FP7 2013 Cooperation Work Programme
Theme 6: Environment (Including Climate Change)

Novel indicators for identifying critical *INFRA*structure at *RISK* from Natural Hazards

Deliverable D5.2

Space-Time Modelling of Structural Behaviour of Critical Infrastructure

Primary Author	Tao Cheng & Khaled Taalab/University College London (UCL)
WP	5
Submission Date	23/10/2016
Primary Reviewer	Zoheir Sabeur/University of Southampton IT Innovation Centre (IT Innov)
Dissemination Level	PU

This project has received funding from the European Union's Seventh Programme for research, technological development and demonstration under grant agreement No 603960.

Project Information

Project Duration:	1/10/2013 - 30/09/2016
Project Coordinator:	Professor Eugene O' Brien eugene.obrien@rod.ie
Work Programme:	2013 Cooperation Theme 6: Environment (Including Climate Change).
Call Topic:	Env.2013.6.4-4 Towards Stress Testing of Critical Infrastructure Against Natural Hazards-FP7-ENV-2013-two stage.
Project Website:	www.infrarisk-fp7.eu

Partners:

Eidgenössische Technische Hochschule Zürich, Switzerland.

Dragados SA, Spain.

Gavin and Doherty Geosolutions Ltd., Ireland.

Probabilistic Solutions Consult and Training BV, Netherlands.

Agencia Estatal Consejo Superior de Investigaciones Científicas, Spain.

University College London, United Kingdom.

PSJ, Netherlands.

Stiftelsen SINTEF, Norway.

Ritchey Consulting AB, Sweden.

University of Southampton (IT Innovation Centre), United Kingdom.

Document Information

Version	Date	Description	Primary Author
Rev01	23/10/2015	Deliverable 5.2 draft for review	Tao Cheng & Khaled Taalab
Rev02	12/04/2016	Deliverable 5.2 draft for review	Tao Cheng & Khaled Taalab
Rev03	03/08/2016	Deliverable 5.2 draft for review	Tao Cheng & Khaled Taalab
Rev04	23/10/2016	Deliverable 5.2 –final for submission	Tao Cheng & Khaled Taalab

Executive Summary

Landslides are a destructive natural hazard which can cause damage and disruption to critical infrastructure networks. They are characterised by different types of movement, cause different types of damage and require different mitigation strategies. Landslides susceptibility maps (LSMs) show the spatial likelihood of landslide occurrence, which can be used to inform planning and mitigation strategies to protect infrastructure, property and reduce risk to lives. This reports presents a two-stage mapping technique to produce LSM and its types by using a data mining method for large heterogeneous regions. Random Forest relates the historical occurrence of landslides and types to a suite of geomorphological conditioning factors which are though to control occurrence and types of landslides. Using the case study in region of Piedmont in northwest Italy, this work demonstrates the efficacy of the data mining approach to create highly accurate LSMs and identify the infrastructure that are highly susceptible to specific types of landslides. This two-stage mapping technique can be used to better inform decision makers looking to reduce the risk posed by landslides hazards.

Technical note

The purpose of this Deliverable is to develop a data mining approach to hazard prediction. A Support Vector Machine (SVM) algorithm was suggested in the proposal as it is an appropriate data-mining technique capable of predicting the likely location of future landslides given a historic record of landslide occurrence and a range of geomorphological data. For this reason, the Deliverable was named 'Software packages of SVM-based modelling', however, during model development other algorithms were tested. Preliminary empirical modelling results showed that the Random Forest (RF) algorithm provided more accurate predictions than SVM, therefore, RF is the algorithm used to complete the work. Although the data mining algorithm used to complete this work differs from the one proposed in the description of work, the overall workflow is unaffected.

Table of Contents

1.0 INTRODUCTION

This report presents the risk mapping of natural hazard (landslide) and its impact to critical infrastructure (transport network) by using the case study in region of Piedmont in northwest Italy. For the risk mapping, it focuses on developing landslide susceptibility map (LSM) and classifying the landslide types by using a data mining approach, Random Forest. By combining the risk map with the transport network, we could identify the potential hazards to critical infrastructure.

Landslides are a natural hazard that can damage structures, jeopardise peoples lives and drastically change the landscape. Landslides can particularly affect critical infrastructure, especially road and rail networks which are likely to be exposed to hazards due to their spatial extent (Bui et al., 2014). Typically, part of the risk management strategy involves the development of a landslide susceptibility map (LSM) which shows the spatial likelihood of landslide occurrence (Fell et al., 2008).

Empirical LSMs are founded on the principle that the location of preceding landslides was determined by a set of conditioning factors: geomorphological variables such as slope, geology and land use (Varnes, 1984). Models and algorithms are applied to these variables to extract a numerical relationship between conditioning factors and the occurrence of landslides. Creating empirical LSMs require data, primarily a landslide inventory. This is data which shows the location of previous landslides within a region, plus any other relevant data such as volume of material involved and speed of movement (Van Western et al., 2008). As well as geomorphological variables, the conditioning factors can include hydrological and man made variables (Ayalew, & Yamagishi, 2005)

The issue here is that developing the numerical relationships between conditioning factors and multiple types of landslides can be difficult as geo-environmental variables will affect the occurrence of different landslides in different ways (Epifânio et al., 2014). This challenge, along with the increasing prevalence of geographical information systems (GIS) has led to the growing use of data mining models for LSM applications (Dai et al., 2001). These include Artificial Neural Networks, Support Vector Machines, Decision Trees and Random Forest. While these models typically require a large amount of data, advantage of a data-mining approach is that they are not subjective and can be applied to large geographical extents.

Most case studies using data mining approaches, however, are at small-medium size (<5000 km^2) or at a single catchment area and deal with a single type or limited number of landslide classes. This work aims to demonstrate that LSMs derived using data mining methods can be applied to a large heterogeneous areas containing a number of diverse landslide typologies. This is achieved by using a two-stage mapping procedure. The first is to produce a statistical LSM which treats all landslides as a single class, showing overall susceptibility. The second is to classify the most probable landslide type for each grid cell in the study area. By combining the two maps, it is possible to identify both highly susceptible areas and attribute the area with a landslide class. This has the benefit of presenting a large volume of information on a single map, which can be easily interpreted by planners and decision makers.

2.0 METHODOLOGY

2.1 Random Forest

The Random Forest algorithm is based on an ensemble of decision trees, which aim to classify data by recursive partitioning based on some explanatory variable. Decision trees use a set of binary rules to predict a target value, based upon a set of training data containing all data on the conditioning factors (represented by the root node). In this instance the target value is binary and represents the presence or absence of landslides. The algorithm determines both the conditioning factor which most accurately separates the data into landslide and non-landslide classes and the threshold value at which to split data.

