This is a peer-reviewed, accepted author manuscript of the following conference paper: Parasyris, A, Stankovic, L & Stankovic, V 2024, Dimensionality reduction for visualization of hydrogeophysical and metereological recordings on a landslide zone. in IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium, IEEE, Piscataway, NJ, pp. 1864-1868. https://doi.org/10.1109/IGARSS53475.2024.10641814

# DIMENSIONALITY REDUCTION FOR VISUALIZATION OF HYDROGEOPHYSICAL AND METEREOLOGICAL RECORDINGS ON A LANDSLIDE ZONE

*Apostolos Parasyris, Lina Stankovic, Vladimir Stankovic*

Electronic and Electrical Engineering
University of Strathclyde
Glasgow, UK

## ABSTRACT

The frequency and intensity of devastating landslides have been increasing worldwide. Timely prediction of slope failure can save lives and protect property. Slope movement is a result of several meteorological and hydrogeophysical variables, such as temperature and moisture content, but this complex relationship is still not well understood. To predict and characterise a slope failure, multiple measurands are usually collected. Since these numerous variables in the predictor set may cause significant increase in complexity, it becomes necessary to use methods that determine the relative importance of measurands that contribute directly to slope failure. To this end, we investigate three methods of visualisation of the feature space and dimensionality reduction, namely Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE) and Linear Discriminant Analysis (LDA), to analyse a range of surface and subsurface measurements from multiple sensors focusing on five stages of slope movement and then make failure predictions using XGBoost regression by setting as predictors two most important components from the extracted features. The results clearly show that LDA better clusters the data points and distinguishes the five different stages of slope movement, including two failures during the period of study encompassing eight years.

***Index Terms—*** Landslide, Dimensionality Reduction, PCA, t-SNE, LDA, XGBoost, Signal Decomposition

## 1. INTRODUCTION

The stability of natural slopes or excavation slopes is determined by the ratio between the magnitude of acting stresses against the shear strength of the geomaterial that constitutes the formation. When many local shear failures take place consequently, they may lead to the generalized shear slip that is called landslide. There are many examples mentioned in the literature on landslides, such as the case of Lodalen railway line, near Oslo in Norway, in 1954 due to the excavation of the original slopes that changed the inclination angle, resulting in 3 wagons facing serious damage and the railway line completely destroyed, or the reactivation 25 years later of an old landslide first occurring in 1925 Shropshire, UK due to intense rainfall, causing disasters in many private citizen homes [1]. [2] describes the mechanisms mainly associated with 3 landslides in the section of the highway "Egnatia Odos" in Greece from Mikro Peristeri to Anthohori. Particularly at the landslide near the entrance of the tunnel of Anthohori, where the intense rainfall on 25-26 August in 2000 played a significant role. Multiple measures were taken such as removal of territorial material to reduce destabilizing forces related to self weight of the mass.

In the era of big data, more and more engineering projects are monitored through sensors that incorporate high technology, placed by scientists to monitor processes, research or ensure safety under operating conditions. Several methodologies based on monitoring have been presented in the literature for various types of assessment such as structural condition assessment of tunnels [3], structural condition assessment of bridges [4], or geophysical investigation [5],[6], [7], in order to ensure safety, to follow a long term behaviour, to be able to take the right measures after an extreme event or for research purposes to understand landslide processes. Underlying features from meteorological measurements influencing slope failure for a north facing high mountain slope experiencing frequent landslides in Ecuador were analysed in [8], concluding that cumulative precipitation over 15 days was by far the most influential factor in landslides. This paper considers 18 meteorological as well as hydrogeophysical measurements taken at subsurface, all measured over a period of eight years on a dynamically moving south facing slope in the UK. In order to visualize the relationship between the 18 measurands, we adopt dimensionality reduction as a pre-processing step to relate to the 5 distinct stages of slope movement. The contributions of this paper can be summarised as follows:

- evaluation and comparison of state-of-the-art dimen-

This work was supported in part by EPSRC New Horizons research programme EP/X01777X/1 and EPSRC Prosperity Partnership research and innovation programme EP/S005560/1. We thank BGS for providing the data as part of EP/X01777X/1. 

sionality reduction approaches for feature extraction on a long-term multi-parameter dataset from a landslide zone

- visualization of transformed components in relation to the five stages of landslide displacement to explain clusters in the transformed feature space
- displacement prediction via XGBoost regression using a subset of explained extracted features

In Section 2.1, we introduce the dataset and measurands, followed by data pre-processing steps. Our methodology is presented in Section 3 and results are shown in Section 4 before concluding in Section 5.

