seen and identified [9], is one of the proxies for quantifying air pollution. Remote sensing technology has been recommended as a means to estimate this proxy [10]. For instance, previous studies on this topic have combined the MODIS product ‘Aerosol Optical Depth (AOD)’ with observation of Aerosol Optical Thickness (AOT) on the ground [11], with Automatic Weather Observing System (AWOS) over the US East Coast [3].
and with numerical weather prediction model [9]. Meanwhile, the accuracy of that models have widely variation, so pure bands of new Himawari-8 are potential as an input to a different visibility modeling with possibly get good accuracy.

The research was designed to develop a visibility estimation model based on Himawari-8 satellite imagery and evaluate its performance using the back-propagation in neural networks (BPNN). The benefit of this research is can be used to quantified the level of air pollution that is related to human life such as transportation safety, human health, or climate studies.

This paper has been arranged in the following order. Section II presents the research data and area, while Section III describes the research methods. Section IV is the results and discussion, containing the accuracy assessment of the visibility estimation model based on the Himawari_8 satellite. Lastly, Section V comprises the conclusion.

2. Research Data and Area

2.1. Data

The data used to develop and assess the accuracy of the visibility estimation model were Himawari_8 satellite data and aerodrome routine meteorological report (METAR) over Sumatera and Borneo Island, Indonesia. Its implemented to the automatic smoke model detection based on findings in [12] and then the visibility estimation model was developed using backpropagation of neural network approach.

This study used Himawari_8 satellite data with a 2-km spatial resolution, 10-minute temporal resolution, and 16 channels. This level 1b data has already been calibrated and validated, meaning that it can be used directly without any further correction [13]. METAR is the information of actual weather observations at an aerodrome made by aeronautical meteorological stations. This information is disseminated to the aerodromes/airports beyond the origin aerodrome. Actual weather conditions, such as wind, visibility, weather phenomena, cloud condition, dewpoint, temperature, and air pressure, are reported in METAR [14]. This research used data samples from both METAR and Himawari-8 satellite data recorded in 2015 from September 1 to 5 and October 1 to 10 at day time. All the data were collected into the matrix dataset after subjected to filtration using CTA E-10 (as a method of selecting smoke pixels) and compared with actual visibility and weather observation from METAR issued by airports in the research area. Then, the dataset was analyzed using principal component analysis to reduce the dimension and dataset complexity. It’s used as an input for developing the visibility estimation model based on BPNN. These data were analyzed in SPSS and Python.

2.2. Research Area

The research area is displayed in Figure 1. Data from 47 airports were used to develop and validate the estimated visibility from the satellite. The visibility estimation model will be detail evaluated to the airspace of Tanjung Pandan airports. Its chosen as the one sample airport due to several reasons, that the following: its located far away from the main sources of forest and land fires and the second the Tanjung pandan island is a small island surrounded by water body. These reasons were related to test the CTA and visibility estimation model performance.
Figure 1. Research Area. White stars are airports. The map of the research area was developed using the base map provided in Phyton and airport location. The red square marks the sample of where the visibility estimation model was implemented (WIKT, the airport of Tanjung Pandan Island)

3. Methods

The research employed (1) smoke classification and separation from undesired backgrounds using the Classification Tree Analysis (CTA) based on the ROI sampling—which have been carried out in Ismanto et al. [12], (2) principal component analysis to reduce the dimensionality of Himawari-8 satellite bands, (3) development of the visibility estimation models using BPNN, (4) calculation of RMSE and $r^2$ (correlation value) to measure the accuracy of the performance of visibility estimation model based on Himawari-8.

As the key ingredient in the development of the visibility estimation model, smoke detection relied on the Classification Tree Analysis (CTA) with Entropy-10. In the previous study, this method was tested for its spatial, logical, and computational performance, and it showed a relatively good result. Smoke appearances and other classes of objects were classified with a high probability of detection (0.53) and overall accuracy (0.78) and supported by a low false alarm rate (0.11) [12]. This research used CTA E-10 mainly to filter smoke pixels. This step produced a dataset containing all pixels of airport locations whose visibility reduced due to smoke. It described the spectral value of Himawari-8 bands of pixels over the airports.