After first level splitting, the algorithm will search for the next splitting variable under each new internal node. Splitting will stop when an internal nodes contains data of only one class (i.e. all the data is either landslide or non-landslide). This becomes a terminal node and is classified as either landslide or non-landslide. If we wanted to predict new data using this model, there would be a lot of misclassification as the model is so finely tuned to a single dataset that it would not be generalise well. To ensure RF models can predict new data, they use an ensemble of decision trees. If the same data was used at the root node and at each split, each decision tree would be identical, making an ensemble pointless. For this reason, RF uses a different subset of the data to train each tree in the ensemble and limits the number of conditioning factors that can be used for splitting at each node. Predictions are made based on the majority vote of all trees in the forest (Breiman, 2001). Bootstrap sampling is used to generate the random subsets of data, which are drawn from the full dataset with replacement.

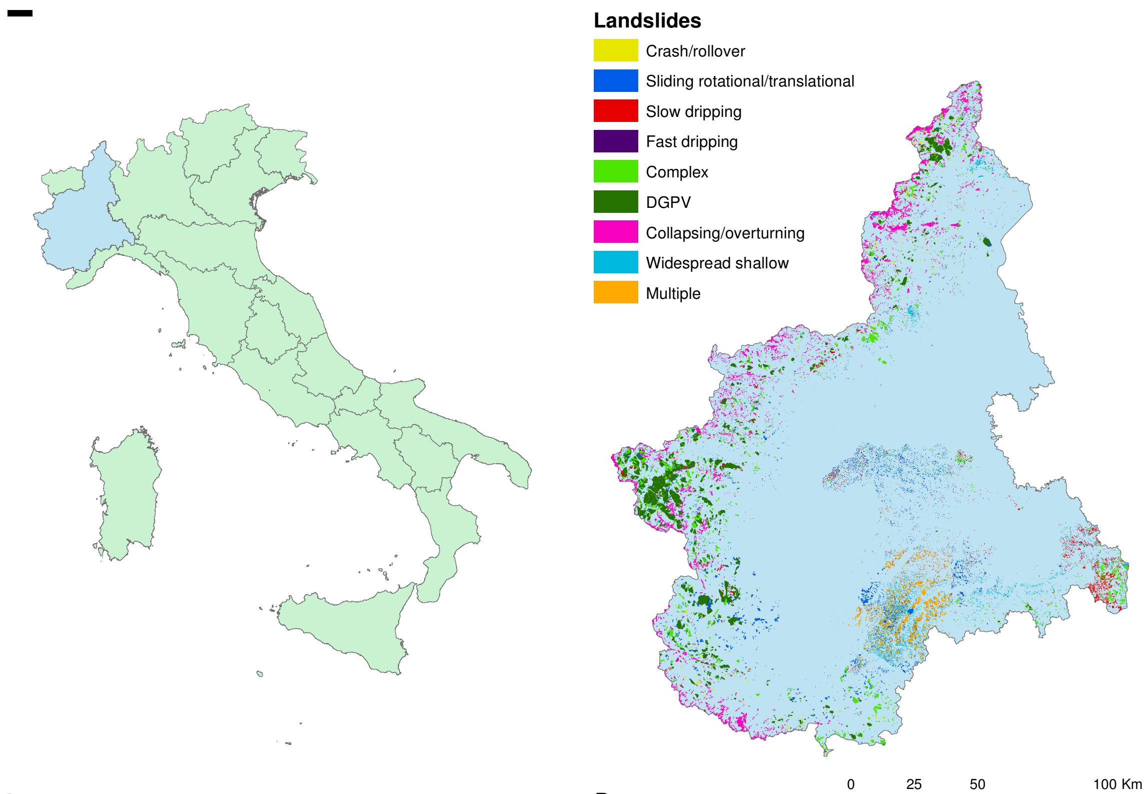
2.2 Data sampling and validation

Assessing the quality of RF models requires validation with independent data. This involves comparing model predictions with observed data. The closer the predictions to the observed data, the better the model is said to perform. Typically, the accuracy is assed using a confusion matrix (e.g. Kavzoglu et al., 2014). For landslide susceptibility there are a number of issues with this approach. Validation requires samples from susceptible and non-susceptible areas to both train and validate the model. Sampling susceptible areas is relatively straightforward. Areas where landslides have occurred in the past can be considered to be susceptible. The areas in and around previous landslides provides the location of the susceptible samples. Non-susceptible locations can be considered as all areas beyond a buffer zone of the previous landslides. However, just because an area is not included within the landslide inventory doesn't mean that it is not susceptible to landslides. It is possible that it has either experienced a landslide in the past that was not recorded in the inventory or it has no yet experienced a landslide but it is susceptible. Despite these caveats, this is still the most common approach to sampling non-susceptible areas (Park et al., 2013)

The geoenvironmental data selected to be used as conditioning factors must be sampled to correspond with each susceptible and non-susceptible point. Inventory data and data on conditioning factors are usually stored as GIS layers made up of grid cells of a given resolution. Once susceptible and non-susceptible areas have been identified, sampling conditioning factors from corresponding spatial locations is a simple GIS operation. This procedure provides the data used to train the RF models. For most data-mining modelling approaches, the total training dataset is split with approximately 70% of samples used to train models and the remainder used for validation.

2.3 Study area

To demonstrate the efficacy of RF for LSM and classification, the method has been applied to the case study region of Piedmont, a 25402 km² region in northwest Italy (Figure 2). This region is particularly appropriate as since 1950, in Italy alone the economic cost of landslides has been more than 53 billion Euros. Piedmont has been identified as being within a landslide 'hotspot' (Jaedicke et al., 2014). For this reason we would like to determine both the susceptibility to landslides and the type of landslides to which an area is susceptible.



A B

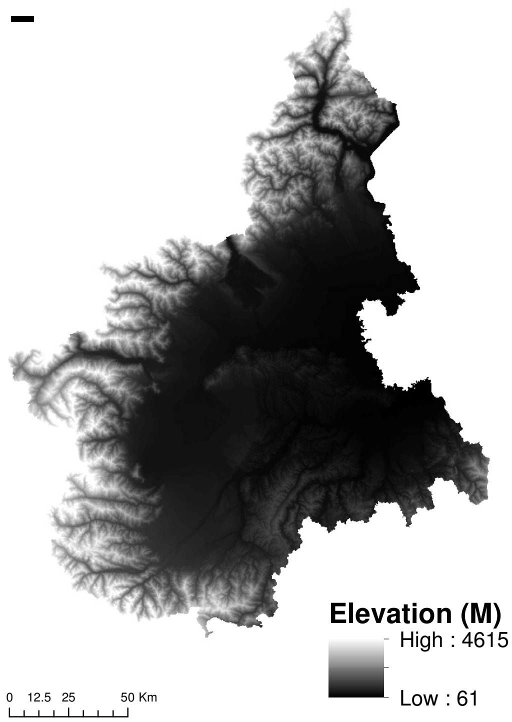
Figure 2: A) Location of Piedmont study area within Italy. B) Location and classification of landslide within the study area.

