Classification Of The Atmospheric Structure Of Mars Using Machine Learning

Kunal Jadhav [1,2]

Supervisor: Graham Sellers [2]

1: Department of Physics, Imperial College London 2: School of Planetary and Space Sciences, The Open University

Introduction

The ExoMars Trace Gas Orbiter is a European Space Agency mission developed in collaboration with Roscosmos, to investigate the presence of trace gases in the Martian atmosphere. One of the on board instruments is the NOMAD (Nadir and Occultation for MArs Discovery) spectrometer, whose UVIS (Ultraviolet and VISible) [1] channel was used in this study, to classify the occultation transmission profiles produced by the attenuation of sun light observed through the Martian atmosphere, as measured by the detector. A pair of supervised and unsupervised algorithms were used to classify and cluster similar profiles, which are explained below. The process of classifying and clustering profiles allows the grouping of these profiles into categories, where each group could be representative of the type of atmospheric feature producing them.

Atmospheric Transmission Profiles

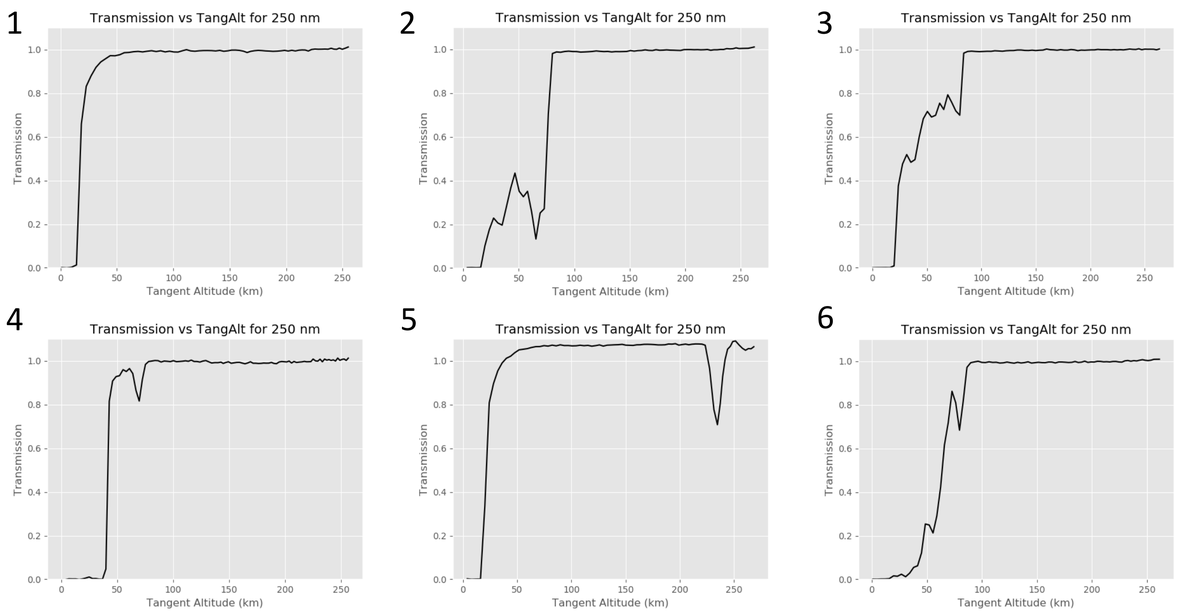

Figure 1: Types of occultation profiles

A binary classification case was assumed, with profiles similar to 1 labelled smooth, and the rest non-smooth. This was developed further to a multi-class problem with the following labels: 1 – smooth; 2 – low altitude structure; 3 - mid altitude structure; 4 – high altitude blip; 5 - extreme altitude blip. These categories were decided based on the altitude at which the main 'structure' of the profile Is observed. An extension would be multi-label classification, e.g. profile 6 could be categorised as low and mid altitude blips.

Curve Length and DBSCAN

In combination with DTW, a similarity measure based on the curve length of profiles was calculated [2], the values of which were stored in a 2D distance matrix. To aid the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [3] algorithm, a peak-finder module was used to detect peaks and valleys of the main 'structure' of each profile. The transmission values at these peaks were altered to certain values to reflect their label, with characteristic curve lengths. This unsupervised algorithm produced reasonable clusters.

Two main parameters to initialise DBSCAN are: ϵ – the neighbourhood radius with respect to a point; and *MinPts* – the minimum number of points defining a dense cluster. A brief overview of the algorithm is as follows:

1. Point *p* is labelled as a core point if at least *MinPts* are within ϵ of itself
2. Point *q* is labelled directly reachable with respect to *p* if *q* is within ϵ of *p*, where *p* is a core point
3. Point *q* is labelled reachable from *p* if a path $p_1,\ldots,p_n$ exists such that $p_1 = p$ and $p_n = q$, where p_{i+1} is directly reachable from p_i. Therefore, all points on the path are core, except possibly *q*.
4. Points which are not reachable from all the others are labelled noise.

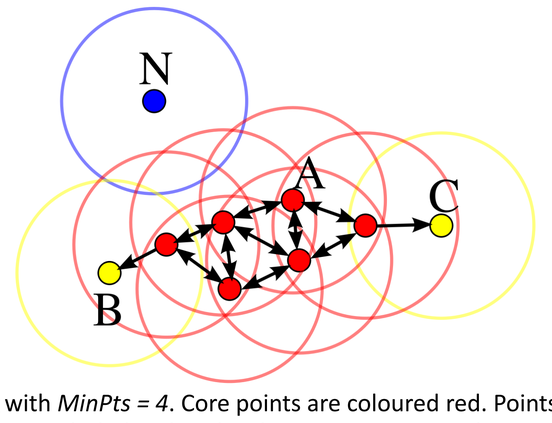

Figure 5: DBSCAN clustering with *MinPts = 4*. Core points are coloured red. Points *B* and *C* are reachable from *A* via other core points and are included within the cluster. Point *N* is neither core nor directly reachable and hence noise. Credit: Wikimedia Commons.