## 2. HOLLIN HILL OBSERVATORY DATASET

### 2.1. Site description

Hollin Hill observatory is a landslide zone that lies to the north of York in UK. It is several hundred metres wide and extends 200 m downslope. Located on the south-facing side of a degraded Devensian ice-margin drainage channel, the slope has an angle of approximately 12°. The slope at Hollin Hill consists of Redcar Mudstone at the base, with an outcrop of the Staithes Sandstone Formation ('Middle Lias') running across the middle section of the slope. A large number of sensors were installed by British Geological Survey (BGS), which together with the COSMOS- UK Network records the various hydroclimatological parameters in the landslide zone, but also inclinometers in the three directions both in the Eastern and Western side of hill. Movement is measured with with marker pegs. We use the following variables measured over the period from 25/03/2014 to 9/03/2022: Displacement, Net radiation, Precipitation, Atmospheric pressure, Air temperature, Wind Speed, Wind direction, Relative Humidity, Heat flux 1, Heat flux 2, Soil temperature 1, Soil temperature 2, Soil moisture 1, Soil moisture 2, Soil temperature at 2cm, Soil temperature at 5cm, Soil temperature at 10cm, Soil temperature at 20cm, Soil temperature at 50cm.

### 2.2. DATA PREPROCESSING

All the non-cumulative meteorological and geophysical data (18 features) are downsampled from half-hour recordings to mean values per day, while all the cumulative data are downsampled to a sum per day. The displacement recordings are transformed to absolute values by substitution of the first recorded point (the reference) from the Leica System and then interpolated at specific regions between small gaps that initially existed in the recordings to obtain a continuous time record. After numerical differentiation, the cumulative displacement time series provide the relative velocity per day. For demonstration purposes, we select sensor 9 which provides recordings with the most significant displacement. The

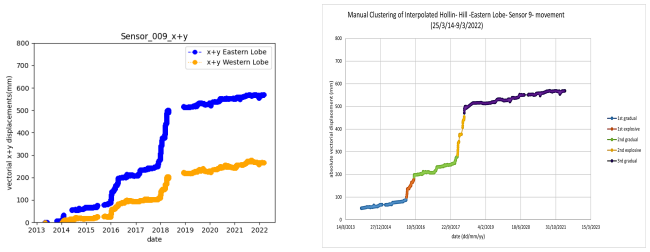



**Fig. 1**. Left to right: Initial vector plane absolute displacements and manual clustering from visual inspection, showing 5 stages of displacement, including 2 explosive failures in 2016 and 2018.

initial displacement plots and the interpolated plots are shown in Fig. 1. In the cumulative displacement time history graph, five characteristic movement stages can be distinguished by manual clustering from visual inspection (Fig. 1, 1st row, right figure). Branches with a lower gradient indicate slower changing ratio of displacement (gradual movement) while the 2 vertical lines indicate rapid movement (explosive movement due to intensive generalised shear slip). In order to analyse how to best predict the gradual and explosive stages, cumulative displacement is decomposed into three parts via statsmodels python library [9] as per Equation 1:,

$$Y = t + s + \epsilon \tag{1}$$

where: $t$ is the trend component which represents the main mode of displacement, $s$ is the seasonal (or periodic) component which represents deformation characteristics influenced by periodic factors and $\epsilon$ is the random (or residual) variable which represents sudden discontinuous changes [10]. As can be seen in Figure 2, the two explosive failures in 2016 and 2018 can be captured by $\epsilon$ (4th row).

## 3. DIMENSIONALITY REDUCTION AND REGRESSION ON EXTRACTED FEATURES

In this section we describe the method used for displacement prediction from the 18 measured features. To reduce complexity, first we perform dimensionality reduction.