Principal Component Analysis (PCA) is one of the methods that can reduce complex data dimensionality by finding the smallest number of components that explain most of the variation in the original data [15]. If the first few PCs (principal components) explain a large portion of the variability our purpose of dimension reduction has been achieved. Usually, two or three PCs [16] already explain most of the variability of the data. The total variance is more than 80%. Based on this analysis, BPNN input data were selected from the most correlate of the independent variable with the PCs.

The general BPNN algorithm is as follows:
1. Defining the initial value of all weight $w_{ij}^{(l)}$,
2. For $t=0,1,2,\ldots,\text{do}$, is the iteration step,
3. Take $\{1,2,\ldots,N\}$, is the number of neuron in each hidden layer,
4. Forward step: counting all the $x_j^{(l)}$,
5. Backward step: counting all the $\delta_j^{(l)}$,
6. Revising the weight: $w_{ij}^{(l)} <= w_{ij}^{(l)} - x_i^{(l-1)} \delta_j^{(l)}$,
7. Conducting iteration is calculating lowest error value using the gradient descent method,
8. The final weight $w_{ij}^{(l)}$, is the final solution for the problems.

Several treatments were used to carry out the BPNN methods using the scikit-learning module of python software. This paper’s action was modifying the combination of the hidden layer, maximum iteration, learning rate initial, and regularization (to avoiding overfitting). These are to find out the best BPNN model to estimate visibility (Table 2). The METAR visibility data were split into several categories, namely class 1 (visibility = below 1,600 meters), class 2 (from 1,601 to 3,200 meters), and class 3 (more than 3,201 meters, but not exceeding 4,800 meters). These categories are based on the thresholds set in the flight rules for aviation operation. All these data were collected into one dataset that contained 16 bands, visibility, and categories of visibility. Sixty-five percent of the data were used for training, while the other thirty percent were for testing.

4. Results and Discussions
The results of this study were (1) the statistical exploration of the dataset, including PCA results and analysis and (2) the BPNN results and discussion.

4.1. Statistical dataset exploration
Before the BPNN process, the statistical PCA was used to reduce/make an alternative input variable that potentially created better BPNN results and performance (Table 1). Table 1 shows the PCA results of the dataset. The first alternative input was using cumulative of total variance more than 80% of the variability of the dataset. Cumulative of total variance reaches more than 80 % by three principal components. The independent variable that have a high correlation with these three PCs were B1,2,3,4,5,6,9,10,11,13,14,15,16. The second alternative was using a cumulative of total variance more than 60 % of the variability of the dataset. It consisted of the first two PCs and the independent variable that have a high correlation were B2,4,5,11,13,14,15. These two scenarios were based on the correlation values of more than 0.7 and 0.8 with the dominant PCs (as representative of all the variability of the dataset). Another scenario was using all the Himawari-8 bands as input for visibility estimation using BPNN.
Table 1. Total Variance of the dataset using the PCA method and First Three Principal Component Matrix

