Unsupervised Classification on SAR Backscatter to Identify Morphological Changes on Perpetually Active Volcanoes

Matthew Cross

August 22, 2024

Abstract

Sangay, Ecuador, has been an excellent candidate for unsupervised clustering of TerraSAR-X imagery. ML techniques have been able to identify the pixels which are prone to events in particular areas of the volcano. The projects final aim is to identify flow lengths and areas associated with the pyroclastic flows and lahars which massively effect the surface morphology.

1 Introduction

Regular SAR images of perpetually active volcanoes provide a basis for the detection of ground surface changes following volcanic events. Interferometric SAR uses the RADAR signal phase to detect ground movement, however in many cases coherence is destroyed due to constant changes, so only the backscatter intensity can be analysed. In the particular case of Sangay the Southernmost volcano in the Ecuadorian Andes, monitoring change in backscatter throughout an event has given clear and identifiable images. After large eruptions, bakcscatter change allows particular events to be identified, however an algorithm for event detection would give constant feedback on the growth of erosion channels and the length and areas covered by pyroclastic flows and lahars.

1.1 Example Images

SAR backscatter is dependent on competing factors, including SAR incident angle/local gradients, surface roughness and dielectric properties. These makes backscatter images difficult to interpret, as often these subtleties are difficult to notice, for example weather patterns can change the dielectric properties of the surface.
However, backscatter signals still provide incredibly clear images of morphological changes, even if the returning signal is difficult to interpret.
An example SAR image of Sangay is show below. This image was taken on the 23rd of September 2020, three days after one of the largest events since 2020.

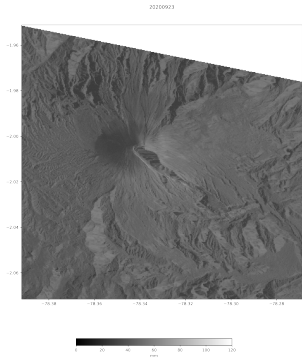


Figure 1: Example SAR image in dB, from which we can analyse backscatter intensity changes

This image is calculated by taking the tenth logarithm of the raw signal. A histogram allows better visualisation of the backscatter intensity distribution in these images. It is clear in the image how the backscatter depends on the ground the local gradient with respect to the line of sight, as clear shadows can be seen due to ground topology.

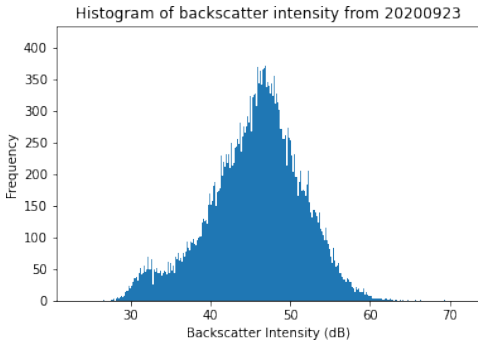


Figure 2: Histogram of backscatter intensity in particular SAR image

Next, the backscatter change ratio and absolute difference between images are calculated as:

$$ratio = 2\frac{(B-A)}{(B+A)}$$

$$difference = B - A$$

Where A is the initial image and B the next taken. The ratio change produces an image taking values between -1 and 1.

2 Method

2.1 Preprocessing of Images

2.1.1 Speckle Filter

SAR data is often very noisy, as can be seen in this above image. This comes from speckle, and occurs when different reflectors in the same pixel interact and give misleading signals. However, this can be filtered out using a Speckle filter, particularly the Gamma-MAP 3x3 speckle filter.

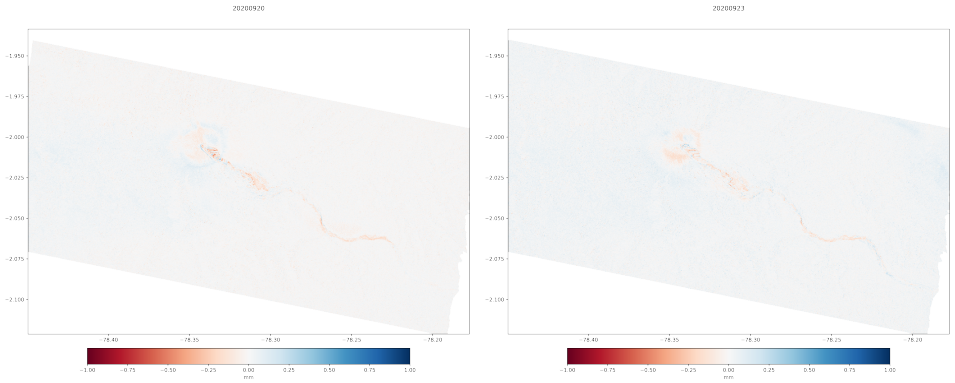
2.1.2 Linear Inversion

The least squares inversion, or linear inversion, is a method which can be used to sharpen satellite images by reducing speckle noise. A linear inversion operates on a specified number, n, of images before an event, and the first image after an event. The inversion essentially maps the n+1 dimensional vectors of each pixel to 2-dimensional space, where the two dimensions are the pixel change after the event, and the pixel value before the event. This allows the images pixels before an event to be combined together as vectors in order to reduce noise in the image and increase variation. It also allows for a much more prominent image to be produced, which in turn will allow the clustering algorithm to identify shapes in the images much more effectively. Understanding the problem with linear algebra allows a better visualisation of the dimensional reduction in play:

$$\begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} * \begin{bmatrix} step \\ prestep \end{bmatrix} = \begin{bmatrix} prestep \\ prestep \\ prestep \\ step + prestep \end{bmatrix}$$

