Classification Tree Analysis (Gini-Index) Smoke Detection using Himawari_8 Satellite Data Over Sumatera-Borneo Maritime Continent Sout East Asia

H Ismanto¹,², Hartono² and M A Marfai²
¹Center of Aeronautical Meteorology, Badan Meteorologi Klimatologi dan Geofisika, Indonesia
²Faculty of Geography, Universitas Gadjah Mada, Indonesia
heri.ismanto@bmkg.go.id

Abstract. Classification tree analysis (CTA) automatic smoke detection was proposed using Himawari_8 Satellite data over Sumatera and Borneo Island Maritime Continent. Day Natural color and aerosol RGB composite were used to make Region of Interest (ROI) sampling of Cumulonimbus (Cb) top, low-mid cloud, smoke, bare soil, cirrus cloud, vegetation, and water. CTA – Gini index supervised classification being constructed with two different band collection as input. The result shows that CTA model 2, using 21 bands collection as input, has better overall accuracy value (about 0.75). Then the CTA model 1 only has an overall accuracy of about 0.63. The future research is still open in comparing the different impurity methods of CTA model, ensemble model of smoke detection using several CTA output models, and also a diurnal or seasonal variation of smoke using CTA model detection.

1. Introduction
Smoke, a specific Indonesian hazard, is becoming an international interest due to spreading out over South East Asian neighboring countries besides Indonesian Area [1–5] and has a negative impact over a various field such as health, economy, tourism and transportsations [6]. The worst transportation disaster ever in smoky condition was Garuda Indonesia (GA152) September 26th, 1997 flight accident. All the passengers were killed due to flight crash in the mountainous area near Medan International Airport [7].

Dry pollutant, smoke, and haze is responsible for visibility restriction that should be considered in term of impacts on safety [8]. Haze is an atmospheric condition where dust, smoke and other dry particles (including fires) obscuring the sky [9]. Regional haze is visibility impairment caused by the cumulative air pollutant emissions from numerous sources over a wide geographic area [8]. The haze particles are so small that they cannot be seen individually, but are still effective in visual range restriction. Smoke has relatively large particles than haze due to the near distance from the sources.

Satellite-based smoke detection methods are divided into two approaches: first was visual approach using RGB composite [10,11] and the second was automatic smoke detections [12-17]. The visual method depends on subjective interpretation. Then the automatic smoke detections are more objective such as Multi-thresholds [12,13]; combinations of Multi-threshold and K-means cluster [14]; multispectral transformation [15] and Neural Network [16,17]. Other statistical methods widely used over fifty years among variate subject [18], Classification Tree Analysis (CTA) be a potential
alternative algorithm to be implemented in the Indonesian tropical region. It used such machine learning to choose the best of the predictor. The best predictor is chosen using a variety of impurity or diversity measures (Gini, towng, ordered two and least–squared deviation) [19,2].

Himawari_8 satellite data have high temporal, spatial and spectral resolution other than previous Japan satellite generation. Every single image has their own variation data of single objects/classification classes. Spectrally, four times more channel equipped in his advance high instrument (AHI) other than MTSAT2 and it can gift an opportunity to make detection of an object more accurately. To optimize the smoke detection using Himawari_8 data over Indonesia area, this research purposing combination of single-channel and multichannel transformation as an input to produce decision tree analysis of Gini Index– CTA.

This research objective is to find out the smoke automatic detection using classification tree analysis (CTA) over Borneo and Sumatera, Tropical Maritime Continent Indonesia area using Himawari_8 satellite data. Comparing the two different input to build CTA-Gini Index model to get the best fit spatially and statistically model to be implemented for automatic smoke detection over Sumatera and Borneo Island Indonesia. The advantages of this research support another research related to climate change, weather forecasting, risk and disaster management, remote sensing, health, and air pollution.

This paper arranges in the following section: data used and methods are introduced in Section II, section III result and discussion contain: RGB day natural color and aerosol analysis, CTA-Gini index classification result and validation. The performance of the algorithm is evaluated using airport observation (METAR) that also addressed in this section. Section IV describes the conclusion of the paper result.

