Estimation of cloud regions in all-sky images by machine learning

Kanta Shimizu[1], Minrui Wang[2], Takashi Y. Nakajima[2]*

[1] *Tokai University, Graduate School, Course of Electrical and Electronic Engineering*

[2] *Tokai University, Research and Information Center*

Global warming, an urgent issue in human society, is progressing. This warming is caused by the increase in greenhouse gases such as carbon dioxide. However, to accurately forecast the future of global warming, it is important to understand the behavior of clouds, which are involved in the Earth's energy balance. According to a paper by Randall et al. (1984), it has been shown that an increase of 4% in the amount of low-level clouds can offset the warming effect equivalent to twice the pre-industrial revolution concentration of carbon dioxide. Thus, this study focuses on cloud cover, one of the parameters of clouds.

In this study, we aim to eliminate observer subjectivity in observational results by objectively observing cloud cover from ground-based all-sky camera and to automate cloud cover observation from the ground. For this purpose, we conducted a verification on the structure of a training model to be used in a machine learning model for developing an algorithm that automatically determines cloud cover from images.

The all-sky camera images were sourced from the TACPAS system owned by Tokai University. TACPAS employs a camera equipped with a 2-pi fisheye lens, capturing the sky every five minutes, 24 hours a day, 365 days a year. Even at night, the illumination from the city lights of Tokyo allows us to discern clouds, making the all-sky camera installed in Tokyo an ideal choice for our research.

The training model stored pixel information and overall image information as explanatory variables, and stored Flag (0: sky, 1: cloud) as the target variable. We used the logistic regression and the XG-Boost in the machine learning model. The logistic regression excels in linear problems, while the XG-Boost excels in nonlinear problems.

As a result, both models were effective and could identify clouds in images both day and night. The logistic regression was effective for images without the sun area during the day and for nighttime images. The XG-Boost was effective for images with the sun area during the day, without the sun, and for nighttime images. In other words, it is considered that when the sun is not in the image, the problem can be approached as either a linear or non-linear one, and when the sun is in the image, it should be approached as a non-linear problem.

From these results, it was found that machine learning using pixel information as a model is an effective method for identifying cloud areas in images. Looking forward, we believe that this method can be utilized to validate the cloud amount data obtained from satellite imagers, such as the GCOM-C SGLI, and then contribute to improving climate models used in simulations.

Keywords: All-sky images, Cloud cover, Machine learning