One difficulty associated with implementing this method on a perpetually active volcano is that if the algorithm is programmed to take the three images before an event, the volcano is very active in this particular timeframe and the time between acquisitions is quite long (for example 20 days), then the resultant inversion would not be accurate as the previous acquisitions would not just include noise but

also another eruption. For this reason, I have adapted my algorithm to take only images which occur after the largest events, and so linear inversions do not overlap with other large events.



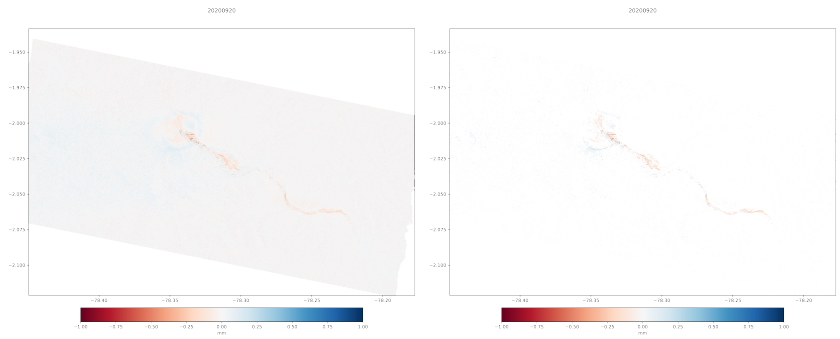
(a) Linear Inversion (b) Ratio Change

Figure 3: Linear Inversion compared to Simple Ratio Change

In this particular case, SAR images were taken on the 1st, 12th and 23rd of September. On the 2nd pyroclastic flows descended the SE flank, along with lahars in the same direction into the Volcan River, generated by heavy rainfall. An explosion on the 20th was noted as the largest explosion during this period, however throughout the week beginning on the 16th pyroclastic flows descended the SE flank daily. Therefore in all of these images the ratio change between acquisitions show similar patterns of pyroclastic flows, and combining them with this method sharpens the images and highlights the areas which on average were most affected. In this way, the variation in the images increases allowing the clustering methods later described to better identify the extents of these flows.

2.1.3 Noise Reduction

After applying a speckle filter and linear inversion, the SAR imagery is still noisy. There are many approaches for getting rid of this noise, however the drawback of hiding particular signals is that they may not be just noise but also signify actual events. However, in order to automatically detect the largest events I included a feature in my algorithm which discards the pixels which take values between a specified threshold, as a percentile. For example, if the specified threshold were 15, then all pixel values within the 15th and 85th percentile would be set to NaNs. This is obviously not a viable method when looking at the backscatter changes due to a number of different effusive events, however when only detecting changes due to pyroclastic deposits and lahars it is a reasonable technique to implement given that these events are associated with the biggest backscatter ratio changes. Further, in a dataset of approximately 400,000 pixels the shapes of the events can still be easily determined even if data is missing. An example of how this implementation can clean up an image is shown below.



(a) Linear Inversion (b) Backscatter Linear Inversion Lower and Upper fifteenth percentile

Figure 4: Linear Inversion with and without cutting low signals

2.2 Clustering Data with KMeans

2.2.1 The KMeans Algorithm

The KMeans clustering algorithm is an unsupervised machine learning method which identifies the best cluster centres in a dataset, given an input guess of initial cluster centres. Often, the input guess of cluster centres is generated randomly if a good initial guess is not known, and this is the method which I have used in this project. The algorithm operates in two steps:

1. Each datapoint in the dataset is assigned to the nearest cluster centre, minimising the distance.

2. For each new group of datapoints surrounding each cluster centre, the average is taken. This set of averages of each group is then assigned as the new cluster centre, and step 1 is repeated.