2. Research Methods
Himawari_8 data were used in this study for the development of the Classification Tree Analysis (CTA) based smoke detection model. METAR airports observation were used for validation of the model. The Himawari 8 satellite is at about 35,800 km above the equator, at 140.7 E longitude. Our analysis is based on Himawari 8 satellite data and meteorological observations and reports over 47 airports for 14 October 2015 at 02.00 UTC as a sample study. Himawari 8 data provided by Meteorological Climatological and Geophysical Agency of Indonesia (BMKG) in SATAID format with 2 km x 2 km spatial resolution. Airport locations and observation data were accessed from http://aviation.bmkg.go.id/.

Himawari 8 data supplied by JMA to BMKG via Himawari Cloud services was used in this study. This level 1b of Himawari 8 satellite data was already radiometric and geometric corrected [21] with 2 km spatial resolution, 10 minutes temporal resolution and 16 channels. The focus area of this research was Sumatera and Borneo Island, Maritime Continent, South East Asia.

Meteorological airport observation (METAR) is aerodrome routine meteorological report (in meteorological code) [22]. This message is mandatory for every airport in the world, especially that conduct international air navigation. It contains meteorological parameters such as actual weather and prevailing visibility. This two meteorological surface parameter will be used as the validation of the CTA smoke automatic detection. Each pixel from CTA smoke automatic detection from Himawari_8 satellite will be validated using actual smoke observation (airport locations).

2.1. RGB and Multispectral analysis
The RGB used in this research are RGB aerosol composite, purposed by Daniel Rosenfeld & Lensky [22] (visible 0.6 µm in Red beam (R), 0.8 µm in Green beam (G) dan IR 10.4 µm (inverse) in Blue beam (B)); RGB natural color (1.6 µm (R); 0.8 µm (G); 0.6 µm (B)) and day microphysics (0.86 µm reflectance (R); 3.9 µm reflectance (G); 10.4 µm radiance (B)). The sampled ROI divide into 7 classes which are class 1: Cb Top, class 2: small cloud at a low and mid-level, class 3: smoke, class 4: bare soil, class 5: thin cirrus, class 6: vegetation, class 7: water.
Figure 1 Feature space of ROI sampling in each class. Three band was the component of RGB aerosol proposed by Rosenfeld and Lensky [22]. Sampling data was taken from October 14, 2015, 02.00 UTC of the Himawari_8 satellite.

The region of interest (ROI’s) feature space revealed in Figure 1. Cumulonimbus cloud (red dot), Water (blue dot) and cirrus cloud over water (purple dot) separated well. In a minor amount of pixel of bare soil (black dot) embedded with vegetation (green dot) and smoke (yellow dot) pixel. Low-Mid cloud (cyan dot) pixel class separated well with smoke, except a few pixels. Figure 1 shows the spectral characteristic of the reflectance and brightness temperature of dense smoke have intermediate values between the cloud and the underlying surface. The reflectance of smoke is usually less than the cloud, but higher than the underlying surface, while the converse is true for brightness temperature. Smoke is hotter than cloud (Cumulonimbus, low-mid cloud and cirrus), but colder than land (especially bare soil).

Data input which used in CTA analysis contains all 16 bands of Himawari_8 satellite data and additional transformation band that proposed by X. Li et al. [17], revealed in Table 1. Two scenarios band collection were used. The first input (Input (1)) was a collection of single and transformation bands proposed by X. Li et al. [17], which implemented in the multi-threshold smoke detection method. The second one (Input (2)) was band collection contains 21 bands (all 16 bands of Himawari_8, inverse B13, and X. Li et al [17] band transformations). Band Transformation T4 and T5 were used to detect smoke, T6 was used to detect cloud and T7 (NDVI) to detect vegetation. These treatment purposes were used to find out the response result and the best threshold (CTA models) from different input bands collection.
2.2. **Classification Tree Analysis (CTA)**

Classification tree analysis, statistically methods used more than fifty years, was used to distinguished a sample of data into several nodes/class based on splitting. It consists of recursive partitions of the dimensional space defined by the predictors into groups that are as homogeneous as possible in terms of response, resulting in a hierarchy of groups within groups. The hierarchy is called a tree, and each group is called a node. The final or terminal nodes called leaves. One splitting algorithm (impurity functions) been used is Gini index, \( \Phi(t) = 1 - \sum_j P^2(j|t) \), where \( P(j|t) \) is the proportion of class j observations in node t (18,19,25). The trees output from this CTA showed in Figure 2. It was used to build the smoke detection. Both results next to be compared spatially and its accuracy.