| Component | Initial Eigenvalues | Component Matrix |
|-----------|---------------------|------------------|
|           | Total | % of Variance | Cumulative % | Band 1 | 2 | 3 |
| 1         | 5.641 | 35.258        | 35.258        | 1 |-0.378 | 0.776 | 0.24 |
| 2         | 5.076 | 31.728        | 66.986        | 2 |-0.372 | 0.800 | 0.237 |
| 3         | 2.181 | 13.628        | 80.614        | 3 |-0.381 | 0.765 | 0.25 |
| 4         | 1.295 | 8.095         | 88.709        | 4 |-0.164 | 0.897 | 0.041 |
| 5         | 0.525 | 3.283         | 91.992        | 5 |-0.217 | 0.876 | 0.025 |
| 6         | 0.503 | 3.142         | 95.133        | 6 |-0.250 | 0.787 | 0.089 |
| 7         | 0.326 | 2.035         | 97.168        | 7 |0.072  | 0.548 | -0.165 |
| 8         | 0.143 | 0.893         | 98.062        | 8 |0.635  | -0.133 | 0.649 |
| 9         | 0.131 | 0.820         | 98.881        | 9 |0.653  | -0.103 | 0.732 |
| 10        | 0.065 | 0.403         | 99.285        | 10 |0.629 | 0.005 | 0.735 |
| 11        | 0.041 | 0.254         | 99.539        | 11 |0.884 | 0.182 | -0.243 |
| 12        | 0.026 | 0.160         | 99.698        | 12 |0.51  | 0.427 | -0.485 |
| 13        | 0.017 | 0.105         | 99.803        | 13 |0.845 | 0.245 | -0.349 |
| 14        | 0.015 | 0.096         | 99.900        | 14 |0.878 | 0.305 | -0.22 |
| 15        | 0.009 | 0.056         | 99.955        | 15 |0.88 | 0.360 | -0.046 |
| 16        | 0.007 | 0.045         | 100.00        | 16 |0.74 | 0.481 | -0.056 |

4.2. The BPNN results for visibility estimation model estimation

The results of the BPNN process using the combination of all bands as the input with hidden layers combination: 16 layers on the left and 3 layers on the right; MI = 200; alpha=0.00001; LRI=0.07 showed that the correlation value (r²) of the model result was 0.75 and the RMSE was 0.67. These results were better than the BPNN result with input B (table 2, column B). And the highest accuracy value of the BPNN model was using input C.

The third BPNN results from the selected bands, which used PC1,2,3 as the input, showed an increase in the model accuracy, as evident from R² = 0.75 and the RMSE=0.56. This achieved at a layer setting 17,6 in the hidden layer with 250 iterations. The results of the BPNN modeling depend on the trial and error in the model tuning. The efficient variable, less hidden layer, is not always making bad results than huge hidden layer amounts. The key idea of the neural network is to find out the best combination of model tuning through neural network algorithms. Even though the results of this paper have a good correlation, but the bias error of the RMSE result was higher than the standard accuracy allowed.

Table 2. BPNN results for several modified input

| Hyper-parameter Combination BPNN | All Himawari 8 bands (A) | The Band have correlation value with PC1 and PC2 (B) | The Band have correlation value with PC1 and PC3 (C) |
|--------------------------------|--------------------------|----------------------------------------------------|----------------------------------------------------|
| R² | RMSE | R² | RMSE | R² | RMSE |
| HL(7,5),MI(200),alp(0.001),LRI(0.08) | 0,65 | 0,78 | | | |
| HL(16,3),MI(200),alp(0.00001),LRI(0.07) | 0,75 | 0,67 | | | |
| HL(17,6),MI(250),alp(0.00001),LRI(0.06) | 0,75 | 0,56 | | | |
4.3. The BPNN results implementation over a small airport
Spatial visibility estimation using the BPNN approach was implemented to the small airport in Tanjung Pandang Island. Its located in the ocean between Sumatera and Borneo Island, a far away area from the fire hotspot sources. The visibility model estimation shows an inline result with the condition of actual horizontal visibility over WIKT airport. Prevailing visibility reported value was 1500 meters due to smoke. Estimation visibility model from Himawari-8 using BPNN methods shows more than 50 % of the airspace of the WIKT aerodrome covered by an area of low visibility (visibility less than 1600 meters), that concentrate along south to the north-east area of the aerodrome. According to the Annex 3 aeronautical meteorology for international air navigation [14], prevailing visibility is “The greatest visibility value, observed in accordance with the definition of ‘visibility’, which is reached within at least half the horizon circle or within at least half of the surface of the aerodrome. These areas could comprises contiguous or non-contiguous sectors.”