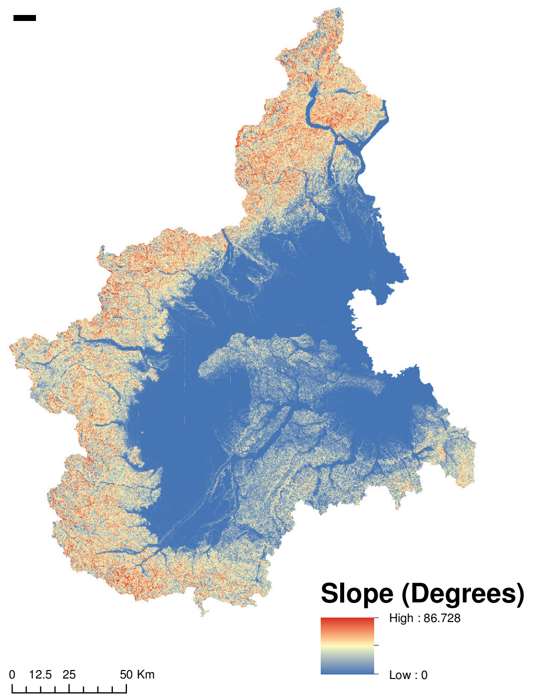
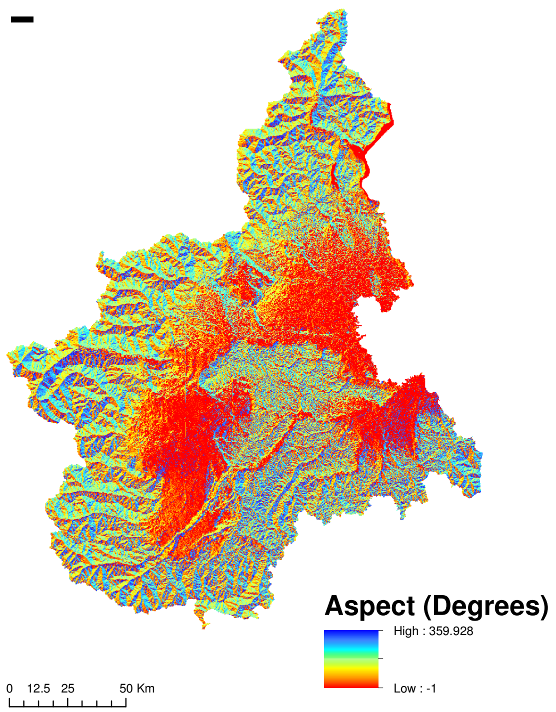
2.4 Landslide Inventory data and conditioning factors

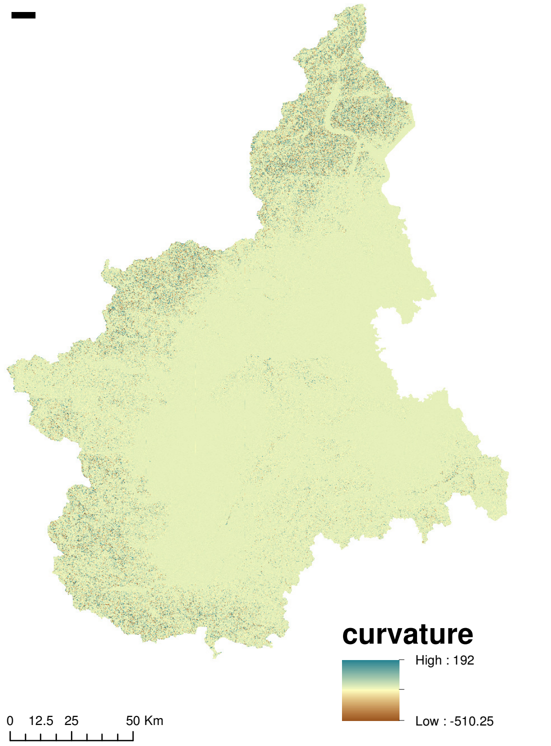
The landslide inventory used in this study is SiFRAP (Sistema Informativo Frane in Piemonte-Landslide information system in Piedmont) is a dataset containing 30439 landslides dating from the early 20th century to 2006, mapped at a scale of 1:10000 (Figure 1B) (Lanteri & Colombo, 2013). This is an update of the IFFI (Inventario dei fenomeni franosi in Italia- Inventory of Landslide in Italy) project (Amanti et al., 2001). Of the 30439 landslides, 20723 have been classified based on the type of mass movement involved. There are nine classes of landslides in the region: crash/rollover, sliding rotational/translational, slow dripping, fast dripping, complex, DGPV (a slow, complex deformation of rock), collapsing/overturning, widespread shallow and multiple. There is no definitive landslide classification, however this is the one devised by the Inventory of Landslide Phenomena in Italy (IFFI) adopeted by SiFRAP. A comprehensive description of the classification taxonomy is available from SiFRAP (2009). The location of the landslides in the inventory are shown in Figure 2B.

