Automated cloud top detection using machine learning algorithm on radio occultation profiles

R. Biondi [1], M. Hammouti [2], P.-Y. Tournigand [3]

(1) Universitá degli Studi di Padova, Dipartimento di Geoscienze, Padova, Italy (riccardo@biondiriccardo.it)
(2) National Research Council, IGAG, Laboratory of Risk Analysis and Emergency Management , Milano, Italy
(3) Physical Geography (FARD), Department of Geography, Vrije Universiteit Brussel, Brussels, Belgium

VESUVIO Project

Abstract

The bending angle anomaly, retrieved from radio occultation profiles, has been widely applied in previous research to determine the cloud top height of tropical cyclones and more recently the top height of volcanic clouds. This methodology has proven to be very accurate in altitude estimations, but it is time consuming. The objective of this work is to develop an algorithm able to acquire near real time RO profiles and to automatically determine the height of dense clouds in the upper troposphere and lower stratosphere.

In this work we collocated the GNSS radio occultations with volcanic clouds and “non eruptive” data to build a statistically relevant training dataset for the machine learning algorithm to detect the cloud. The algorithm receives in input the cloud detection from different instruments and different atmospheric parameters profiles, learns the atmospheric parameter variations as a function of the altitude and how to identify the discontinuities due to the presence of a cloud, and provides in output an estimation of the presence of the volcanic cloud. The dataset includes about four thousand collocations between RO and clouds in the period 2008-2015. From the entire dataset we have randomly selected 80% of the collocations used as training, 20% used as testing, and applied the cross-validation technique in order to validate the model performance to an independent dataset that was not used in estimating it. The output of the algorithm will contribute to the development of an already existing early warning system. As a next step we will use the cloud top height estimation from the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) backscatter and train the algorithm to estimate it. The same methodology will be soon applied also to the tropical cyclones.

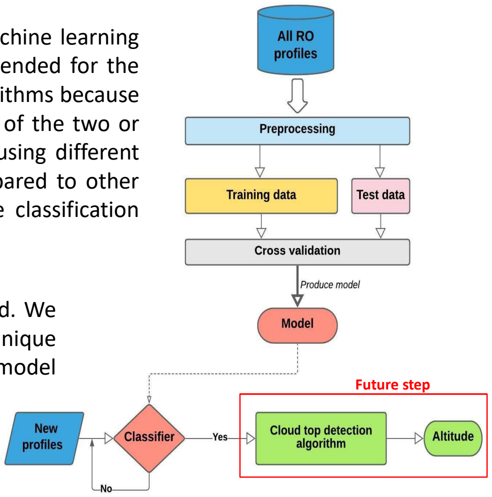
Methodology

Support Vector Machines (SVMs) are a set of supervised learning methods used for common tasks in machine learning (e.g., classification, regression, and outliers detection). The use of SVMs for classification problems is intended for the binary classification setting in which there are two classes. SVMs are different from other classification algorithms because of the way they choose the decision boundary that maximizes the distance from the nearest data points of the two or more classes. Moreover, they are versatile, as they can be applied for both linear and non-linear data, using different kernel functions specified for the decision function. SVMs usually find more accurate results when compared to other classification algorithms because of their ability to handle smaller and complex datasets. However, the classification accuracy can be improved by increasing samples number.

Training algorithm

The datasest consists in two classes of RO profiles: ROs collocated with volcanic clouds and not co-located. We have considered 80% of datasets as training set, and 20% as test set. We apply the cross validation technique (stratified k-fold method) in order to control problems like overfitting, and generally, for assessing the model performance in classifying new data (profiles) that were not used in the estimation.

We have iterated various models (kernel functions) of SVM to find a better performing model considering linear, polynomial (2° and 3° degree), and Radial Basis Function.



All RO profiles

Preprocessing

Training data

Test data

Cross validation

Produce model

Model

Future step

New profiles

Classifier

Yes

No

Cloud top detection algorithm

Altitude

Cloud top detection (on-going analysis)

The SVM model has been chosen as classifier to sort new anomaly profiles, which may be collocated or not with volcanic clouds. If there is a profile collocated with the volcanic cloud, a secondary automated algorithm is used to detect the cloud top height from this profile. Particularly, the algorithm detects the peak with maximum amplitude with respect to the local minimum (just before the peak) related to the volcanic cloud top height.