Figure 3. Spatial performance of visibility estimation model over WIKT Airport on October 24, 2015, 05.00 UTC (13.00 local time)

5. Conclusion
The backpropagation of the neural network approach of the categorical visibility estimation model has a correlation value reach 0.75 with bias error RMSE 0.56. The best combination Himawari-8 bands as input were band 1,2,3,4,5,6,9,10,11,13,14,15,16. The main concept of BPNN is modifying/tuning the neural network hyperparameter as efficient as possible. Even though the categorical visibility Hima_8-BPNN model error is higher than international standard accuracy allowed, it can gift the spatial visibility situation along aerodrome. This result is very useful for situational awareness for all aviation stakeholders, such as aeronautical meteorological service providers, air navigation service providers, aviation operators, and aerodrome operators. It’s very useful as an effort to achieve the safety of aviation operation.

Acknowledgments
This paper is funded by Hibah-RTA and is one of the results of the research in the doctoral program at the Faculty of Geography, Universitas Gadjah Mada.

References
[1] Huijnen V, Wooster M J, Kaiser J W, Gaveau D L, Flemming J, Parrington M and van Weele M 2016 Fire carbon emissions over maritime southeast Asia in 2015 largest since 1997
[2] Ismanto H, Hartono H and Marfai M A 2019 Spatiotemporal visibility characteristics impacted by
The 3rd Environmental Resources Management in Global Region

IOP Conf. Series: Earth and Environmental Science 451 (2020) 012010
doi:10.1088/1755-1315/451/1/012010

...forest and land fire over airports in Sumatera and Borneo Island, Indonesia Quaest. Geogr. 38
1–12

[3] Kessner A L, Wang J, Levy R C and Colarco P R 2013 Remote sensing of surface visibility from
space: A look at the United States East Coast Atmos. Environ. 136–47

[4] Thuiller W, Araujo M B and Lavoral S 2003 Generalized models vs. classification tree analysis:
Predicting spatial distribution of plant species at different scales J. Veg. Sci. 669–80

[5] Dominick G M, Papas M A, Rogers M L and Rakowski W 2015 Classification tree analysis to
examine influences on colorectal cancer screening Cancer Causes & Control pp 443–54

[6] Shaaban K and Pande A 2016 Classification tree analysis of factors affecting parking choices in
Qatar Case Study Transp. Policy 88–95

[7] Ismanto H, Hartono H and Marfai M A 2019 Classification tree analysis (Gini-Index) smoke
detection using Himawari_8 Satellite data over Sumatera-Borneo Maritime Continent South
East Asia IOP Conference Series: Earth and Environmental Science Series: Earth and
Environmental Science pp 1–9

[8] Loh W Y 2014 Fifty years of classification and regression trees Int. Stat. Rev. 329–48

[9] Brunner J, Pierce R B and Lenzen A 2016 Development and validation of satellite-based estimates
of surface visibility Atmospheric Measurement Techniques pp 409–22

[10] Hyslop N P 2009 Impaired visibility: the air pollution people see Atmos. Environ. 43 182–95

[11] Nichol J E, Wong M S and Wang J 2010 A 3D aerosol and visibility information system for urban
areas using remote sensing and GIS Atmos. Environ. 2501–6

[12] Ismanto H, Hartono H and Marfai M A 2020 Visibility estimation due to forest fires smoke using
backward elimination multiple regression of Satellite Himawari-8 data over Sumatera and
Borneo Islands Indonesia IOP Conference Series: Earth and Environmental Science
(Yogyakarta) pp 1–9

[13] Bessho K, Date K and Hayashi M 2016 An Introduction to Himawari 8/9-Japan’s New-
Generation Geostationary Meteorological Satellites J. Meteorol. Soc. Japan 151–83

[14] International Civil Aviation Organization (ICAO) 2015 Annex 3 Aeronautical Meteorological
Services for International Air Navigation Seventy (Montreal: International Civil Aviation
Organization (ICAO))

[15] Jackson B and Bund B 1983 Multivariate Data Analysis: An Introduction (New York: McGraw-
Hill)

[16] Masoumeh K D and Amir E 2016 Use of principal component analysis in accuracy of
classification maps (case study: North of Iran) Res. J. For. 10 2016)