| Wavelength (\( \mu m \)) | Band / Spectral Transformation | Input (1) | Input (2) | Physics Properties |
|--------------------------|--------------------------------|-----------|-----------|-------------------|
| 0.47                     | × (T1)                         | × (B1)    |           | Vegetation,aerosol |
| 0.51                     | × (T2)                         | × (B2)    |           | Vegetation,aerosol |
| 0.64                     | × (B3)                         | × (B4)    |           | Low cloud, fog    |
| 0.86                     | × (B5)                         | × (B6)    |           | Vegetation,aerosol |
| 1.6                      | × (B7)                         |           |           | Low cloud, fog, forest fire |
| 2.3                      | × (B8)                         | × (B9)    |           | Mid and upper-level Moisture |
| 3.9                      | × (B10)                        |           |           | Mid-level moisture |
| 6.2                      | × (B11)                        |           |           | Cloud, SO2 |
| 6.9                      | × (B12)                        |           |           | Ozone content |
| 7.3                      | × (B13)                        |           |           | Cloud top information |
| 8.6                      | × (B14)                        |           |           | Cloud imagery, SST |
| 10.4                     | × (B15)                        |           |           | Cloud imagery, SST |
| 11.2                     | × (B16)                        |           |           | Cloud top height |
| 12.4                     | × (T3)                         | × (B17)   |           | Cloud top information |
| 13,3                     | × (T4)                         | × (B18)   |           | Smoke/aerosol |
| 13 (Inverse)             | × (B19)                        |           |           | Smoke/aerosol |
| (Li 2015).1 (B1-B4)/(B1+B4) | × (T5)                        | × (B20)   |           | Cloud |
| (Li 2015).2 (B1-B6)/(B1+B6) | × (T6)                        |           |           | Vegetation |
| NDVI (B4-B3)/(B4+B3)     | × (T7)                         | × (B21)   |           | Vegetation |

2.2. **Classification Tree Analysis (CTA)**

Classification tree analysis, statistically methods used more than fifty years, was used to distinguished a sample of data into several nodes/class based on splitting. It consists of recursive partitions of the dimensional space defined by the predictors into groups that are as homogeneous as possible in terms of response, resulting in a hierarchy of groups within groups. The hierarchy is called a tree, and each group is called a node. The final or terminal nodes called leaves. One splitting algorithm (impurity functions) been used is Gini index, \( \Phi(t) = 1 - \sum_j P^2(j|t) \), where \( P(j|t) \) is the proportion of class j observations in node t (18,19,25). The trees output from this CTA showed in Figure 2. It was used to build the smoke detection. Both results next to be compared spatially and its accuracy.
Figure 2 Classification tree Himawari_8 results. The sample was taken October 14, 2015, 02.00 UTC, using the Gini Index impurity method. A) based on 7 band collection as input, B) based on 21 band collection as input.

2.3. Model validation
The validation of the smoke detection is comparing the classification pixels of the CTA model in the same location with the airport and the airport's actual weather observation. By using “extract by multiple points” of a GIS software tool, the classification Himawari_8 satellite pixels was extracted to be a table. This dataset then to be compared with the airport's actual weather observation. Overall accuracy was used as the quantification method. The value range is 0 to 1 as the highest value that means the accuracy is 100%.

3. Result and Discussion
Figure 3 shows the RGB composite day natural color (A) and RGB aerosol (B) overlaid with the smoke assignment from Himawari_8 satellite data. Smoke cannot see clearly using RGB natural day color. Other, RGB aerosol showed that: ground seen in green, sea surface was dark, and smoke
appeared as the fuzzy brown features over the sea background or yellow fuzzy features over land. Smoke was intense in the southern part of Borneo and was less in the north of Borneo Island. Smoke has a smooth texture but cloud look more undulating and separated in patches. It because cloud tops of convective have had strong vertical development that contains a lot of ice particle, while a high variation of water vapor and vertical updraft convection makes small low cloud looks in patches. Clouds were white and light blue for the cold tops and cirrus.

![RGB composite: Day Natural Color (A), Aerosol Rosenfeld & Lensky (B). Satellite data on October 14, 2015, 02.00 UTC.](image)

Figure 3. RGB composite: Day Natural Color (A), Aerosol Rosenfeld & Lensky [23] (B). Satellite data on October 14, 2015, 02.00 UTC.