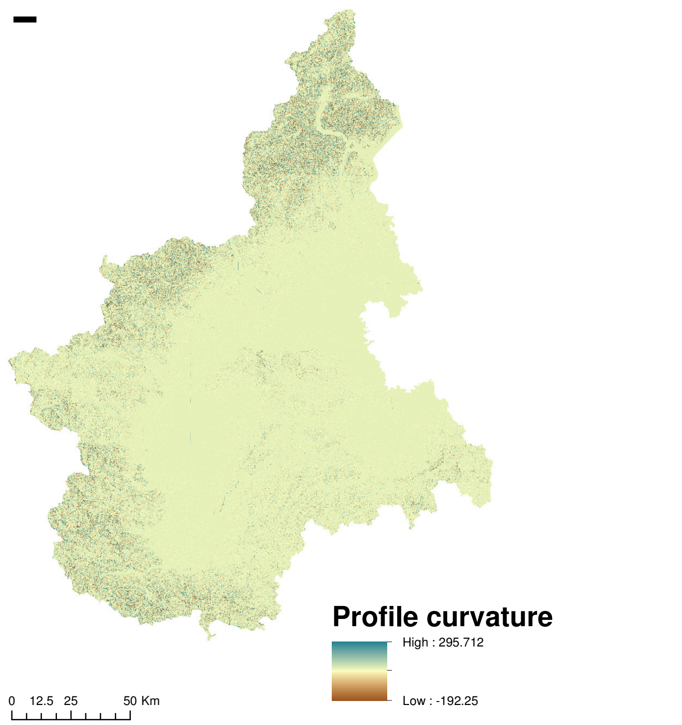
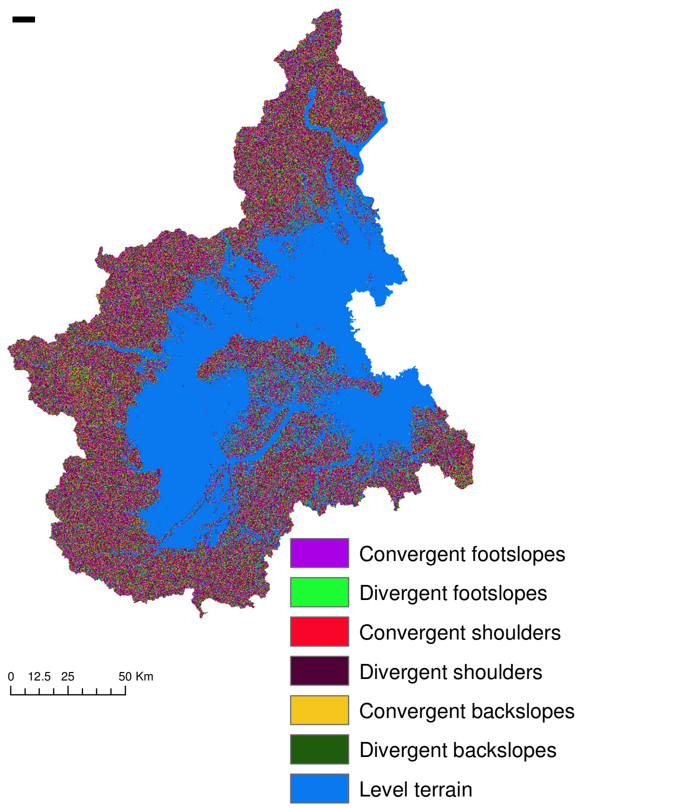
2.5 Conditioning factors

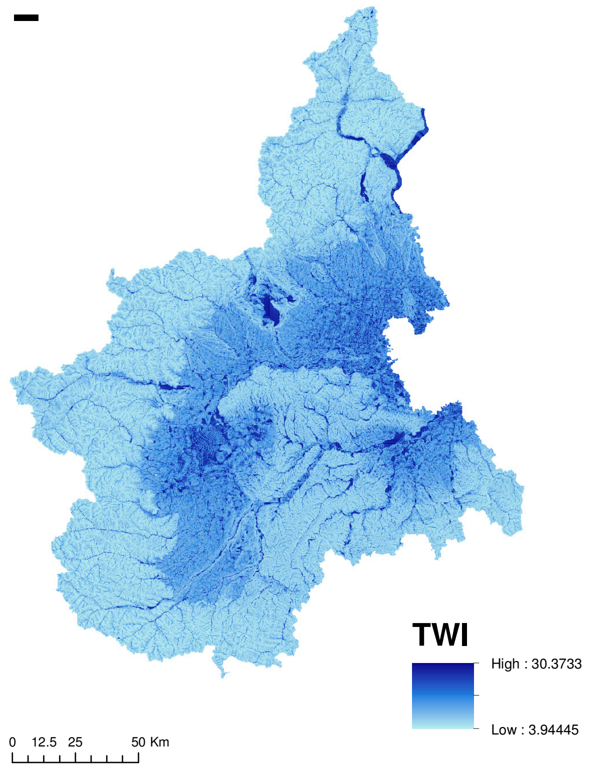
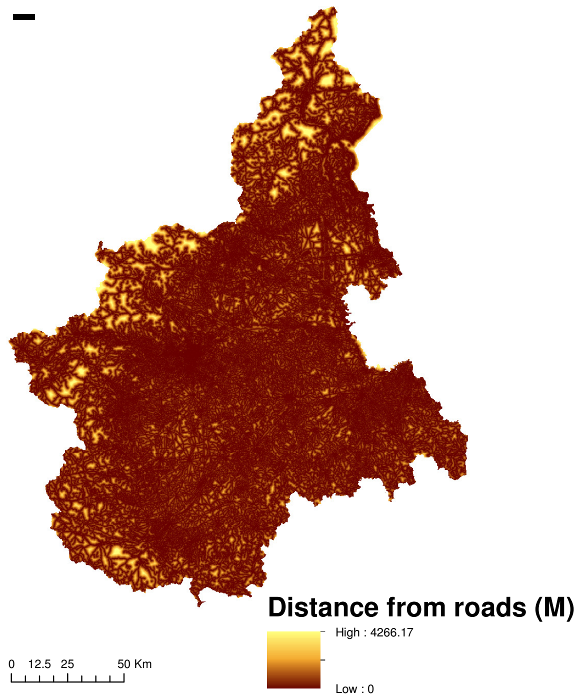
The range of conditioning factors used in this study are shown in Table 2.