Results

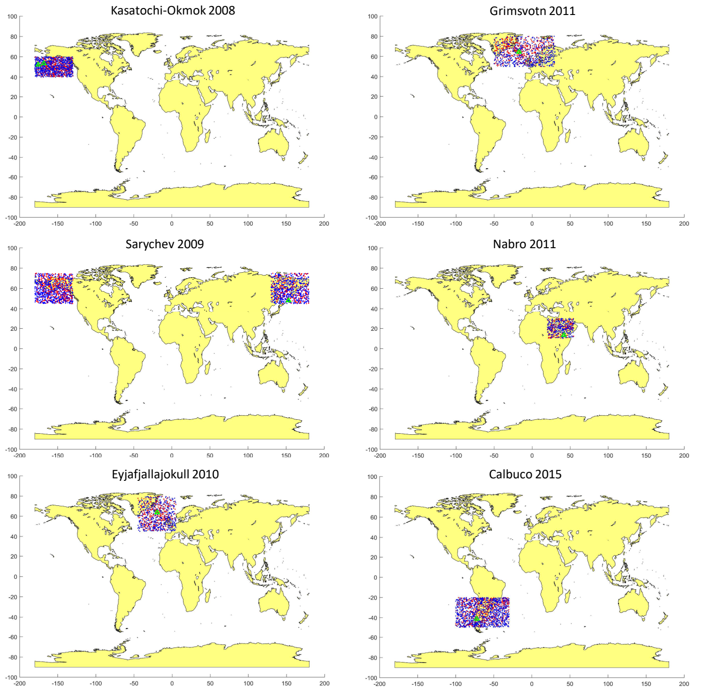
Eruption	# collocations	Bangle	Temp	Bangle anomaly (Lat 5°)	Temp anomaly (Lat 5°)	Bangle anomaly (Box 2.5°)	Temp anomaly (Box 2.5°)
Kasatochi/Okmok - 2008	742	0,59	0,55	0,74	0,72	0,75	0,76
Sarychev - 2009	1164	/	/	0,77	0,77	0,77	0,79
Eyjafjallajokull - 2010	407	/	/	0,76	0,80	0,76	0,80
Grimsvotn - 2011	509	/	/	0,81	0,85	0,85	0,84
Nabro - 2011	282	0,59	0,56	0,72	0,69	0,72	0,74
Calbuco - 2015	977	0,6	0,6	0,71	0,73	0,68	0,75

We used two different climatologies as reference to compute the anomalies: one based in latitudinal bands of 5° and one based on boxes of 2,5° in lat/lon.

The bending angle performs better than temperature, however the use of absolute values does not provide a good classification performance. The use of the anomaly computed versus monthly climatologies, instead, provides better results with an accuracy up to 85% for the Grimsvotn volcanic cloud. The use of different climatologies to compute the anomaly sligthly impacts the algorithm performances.

Eruption	# collocations	Bangle	Temp	Bangle anomaly (Lat 5°)	Temp anomaly (Lat 5°)	Bangle anomaly (Box 2.5°)	Temp anomaly (Box 2.5°)	Bangle anomaly (Lat 5°) Clearsky	Temp anomaly (Lat 5°) Clearsky
Nabro - 2011	282	0,59	0,56	0,72	0,69	0,72	0,74	0,76	0,81

In case of the Nabro 2011 eruption we have also selected clear sky profiles to check if the algorithm performances can be further improved. The improvement is relevant especially in case of using the temperature as target.



° volcanic cloud ° non-volcanic environment

The RO in non-volcanic environment are randomly selected in the same month of the eruption, but different years. For training the algorithm the number of non-volcanic RO is the same as the number of «volcanic RO».

The algorithm is not optimized yet, so the performance can still be improved. However, it is interesting to notice that the use of bending angle or temperature profile does not affect much the results, but the use of the anomaly (instead of the absolute value) is fundamental to get a higher accuracy. The results are even better if the algorithm is trained and tested with clear-sky data.

This study is part of the Volcanic clouds dEtection and monitoring for Studying the erUption impact on climate and aVIatiOn (VESUVIO) project, funded within the Programme Supporting TAlent in ReSearch (STARS) @ University of Padova.