Since the 18 features are expressed in different units and in various ranges, before performing dimensionality reduction, we standardise the range of all continuous initial variables so that each one contributes equally to the analysis. The standardisation step, i.e., obtaining zero mean and unit variance, is the critical step for the hydrogeophysical data, since large variances tend to suppress the information carried by small ones which often explain the important complex processes.

We compare three commonly used dimensionality reduction methods: Principal Component Analysis (PCA), t-Stochastic Neighbor Embedding (t-SNE), and Linear Discriminant Analysis (LDA). PCA linearly transforms the data

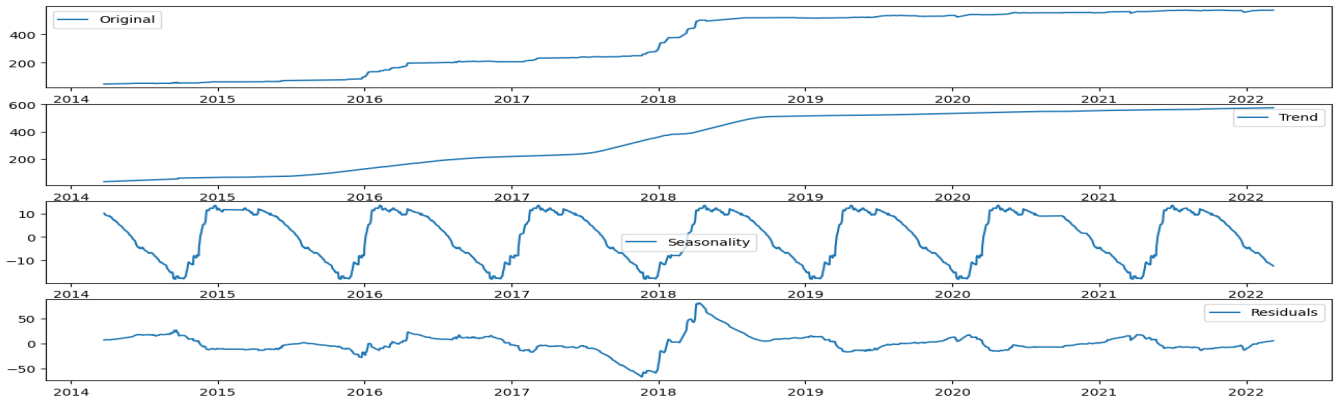



**Fig. 2**. Cumulative displacements time series decomposition from the original absolute displacement (top), to trend $t$ (2nd row), seasonal $s$ (3rd row) and random $\epsilon$ component (4th row).

onto a new coordinate system, defined by principal components. By keeping only the most important principal components that capture the largest variation in data, the dimensionality reduction is achieved. t-SNE is a popular method for visualising higher dimensionality data, widely adopted for presenting the data in the 2D plane or 3D space. LDA focuses on finding the valid projection of the data that minimises the inter-class variance and maximises the distance between the projected means of the classes [11]. Through eigenvalue decomposition, and projection of the data to new axes, the algorithm searches for the best separating border for the classes.

### 3.1. Random Displacement Prediction through XGBoost Regression

After dimensionality reduction, we attempt to evaluate how effective the extracted features from PCA, t-SNE and LDA are in predicting the stages of failure, specifically the random or residual term $\epsilon$ from Equation 1, which most distinguishes the explosive and gradual stages of failure. XGBoost Regressor is used for making predictions using a training/testing split ratio of 50/50 such that the testing set contains predictors for at least one gradual and one explosive displacement.

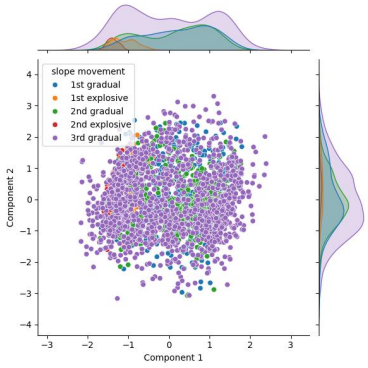
## 4. RESULTS

To make it easier to visualise the results, in this study, we focus on the prediction of random displacement using as predictors only the first two components of each of the dimensionality reduction methods discussed in the previous section. Note that the first two principal components explain 65% of the total variance. The parameters used with t-SNE are: perplexity=50, random-state=0, n-iter=5000 and learning rate=300.