The algorithm is repeated a large number of times in order to find the best cluster centres for the dataset. More explicitly, the algorithm minimises the loss function:

$$\mathrm{L}_{KMeans} = \sum_{j=1}^{k} \sum_{i \in D_j} |x_i - \mu_j|^2$$

Where μ_j is the cluster centre which each point x_i is assigned.
KMeans is an efficient algorithm for large datasets, and it is also simple to work with. However, it sensitive to initial conditions, and so often it is best to repeat the algorithm many times with different initial conditions in order to find a more general minima of the loss function. One important assumption of KMeans is that the data is clustered around the cluster centres spherically. In this project, and for the majority of unsupervised clustering tasks implemented on series of backscatter images, this is not entirely true. For example, after different volcanic events the backscatter changes associated with each pixel will not have a gaussian distribution which can be seen in the data. In the case of the SAR data from Sangay, the majority of events do have similar shapes and cover similar areas, particularly the consistent increase in the backscatter on the South-Eastern surface after the defined large events **. However, on the 12th of December 2021 an eruption occurred which led to only backscatter decreases on the Northern side of the volcano likely due to the new lava flow reported on the North flank of the volcano on the 2nd December 2021. This event does not correlate with those on the South-Eastern flank which remained almost unaffected, and hence the shape of this event envisioned in greater dimensional spaces when given to Kmeans would not be spherical, and so potentially another clustering algorithm could lead to better results. Generally, though, KMeans has given good results identifying well the different areas affected by volcanic events.

2.2.2 Formatting Data for Clustering

Now that the n large events have been isolated and cleaned mostly from noise, they can be transformed into a format which KMeans can be applied to, so each pixel has n features. To allow for better differentiation between pixels of different topographical properties the DEM, roughness, aspect and slope values for each pixel is appended as a feature of each pixel, which has in turn led to the isolation of clusters which are representative of similar physical features which are not affected by the backscatter changes associated with volcanic events. This is important because if this were not the case then the remaining noise present in pixels which are far from the areas of interest could correlate with the backscatter changes by chance leading to less accurate area and length calculations. Another feature which is important to include is the spatial coordinates of the pixel, which allows for the differentiation between pixels which are physically close together, as often an increase in backscatter can be indicative of both pyroclastic deposits and tephra deposits even if they occur in completely different areas, and their distributions are very different. This leads to a total of six additional features, alongside the backscatter ratio change values of the large events, and so the clustering takes place in n+6-dimensional space. On Sangay nine large events were selected for clustering, and so each pixel has associated 15 features, with each SAR image consisting of approximately 400,000 pixels.

Another feature which I implemented earlier on but was later dropped is the radial distance from the volcano centre. The reasoning behind the inclusion is that the ashfall would have some radial

corelation, however in reality the weather patterns push the ashfall almost always to the West towards the coast.
The nature of ashfall on Sangay has been described in detail in many papers, with isopachs produced highlighting the average tephra depth in surrounding regions, for example in the below figure **Flores A. Javier, Universidad Central del Ecuador 2021. One further implementation which would be useful in identifying ash is one hot encoding the depth of ashfall in these different areas, or manually inputting the depth in the highlighted regions and setting all else to zero. The latter would work best in this particular case, as KMeans is not compatible with one-hot encoded data.

2.2.3 Initial Results

Without using the noise reduction technique described above, and without a speckle filter, the following result was obtained. In this figure the pixel closest to the cluster centres is shown in green, in order to understand from the map what events the clusters have identified.

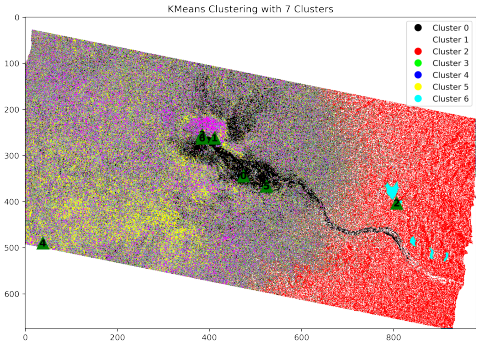


Figure 5: Initial results of kmeans with DEM data and no speckle filter

2.2.4 Reducing Noise by Neglecting Middle Percentiles

With noise reduction but no spatial feature, the following results were obtained.

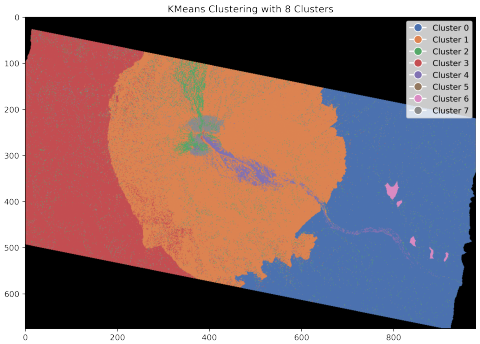


Figure 6: Results after neglecting low values and weigthing DEM data more suitably

This configuration identified the areas of interest much more effectively, however it still did not make a distinction between the regions effected by lahars and the regions effected by pyroclastic flows, which are much closer to the SE flank. To solve this problem, including the spatial feature should allow a better distinction, and particularly increasing the weighting should help too.

2.2.5 Results

After the data has been clustered in this way, it is now possible to use the identified cluster centers on all of the backscatter change images from the entire set, i.e. all of the images taken between 2020 and 2024.

3 Conclusion

4 Future Research

The main goal of this project

References