Figure 4 shows the CTA – Gini Index result based on different input band collection. CTA- Gini index with 7 input band collection spatially showed the overestimates result. Wide area, including the South China Sea; Java Sea; Hindian Ocean, were assigned as smoke. Another CTA-Gini index, using 21 bans collection, spatially showed the underestimates result. It shows only a few smoke assigned area that visually seen like revealed in the RGB aerosol in figure 3 (lower).

| Table 2 Contingency table and overall accuracy calculation |
|----------------------------------------------------------|
| Smoke Himawari_8 Classifications | Yes | No |
| Smoke Airport Observation | Yes | A | B |
| No | C | D |

A, B, C, and D denote the number of occurrences for each case. Overall accuracy is defined as Overall Accuracy = (A+D)/(A+B+C+D) (26)

Validation research methods here was using overall accuracy. “Miss” and “Hit” of smoke detection calculate using contingency table (Table 2). Predicted class by CTA-Himawari_8 counted as ‘Miss’
when vegetation/bare soil revealed and the actual observation over the airport is smoke obscuration or smoke revealed in CTA results but no smoke in actual airport/s observation. And counted as ‘Hit’ if predicted class and the airport observation same result, smoke exists or no smoke exist both in CTA result and airport/s observation. This validation not counted if the cloud revealed in the CTA model of Himawari_8 satellite data because the existence of a cloud is blocking the smoke information. The CTA model 1 (using 7 transformation bands as input) and CTA model 2 (using 21 of combination pure and transformation bands were compared using overall accuracy. The results show that CTA model 2 was more accurate other than CTA model 1. The overall accuracy of CTA model 2 is 75%. Then the overall accuracy of the CTA model 1 is 62.8%.

![Image](image_url)

**Figure 4.** CTA – Gini Index result using two input band collection of Himawari_8 satellite data. (A) CTA – Gini Index result based on 7 collection bands (input 1) based on X. Li et al. (17); (B) CTA – Gini Index result based on 21 collection bands (input 2).

### 4. Conclusions

CTA – Gini index is a potential method of automatic smoke detection. The result of this method depends on several factors, that are: RGB analysis to get ROI sampling, data band collection input, and the impurity methods are chosen (gini, entropy or ratio). The CTA model 2 is a potential method of an automatic smoke detection due to relatively high accuracy. This result very useful to be implemented in several topics such as meteorology, climatology, healthy, transportation and pollution monitoring due to low computation effort in high temporal satellite data. The future research is still open in comparing the impurity methods of CTA model, ensemble model of smoke detection using several CTA output models, and also a diurnal or seasonal variation of smoke using CTA model detection.
Acknowledgments
Special thanks to Dr. Retnadi Heru Jatmiko and Restu from Faculty of Geography UGM to the valuable discussion for the analyzing methods of these research. And Mr. Andersen Panjaitan center of Public Meteorology BMKG Indonesia for the WIS Himawari_8 data used in this paper.