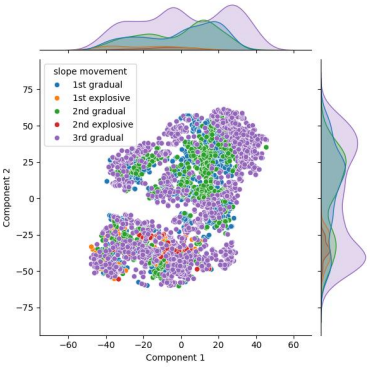
We label each data point with a label corresponding to the 5 stages of failure as per Figure 1 (right), i.e., 1st gradual, 1st explosive, 2nd gradual, 2nd explosive, and 3rd gradual. The labelled data, after being transformed onto 2 components by PCA, t-SNE and LDA, is visualised in Figure 3. We aim to be able to differentiate all the data points corresponding to each stage of failure.

Using 2 principal components (Figure 3a) and because PCA tries to preserve the global structure of the data, local structures can get lost in this high dimensional dataset; thus, it is difficult to differentiate clearly the data points corresponding to the 5 stages of movement. Like PCA, t-SNE is also unsupervised but unlike PCA, t-SNE is non-linear, tries to preserve the local structure of the data and is more complex especially for our large dataset. Therefore, as expected, we can observe slightly more distinct clusters, but the data points corresponding to the 3rd gradual movement are widely scattered over the set. Finally, unlike PCA and t-SNE, LDA is a supervised dimensionality reduction approach that aims to find the best combination of features that best distinguishes the 5 classes. However, it cannot be used when there are no labels for the data points. As shown in Figure 3c), we can observe a more distinct separation of each of the five clusters, which is expected since LDA projection is based on minimising the inter-class variance using data labels, but the 1st and 2nd explosive clusters still cannot be separated.

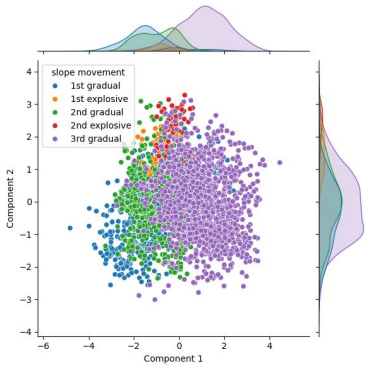
Next we predict the random term $\epsilon$ of the absolute movement $Y$ as shown in the last row of Figure 2, by using as predictors only the first two components in the transformed space for each of the three dimensionality reduction methods analysed previously. The results are shown in Figure 4, where we can observe that the 2 components of LDA were the only ones that captured the peak value on the major failure of 2018 (2nd explosive failure) as shown in Figure 4(d) and comparable with the regression using original 18 measurands as features as shown in Figure 4(a). Additionally, compared to the regression using 18 original features as predictors, we can see that both root mean square error (RMSE) and mean absolute error (MAE) values against ground truth are closest with LDA 2-component training.




(a) 2 component PCA visualisation

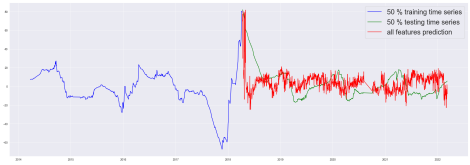



(b) 2 component t-SNE visualisation

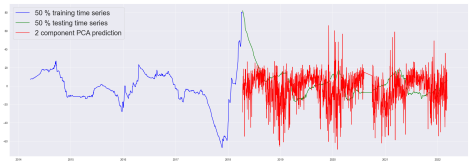



(c) 2 component LDA visualisation

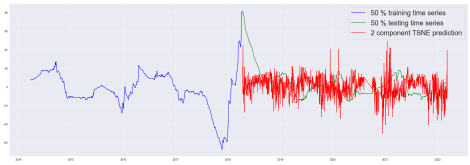
**Fig. 3**. Data visualisation after dimensionality reduction based on clustering of 5 stages of slope movement.