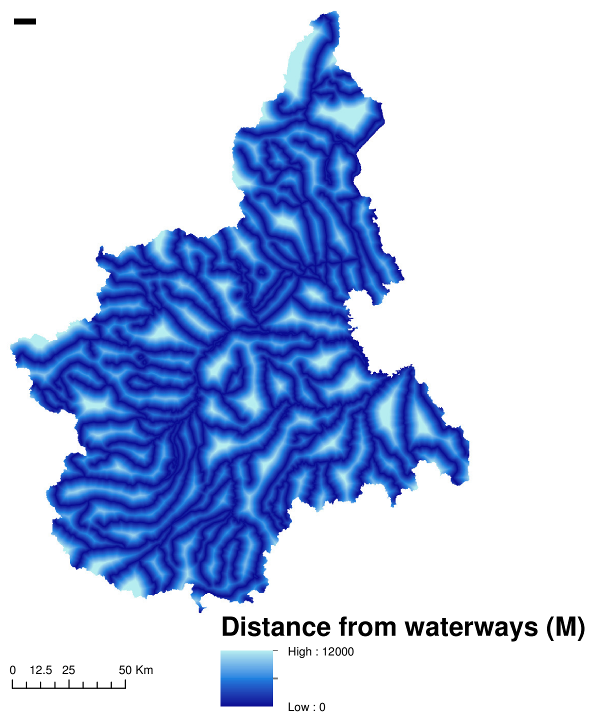
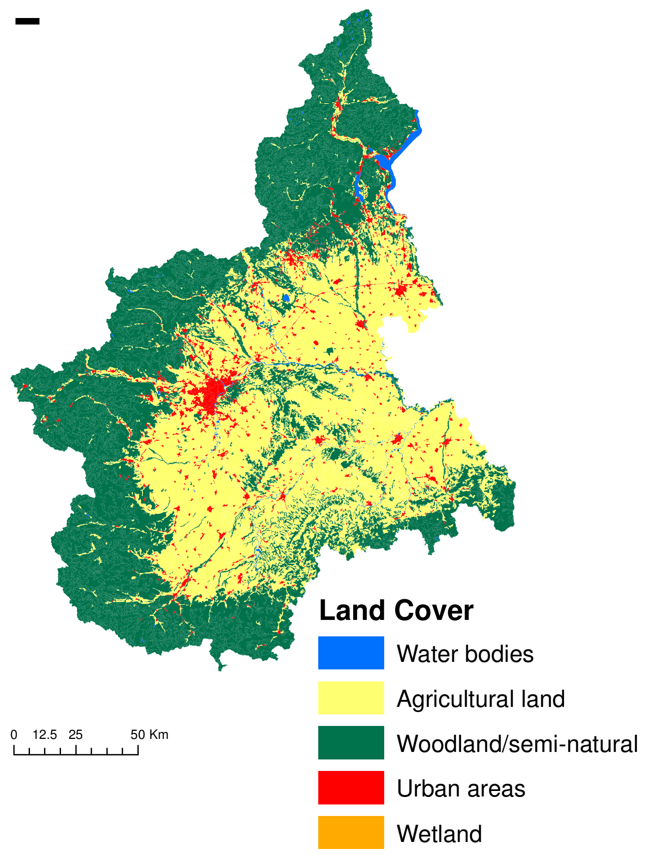
Map	Description
	Elevation is commonly used in LSM as it is usually indicative of climatic and vegetation patterns. The DEM used in this study is a 20 m resolution raster showing elevation above sea level. Ranging from 61-4615 meters above sea level.

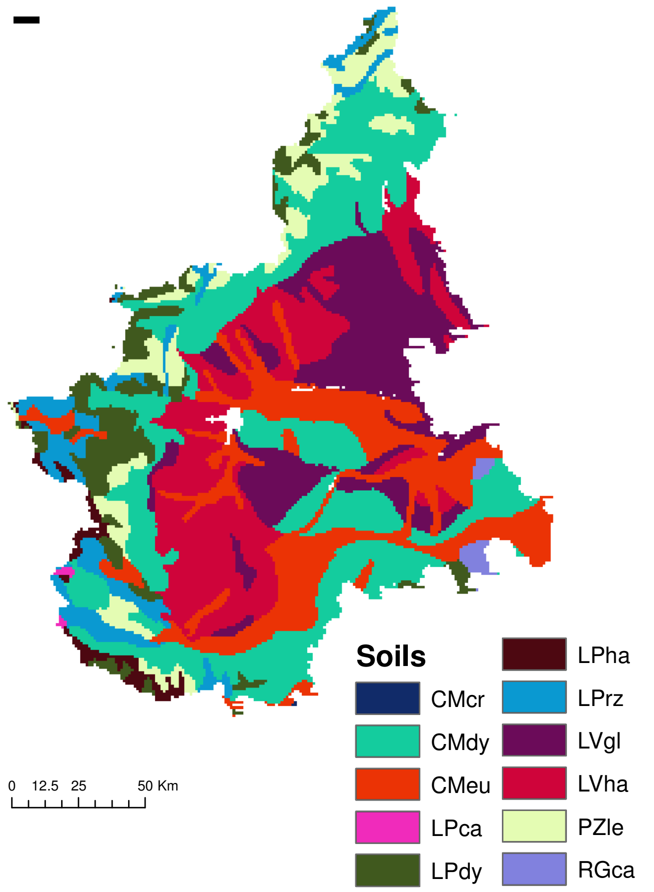
	Slope is the angle formed between any part of the surface of the earth and the horizontal. The angle is a prominent controlling factor on the shear stress experienced by earth and rock mass on a slope. This is on a 20 m grid derived from the DEM ranging between 0°-87°.
	The aspect of each grid cell will have a bearing on the amount of rainfall and intensity of rainfall it experiences, as well as the amount and intensity of solar radiation. Aspect is defined as the compass direction of a slope. From 0° to 360°, flat areas are assigned -1.

	Curvature can be thought of as the slope of a slope. This will affect both stress on the material on the slope and the movement of water across the slope surface. Derived from the DEM on a 20m grid. The values range from -510 to 192
	The slope perpendicular to the direction of maximum slope. A positive value indicates the cell is part of a sidewardly concave slope. A negative value indicates that the slope is sidewardly convex. In the study areas this ranges from 59.9 to -214.5

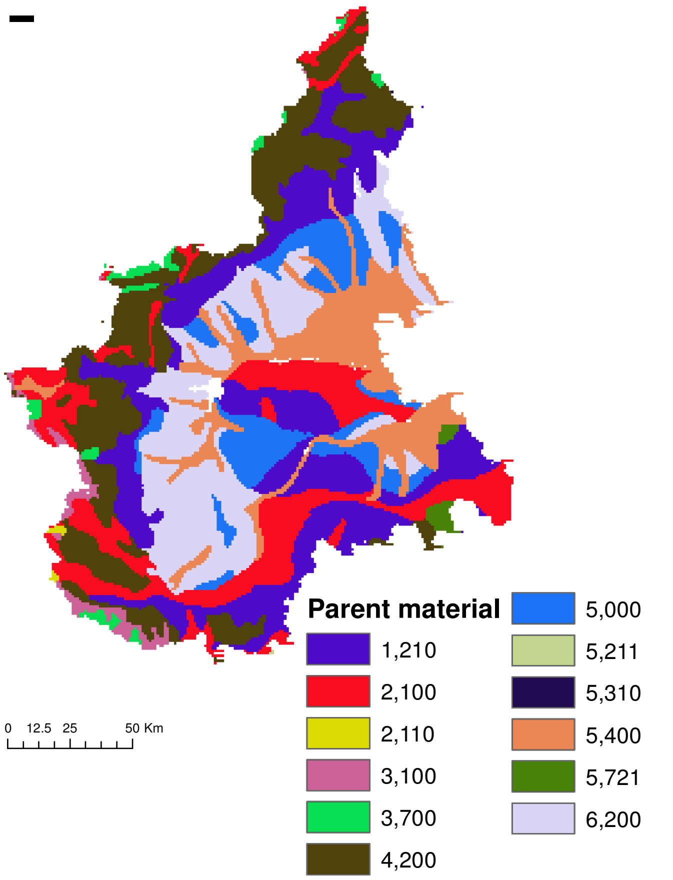
	This is the rate at which the slope gradient changes parallel to the direction of maximum slope. A positive value indicates the cell is part of an upwardly concave slope. A negative value indicates that the slope is upwardly convex. In the study areas this ranges from 295.7 to -192.3.
	Pennock landform classification, divided the study areas into 7 classes of three dimensional landform elements. (Pennock et al., 1987). Landform has been shown to strongly influence LSM (Schulz, 2007).