References

[1] Aouizerats B van der Werf G R Balasubramanian R and Betha R 2015 Importance of transboundary transport of biomass burning emissions to regional air quality in Southeast Asia during a high fire event Atmos Chem Phys. 15(1):363–73.
[2] Behera S Betha R Huang X and Balasubramanian R 2015 Characterization and estimation of human airway deposition of size-resolved particulate-bound trace elements during a recent haze episode in Southeast Asia. Environ Sci Pollut Res. 22(6):4265–80. Available from: http://dx.doi.org/10.1007/s11356-014-3645-6.
[3] Gaveau D L a Salim M a Hergoualc’h K Locatelli B Sloan S Wooster M et al. 2014 Major atmospheric emissions from peat fires in Southeast Asia during non-drought years: evidence from the 2013 Sumatran fires Sci Rep 4:1–7.
[4] Reid J S Hyer E J Johnson R S Holben B N Yokelson R J Zhang J et al. 2013 Observing and understanding the Southeast Asian aerosol system by remote sensing: An initial review and analysis for the Seven Southeast Asian Studies (7SEAS) program Atmos Res 122:403–68. Available from: http://dx.doi.org/10.1016/j.atmosres.2012.06.005.
[5] Radojevic M 2003 Chemistry of Forest Fires and Regional Haze with Emphasis on Southeast Asia Pure Appl Geophys. 160:157–87.
[6] Murdiyarso D Lebel L Gintings a N Tamubolon S M H Heil A and Wasson M 2004 Policy responses to complex environmental problems: Insights from a science-policy activity on transboundary haze from vegetation fires in Southeast Asia Agric Ecosyst Environ. 104:47–56.
[7] Aiken S R 2004 Runaway fires, smoke-haze pollution, and unnatural disasters in Indonesia Geogr Rev. 94:55–79
[8] Achtemeier G Jackson B and Brenner J D 2001 Smoke Management Guide for Prescribed and Wildland Fire (National Wildfire Coordination Group)
[9] Deza M M and Deza E 2013 Encyclopedia of Distance (Second Edi)
[10] Chung Y S and Kim H S 2008 Satellite monitoring of forest fires and associated smoke plumes occurring in Korea Air Qual Atmos Heal. 1:111–8.
[11] Lensky I M and Rosenfeld D 2008 Physics Clouds-Aerosols-Precipitation Satellite Analysis Tool (CAPSAT) Atmos Chem Phys. 6739–53.
[12] Xie Y Qu J J Xiong X Hao X Che N and Sommers W 2007 Smoke plume detection in the eastern United States using MODIS. Int J Remote Sens. 28(10):2367–74. Available from: http://www.tandfonline.com/doi/abs/10.1080/01431160701236795
[13] Wang W Qu J J Hao X Liu Y and Sommers T 2007 An improved algorithm for small and cool fire detection using MODIS data: A preliminary study in the southeastern United States Remote Sens Environ. 108:163–70.
[14] Wang J Song W Wang W Zhang Y and Liu S 2011 A new algorithm for forest fire smoke detection based on MODIS data in Heilongjiang province IEEE 5–8.
[15] Nagatani I Kudoh J and Kawano K 2012 A new technique for visualization of forest fire smoke plume using MODIS data Tohoku Institute of Technology. IEEE 2380–3.
[16] Li Z Q Khananian A Fraser R H and Cihlar J 2001 Automatic detection of fire smoke using artificial neural networks and threshold approaches applied to AVHRR imagery IEEE Trans Geosci Remote Sens. 39(9):1859–70.

[17] Li X Song W Lian L and Wei X 2015 Forest Fire Smoke Detection Using Back-Propagation Neural Network based on Modis Data Remote Sens. 7(100501):4473–98.

[18] Loh W-Y 2014 Fifty Years of Classification and Regression Trees Int Stat Rev 82(3):329–48. Available from: http://doi.wiley.com/10.1111/insr.12016.

[19] Shaaban K and Pande A 2016 Classification tree analysis of factors affecting parking choices in Qatar. Case Stud Transp Policy 4(2): 88–95. Available from: http://dx.doi.org/10.1016/j.cstp.2015.11.002.

[20] Kurt I Ture M and Kurum A T 2008 Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease Expert Syst Appl. 34(1):366–74.

[21] Bessho K Date K and Hayashi M 2015 An Introduction to Himawari-8/9 — Japan ’ s New-Generation Geostationary Meteorological Satellites J Meteorol Soc Japan 2 (December 2015). Available from: https://www.jstage.jst.go.jp/article/jmsj/94/2/94_2016-009/_article.

[22]  ICAO Doc 8896 2011 Manual of Aeronautical Meteorological Practice. (Ninth Edit. Montreal, Quebec, Canada).

[23] Rosenfeld D and Lensky I M 1998 Satellite-Based Insights into Precipitation Formation Processes in Continental and Maritime Convective Clouds Bull Am Meteorol Soc. 79 (November 1998).

[24] Rosenfeld D and Lensky IM 1998 Satellite-Based Insight into Precipitation Formation Processes in Continental and Maritime Convective Clouds Bull Am Meteorol Soc. 79(November 11).

[25] Li X and Claramunt C 2006 A Spatial Entropy-Based Decision Tree for Classification of Geographical Information Trans GIS 10(3):451–67. Available from: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1467-9671.2006.01006.x

[26] Ismanto H Hartono H and Marfai M 2018 Smoke detections and visibility estimation using satellite himawari_8 data over sumatera and borneo island indonesia Spat. Inf. Res. Springer p1-12