References

[1] M. R. Patel et. al., "The NOMAD spectrometer suite for NADIR and solar occultation observations on the ExoMars Trace Gas Orbiter" (2011)
[2] A. Andrade-Campos et. al., "Novel criteria for determination of material model parameters" (2012)
[3] Martin Ester et. al., "A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise" (1996)
[4] Stan Salvador et. al., FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space (2004)
[5] http://www.mathcs.emory.edu/~lxiong/cs730_s13/share/slides/searching_sigkdd2012_DTW.pdf

Dynamic Time Warping (DTW) and K Nearest Neighbours (KNN)

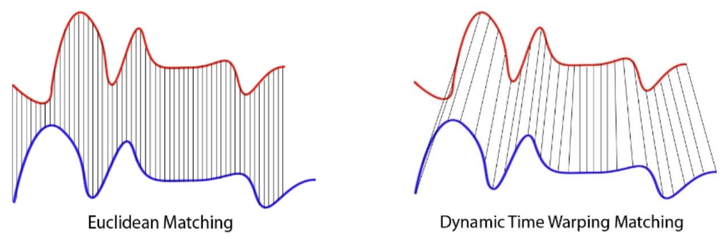

Figure 2: Difference between the matching produced by Euclidean and DTW. Credit : Databricks

In contrast to generic Euclidean mapping, DTW allows signals of varying lengths and 'speeds' or 'phase lags', to be compared as seen in Figure 2. DTW was used as the profiles are essentially time-series of different lengths [4]. Figure 3 shows the difference between the paths and hence mappings produced by Euclidean and DTW:

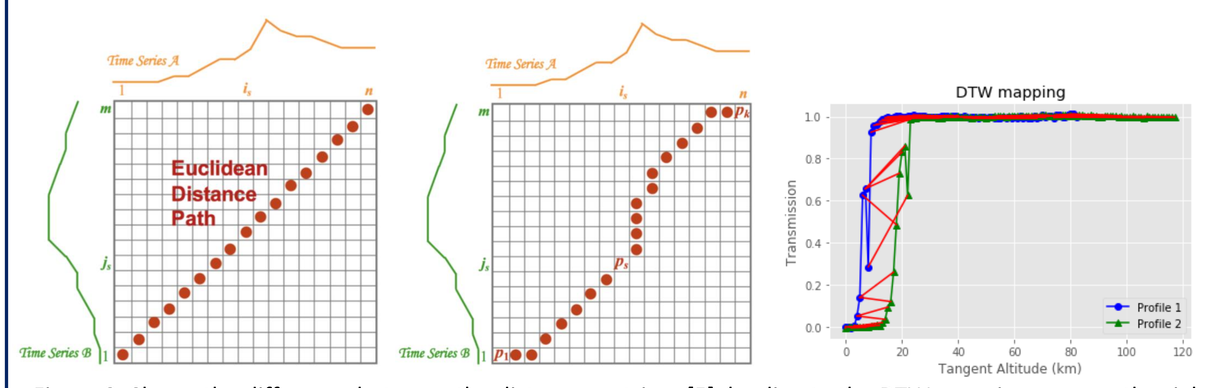

Figure 3: Shows the difference between the distance matrices [5], leading to the DTW mapping seen on the right

DTW deals with phase lags between series by providing a similarity measure by:

1. Dividing the two time-series into equal points
2. Calculating the Euclidean distance between the i^{th} point in series 1 to all *j* points in series 2, and storing the minimum distance in a matrix at *[i,j]*
3. Repeating step 2 but now with series 2 as the initial reference
4. Summing the minimum distances stored from the first point in series 1 to the last point in series 2

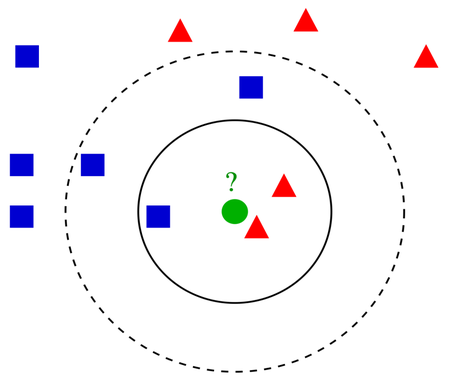

Figure 4: A demonstration of KNN: If K = 3 (solid circle), the green circle is classified as a red triangle. If K = 5 (dashed circle), the green circle is classified as a blue square. Credit: Wikimedia Commons.

The supervised learning algorithm consisted of a simple implementation of KNN, used on a ascending sorted DTW distance matrix, by assigning a profile's predicted label as the mode of the first *K* labels in its row in the matrix. Hyperparameter tuning was achieved by comparing training and test data sets to produce the highest precision.

Conclusion

Supervised Learning: For a binary case of smooth and non-smooth profiles, a healthy precision of 87% was obtained. However, the multi-class case resulted in a modest precision of 67%. This was mainly due to the ambiguity in the class a profile could belong to, as the profiles fell under a broad spectrum of classes. These classifications can be applied as flags on the data to help inform further investigation.
Unsupervised Learning: Considering the profiles could be classified under a spectrum of classes, this was the best approach. Singular label profiles were clustered together and profiles which could have multiple labels were clustered together.
Further Developments: A similar/modified form of this method could be applied to other time series-like and spectral data obtained by NOMAD.

Acknowledgements: I would like to thank the ExoMars team at The Open University, most notably my supervisor - Graham Sellers; Jon Mason and Manish Patel for all their assistance and for such an amazing experience.