	Topographic wetness index represents a hypothetical measure of the accumulation of water flow at any point within a river basin. This can be considered to represent the distribution of soil moisture in the region. This was derived from the DEM on a 100m grid. Values range from 3.9-30.4.
	Building roads can destabilise slopes, leaving them predisposed to landslides. Furthermore, the vibrations caused by traffic can become a triggering mechanism. This was derived from the OpenStreetMap road network map of Italy using a GIS operation to calculate distance from a line. This produced a raster grid of 100 m resolution.

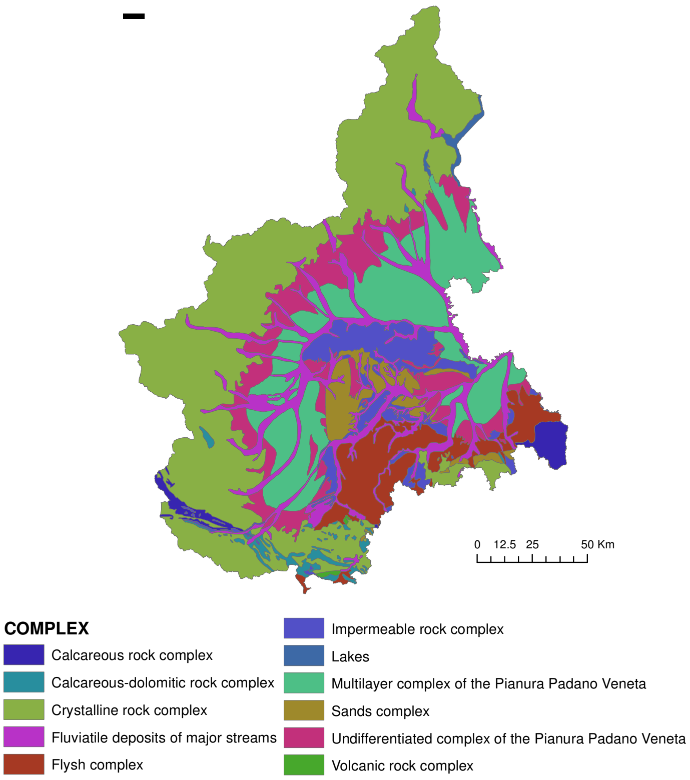
	Proximity to the stream network has been shown to influence susceptibility. Streams have the power to erode soil and apparent material (Gomez & Kavzoglu, 2005). This was derived from a river network map of Italy (available from ISPRA) using a GIS operation to calculate distance from a line. This produced a raster grid of 100 m resolution
	Represents vegetation and how the land is used, both of which can influence susceptibility. We use the CORINE land cover map 2006. A 1:100,000 scale land cover map divided into 16 land cover classes in the region, produced by interpretation of Landsat TM and SPOT HRV satellite imagery

	Soils influence how water moves across the landscape. Some soils are more cohesive, others more prone to erosion. This will affect the conditions which trigger landslides and the type of landslide which occurs. World Reference Base (WRB) soil classification on a 1 km grid, divided in 15 classes in the region (Van Liedekerke et al., 2006)

<table>
<tr>
<td>



</td>
<td>Lithology represents the geomechanical properties of bedrock and is a controlling factor in the structural and chemical propertied of soil. This study uses a 1 km raster grid showing the dominant parent material, divided in 12 classes in the region (Van Liedekerke et al., 2006)</td>
</tr>
<tr>
<td>



</td>
<td>A classification based on hydrogeological formations, which contain similar geological, hydrogeological, productivity and hydrogeochemical facies. This map is produced by ISPRA at 1:10000 scale, divided into 11 classes within the region.</td>
</tr>
</table>

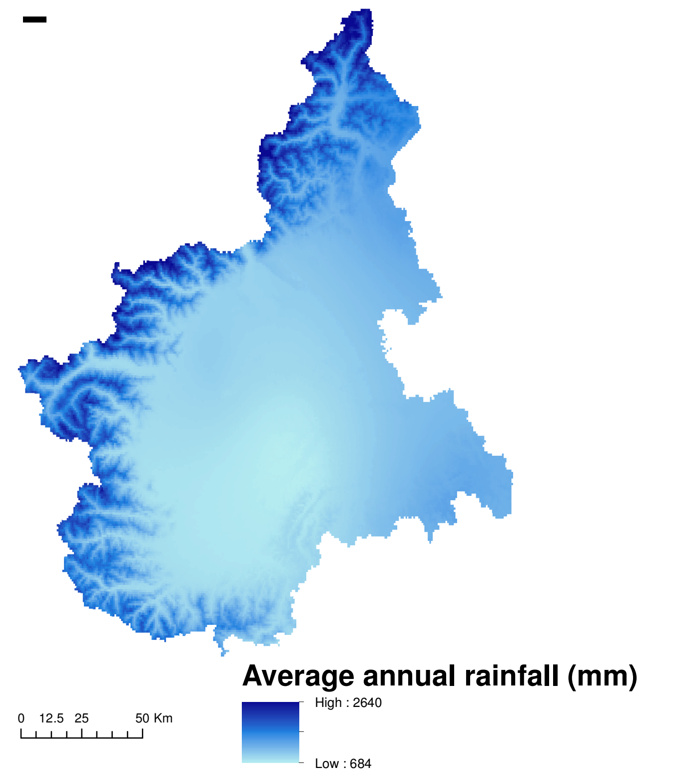
	Rainfall is generally considered as a triggering mechanism for landslides, however, it is rarely included in LSM. In a study area this large, the rainfall will be spatially variable and should therefore be considered as a predisposing factor (Catani et al., 2013). Here the average annual rainfall on a 20 km² grid ranging from 684-2640 mm y^{-1}.

Table 2: landslide conditioning factors maps and description

3.0 RESULTS

3.1 Landslide susceptibility

The LSM produced by RF are shown in Figure 3. Figure 3 shows a classification, where the region is classified as either susceptible to landslides (red) or not susceptible to landslides (green). Across the region, susceptibility is highest in the mountainous areas in the north, west and south of the region and lowest in the alluvial plain in the centre and east. Validation of the LSM using independent data showed that the classification accuracy of the map was over 88%, which suggest that the RF method can be used to create highly accurate LSMs.