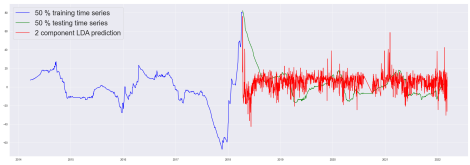

(a) all features, RMSE=17.95, MAE=13.66




(b) 2 PCA components, RMSE=25.79, MAE=19.74




(c) 2 t-SNE components, RMSE=20.32, MAE=15.27




(d) 2 LDA components, RMSE= 20.00, MAE=13.90

**Fig. 4**. XGBoost Regression for Random Displacement time series, based on a) all 18 original features, b) 2 PCA Components, c) 2 t-SNE Components and d) 2 LDA components.

## 5. CONCLUSIONS

In order to reduce prediction complexity for an active landslide zone in UK for a period of eight years, we adopt dimensionality reduction to two dimensions, through LDA, PCA and t-SNE methods, and evaluate their effectiveness through visually comparing how distinctly they distinguish the five stages of slope movement. As unsupervised approaches, t-SNE provides better separation of the 5 stages of displacement compared to PCA. Supervised LDA makes a clearer distinction of the stages of slope failure and therefore better at explaining the transformed feature space. The next step in evaluation of the dimesionality methods is predicting via XGBoost regression gradual and explosive stages of slope failure using as predictors the first two components of PCA, t-SNE and LDA. Both visual plots of the predicted random term from the decomposed displacement, which most clearly highlights explosive failure, and the RMSE and MAE results show that the first two components of LDA are as effective as using all original 18 features in predicting random displacement with the smallest RMSE and MAE observed against ground truth compared to PCA and t-SNE.

## 6. REFERENCES


[1] Graham Barnes, *Soil Mechanics*, Bloomsbury Publishing, Sept. 2017, ISBN 9781137512215.

[2] Spyros Kavounides, *Landslides in Greece*, Parasotiriou Publishing, Sept. 2020, ISBN 978-960-491-146-2.

[3] A. Parasyris and D. Bairaktaris, "Innovative methodology for advanced structural condition assessment of tunnels," in *Expanding Underground - Knowledge and Passion to Make a Positive Impact on the World*, pp. 2493–2500. CRC Press, London, 1 edition, Apr. 2023.

[4] Zhiguo He, Wentao Li, Hadi Salehi, Hao Zhang, Haiyi Zhou, and Pengcheng Jiao, "Integrated structural health monitoring in bridge engineering," *Automation in Construction*, vol. 136, pp. 104168, Apr. 2022.

[5] A. Parasyris, L. Stankovic, S. Pytharouli, and V. Stankovic, "Near surface full waveform inversion via deep learning for subsurface imaging," in *Expanding Underground - Knowledge and Passion to Make a Positive Impact on the World*, pp. 2829–2836. CRC Press, London, 1 edition, Apr. 2023.

[6] Jim Whiteley, Cornelia Inauen, Paul Wilkinson, Philip Meldrum, Russell Swift, Oliver Kuras, and Jonathan Chambers, "Assessing the risk of slope failure to highway infrastructure using automated time-lapse electrical resistivity tomography monitoring," *Transportation Geotechnics*, vol. 43, pp. 101129, Nov. 2023.

[7] S. Uhlemann, A. Smith, J. Chambers, N. Dixon, T. Dijkstra, E. Haslam, P. Meldrum, A. Merritt, D. Gunn, and J. Mackay, "Assessment of ground-based monitoring techniques applied to landslide investigations," *Geomorphology*, vol. 253, pp. 438–451, Jan. 2016.

[8] Byron Guerrero-Rodriguez, Jaime Salvador-Meneses, Jose Garcia-Rodriguez, and Christian Mejia-Escobar, "Improving Landslides Prediction: Meteorological Data Preprocessing Based on Supervised and Unsupervised Learning," *Cybernetics and Systems*, vol. 0, no. 0, pp. 1–25, 2023, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01969722.2023.2240647.

[9] Skipper Seabold and Josef Perktold, "statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.

[10] Qi Liu, Guangyin Lu, and Jie Dong, "Prediction of landslide displacement with step-like curve using variational mode decomposition and periodic neural network," vol. 80, pp. 3783–3799. May 2021.

[11] Petros Xanthopoulos, Panos M. Pardalos, and Theodore B. Trafalis, *Robust Data Mining*, SpringerBriefs in Optimization. Springer, New York, NY, 2013.