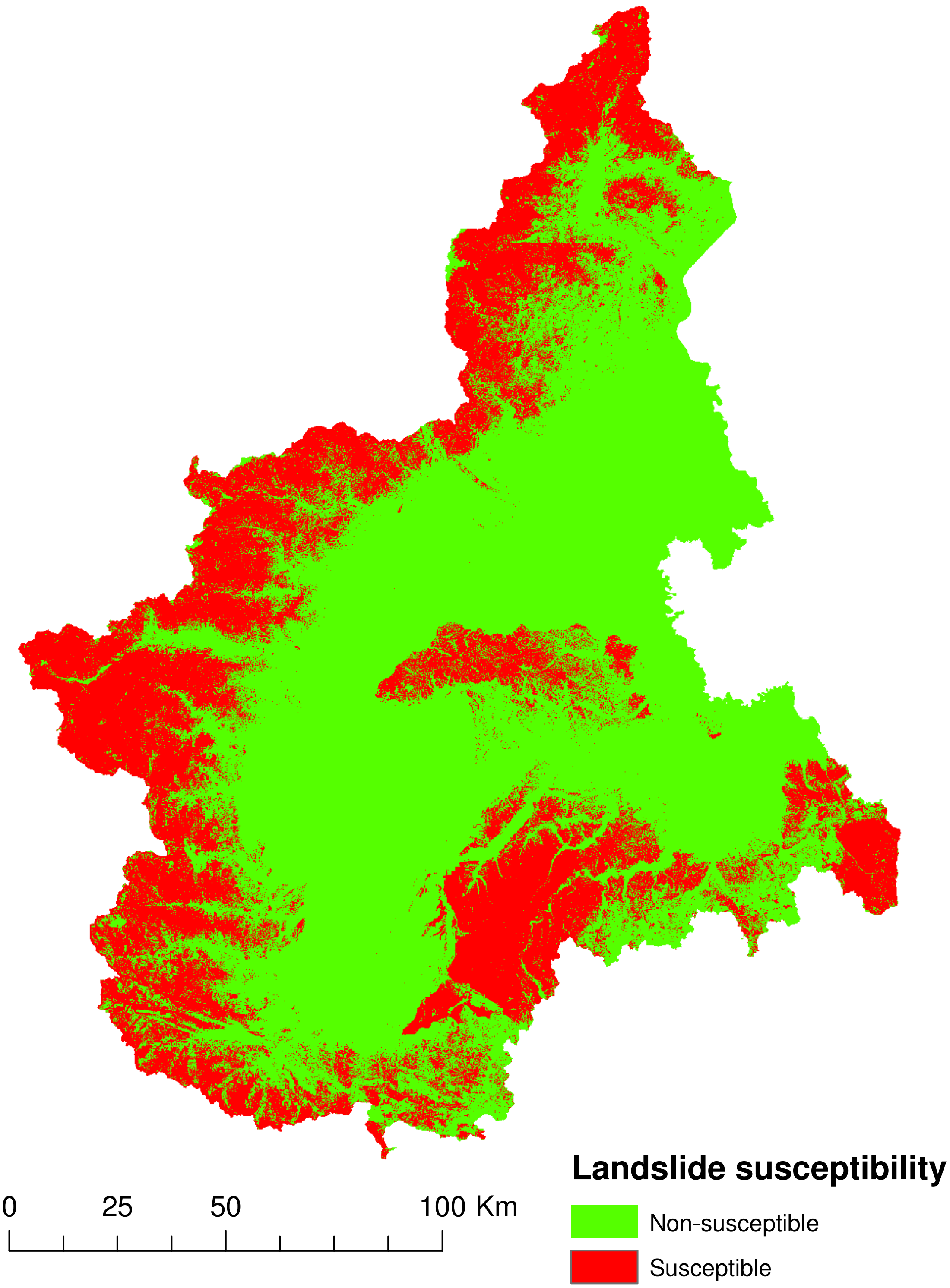


Figure 3: The landslide susceptibility map for Piedmont

Susceptibility maps can be used as part of hazards mitigation strategies. The amount of information shown on a LSM can be increased if we use RF to classify the probable types of landslides that will occur within a given region. Figure 4 shows the susceptible areas of Piedmont and the types of landslides to which these areas are susceptible. For example, these maps can be used to identify sections of the road network which would benefit from barriers to stop large rocks and debris affecting the road. As resources are limited, protecting the roads in this way needs to be focused on the most exposed areas. A landslide susceptibility map alone will not always be the most effecting tool for planning mitigation strategies as in this instance, it is necessary to identify specific types of landslides. For example, is a road passed through an area that was highly susceptible to slow dripping landslides, it probably would not benefit from these specific preventative measures.

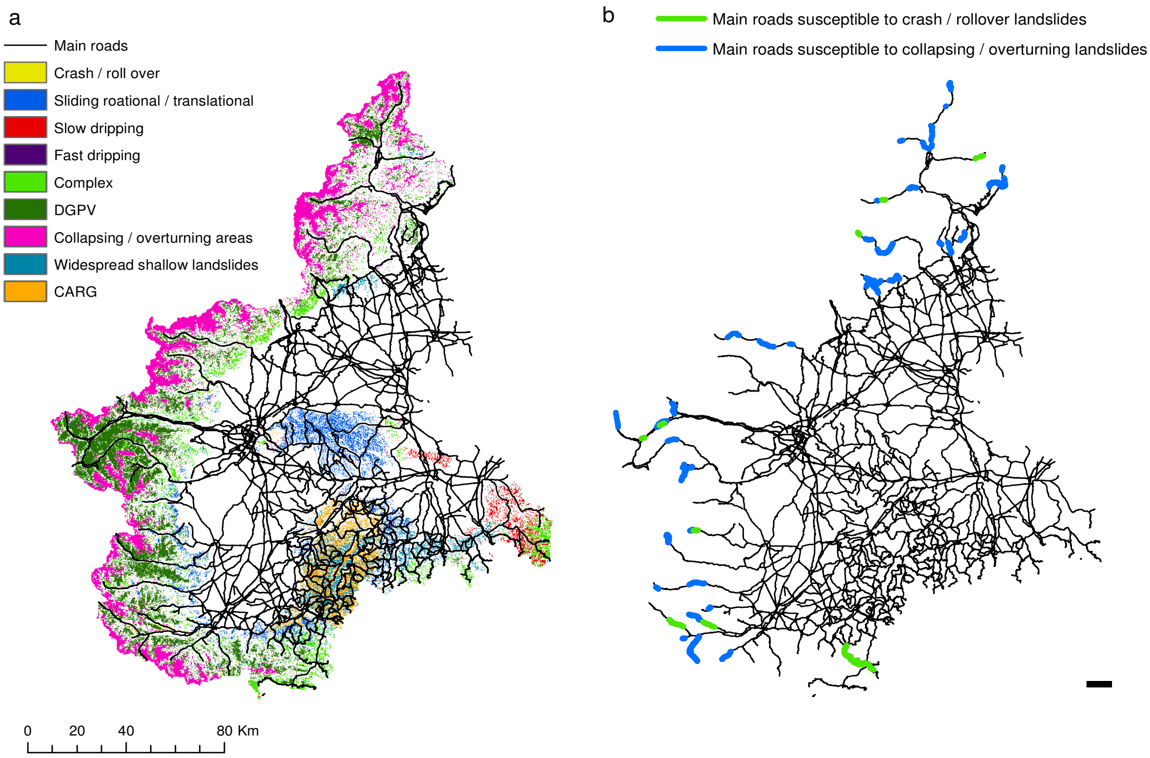


Figure 4: Example of landslide classification mapping to identify roads exposed to collapsing/overturning and crash/rollover type landslides.

The advantage of this approach is that it conveys a lot of information in a way that is easy to interpret, which is critical for LSM applications (Guzzetti et al., 1999). Moreover, it can be tailored to focus on any area or to address any types of landslide-related problem. For example, when planning new developments it can be used to help select suitable routes for new infrastructure networks or help inform planners of the type of hazards that new structures may need to withstand.

3.2 Model run times

Models were trained in R using the RandomForest package (Liaw & Wiener, 2002). The computer used to train the models had an Intel(R) Xeon(R) CPU E5520 @ 2.27GHz, 2261 Mhz, 4 Core(s), 4 Logical Processor(s) and 12.0 GB installed physical memory (RAM). Training the model on a data frame with 335544 rows and 20 columns took 792.56 seconds. Making predictions for 2538135 records took 105.77 seconds.

4.0 CONCLUSION

Landslides are a destructive natural hazard which can cause damage and disruption to critical infrastructure networks. Landslides susceptibility maps (LSMs) show the spatial likelihood of landslide occurrence. These maps are used to inform planning and mitigation strategies which are responsible for protecting infrastructure, property and reducing risk to lives This work shows the development of an empirical LSM using a Random Forest data mining algorithm. This method uses the location of historical landslides and a suite of geomorphological data to predict the spatial likelihood of landslide occurrence in the Piedmont region. Using a Random Forest algorithm, this work demonstrates the efficacy of the data mining approach to create highly accurate LSMs for large, heterogeneous regions.

Not all landslide are the same. They are characterised by different types of movement, cause different types of damage and require different mitigation strategies. To address this issue, RF also used to create a classification map which, in conjunction with the LSM, could be used to identify areas that are highly susceptible to specific types of landslides. This two-stage mapping technique can be used to better inform decision makers looking to reduce the risk posed by landslides hazards.

5.0 REFERENCES

Amanti, M., Bertolini, G., & Ramasco, M. (2001). The Italian landslides inventory–IFFI Project. In Proceedings of III Simposio Panamericano de deslizamientos (pp. 1-2).

Ayalew, L., & Yamagishi, H. (2005). The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. *Geomorphology*, *65*(1), 15-31.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Bui, D. T., Ho, T. C., Revhaug, I., Pradhan, B., & Nguyen, D. B. (2014). Landslide susceptibility mapping along the national road 32 of Vietnam using GIS-based j48 decision tree classifier and its ensembles. In Cartography from pole to pole (pp. 303-317). Springer Berlin Heidelberg.

Catani, F., Lagomarsino, D., Segoni, S., & Tofani, V. (2013). Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. Natural Hazards and Earth System Science, 13(11), 2815-2831.

Dai, F. C., Lee, C. F., Li, J., & Xu, Z. W. (2001). Assessment of landslide susceptibility on the natural terrain of Lantau Island, Hong Kong. Environmental Geology, 40(3), 381-391.

Epifânio, B., Zêzere, J. L., & Neves, M. (2014). Susceptibility assessment to different types of landslides in the coastal cliffs of Lourinhã (Central Portugal).Journal of Sea Research.

Fell, R., Corominas, J., Bonnard, C., Cascini, L., Leroi, E., & Savage, W. Z. (2008). Guidelines for landslide susceptibility, hazard and risk zoning for land use planning. Engineering Geology, 102(3), 85-98.

Gomez, H., & Kavzoglu, T. (2005). Assessment of shallow landslide susceptibility using artificial neural networks in Jabonosa River Basin, Venezuela. Engineering Geology, 78(1), 11-27.

Guzzetti, F., Carrara, A., Cardinali, M., & Reichenbach, P. (1999). Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy. Geomorphology, 31(1), 181-216.

Jaedicke, C., Van Den Eeckhaut, M., Nadim, F., Hervás, J., Kalsnes, B., Vangelsten, B. V., Smith, J.T., Tofani, V., Ciurean, R., Winter, M.G., Sverdrup-Thygeson, K., Syre, E. and Smebye, H. (2014). Identification of landslide hazard and risk 'hotspots' in Europe. Bulletin of Engineering Geology and the Environment, 73(2), 325-339.

Kavzoglu, T., Sahin, E. K., & Colkesen, I. (2014). Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. Landslides, 11(3), 425-439.

Lanteri, L., & Colombo, A. (2013). The integration between satellite data and conventional monitoring system in order to update the Arpa Piemonte landslide inventory. In Margottini, C., Canuti, P., & Sassa, K. (Eds.) *Landslide science and practice* (pp. 135-140). Springer: Berlin Heidelberg.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Park, S., Choi, C., Kim, B., & Kim, J. (2013). Landslide susceptibility mapping using frequency ratio, analytic hierarchy process, logistic regression, and artificial neural network methods at the Inje area, Korea. Environmental earth sciences, 68(5), 1443-1464.

Pennock, D. J., Zebarth, B. J., & De Jong, E. (1987). Landform classification and soil distribution in hummocky terrain, Saskatchewan, Canada. Geoderma,40(3), 297-315.

Schulz, W. H. (2007). Landslide susceptibility revealed by LIDAR imagery and historical records, Seattle, Washington. Engineering Geology, 89(1), 67-87.

SiFRAP (2009). Guida alla lettura della scheda frane SIFRAP.*Servizio WebGIS Sistema Informativo Frane in Piemonte-http://gisweb. arpa. piemonte. it/arpagis/index. htm*.

Van Liedekerke, M., Jones, A., & Panagos, P. (2006). ESDBv2 Raster Library-a set of rasters derived from the European Soil Database distribution v2. 0.European Commission and the European Soil Bureau Network, CDROM, EUR,19945.

Van Westen, C. J., Castellanos, E., & Kuriakose, S. L. (2008). Spatial data for landslide susceptibility, hazard, and vulnerability assessment: An overview.Engineering geology, 102(3), 112-131.

Varnes, D. J. (1978). Slope movement types and processes. *Transportation Research Board Special Report*, (176).