# Delineating flood prone areas using a statistical approach

**I. Marchesini[1], M. Rossi[1], P. Salvati[1], M. Donnini[1], S. Sterlacchini[2], and F. Guzzetti[1]**

[1]**CNR IRPI, via della Madonna Alta 126, 06128, Perugia, Italy.**
**Corresponding author: Ivan Marchesini** *ivan.marchesini@irpi.cnr.it*
[2]**CNR, IDPA, Piazza della Scienza 1, 20126 Milano, Italy.**

## ABSTRACT

Floods are frequent and widespread in Italy and pose a severe risk for the population. Local administrations commonly use flow propagation models to delineate the flood prone areas. These modeling approaches require a detail geo-environmental data knowledge, intensive calculation and long computational times. Conversely, statistical methods can be used to asses flood hazard over large areas, or to extend the flood hazard zonation to the portion of the river networks where hydraulic models have still not been applied or can be applied with difficulties. In this paper, we describe a statistical approach to prepare flood hazard maps for the whole of Italy. The proposed method is based on a multivariate machine learning algorithm calibrated using in input flood hazard maps delineated by the local authorities and terrain elevation data. The preliminary results obtained in several major Italian catchments indicate good performances of the statistical algorithm in matching the training data. Results are promising giving the possibility to obtain reliable delineations of flood prone areas obtained in the rest of the Italian territory.

Keywords: Flood, DEM, Hazard, Statistical model, Zonation, Machine Learning Algorithm

## INTRODUCTION

Flood Hazard Maps (FHMs) delineate flood prone areas and are fundamental for a proper land and development planning. FMHs are commonly produced exploiting 1D or 2D flow propagation models, capable of determining potential inundated areas corresponding to different peak discharges, for different return periods. Flow propagation models require detailed descriptions of the topographic surface and multiple river profiles, intensive calculation and long computational times. In Italy, FHMs were derived by River Basin Authorities (RBA) and by regional and provincial administrations, in accordance with the "EU Directive on the assessment and management of flood risks [2007/60/EC]". that have used different propagation models (i.e. different modeling schema) often adopting different boundary conditions. This limits the comparability of the different FHMs, across the Italian territory. Moreover, the RBAs have prepared FHMs only for parts of the major rivers under their direct administrative responsibility. As a result, parts of the major rivers and most of the minor tributaries are not covered by FHMs. ISPRA (2015) has mosaicked all the available FHMs to produce three national flood hazard maps showing (i) areas of high flood hazard ($HH_F$) for expected flood return periods from 20 to 50 years, (ii) areas of medium flood hazard ($MH_F$) for return periods from 100 to 200 years, and (iii) areas of low flood hazard ($LH_F$) for rare flood events. A considerable amount of literature is available on the use of statistical approaches for mapping areas prone to natural hazards. As an example, a number of machine learning methods were used to determine areas susceptible to landslides (Rossi et al., 2010). Few experiments were done to delineate potentially flood prone areas using statistical approaches (Degiorgis et al., 2012; De Risi et al., 2015; Manfreda et al., 2015). In particular, Degiorgis et al. (2012) described an approach to flood mapping that exploited a linear Support Vector Machine (SVM) classifier, coarse-resolution terrain elevation data (hydrosheds.cr.usgs.gov/index.php), and two morphometric variables i.e., the difference in elevation and the distance to nearest stream, measured along the drainage path. Similarly, we adopted a supervised multivariate machine learning statistical approach, exploiting terrain elevation data, to prepare FHMs for the whole river network of Italy and we present the preliminary results obtained for a subset of some of

the major Italian river catchments.

## 1 METHOD

In our experiment, to derive the statistical classification model, we used, as independent data, the same morphometric variables proposed by Degiorgis et al. (2012) corresponding to the difference in elevation (Hp) and the distance (Lp) to nearest stream, measured along the drainage path. Differently from Degiorgis et al. (2012) , we used a Logistic Regression (LR) classification model to derive the FHMs. In addition, we used two different sources of terrain elevation data: (i) the 25 m resolution EU-DEM (http://www.eea.europa.eu/data-and-maps/data/eu-dem), provided by the European Environmental Agency, and (ii) the 10 m resolution TINITALY/01 DEM (with 4.3 m vertical accuracy) (Tarquini et al., 2007) compiled by the Italian Istituto Nazionale di Geofisica e Vulcanologia. Lastly, we exploited the national FHMs mosaicked by ISPRA (2015) for the calibration of the LR models (i.e. used as grouping variable in the supervised analysis). We used GRASS GIS (GRASS Development Team, 2016) and R (R Core Team, 2013) to perform the necessary geographical and statistical analyses.

Figure 1 portrays the simplified logical framework (and the algorithm) adopted for the production of the LR classification models for flood hazard zonation, and the preparation of FHMs. The framework assumes that the algorithm is applied to the entire river network in a single RBA, using one of the two available DEMs. The inputs to the algorithm are: (i) the selected DEM, (ii) the available FHMs for the selected RBA (in the same raster format and resolution of the DEM). In addition, a flow accumulation threshold parameter value need to be specified for the extraction of the river network. First, the river network is extracted from the DEM and the river network originating points are identified using the flow accumulation threshold (Metz et al., 2011). Next, the grid cells representing the river network inside the high flood hazard ($HH_F$) zones shown in the FHMs are selected, and the portions of the catchments draining into these cells are identified and used successively for calibration of the LR models (calibration areas). Maps of Lp and Hp are also prepared. In each selected catchment, we calibrated three different LR models, using Lp and Hp as independent variables (i.e. predictors) and the different dependent grouping variables corresponding to the FHMs provided by ISPRA (2015). To determine the grouping variables, the first LR model exploits the $HH_F$ zones and sets to 1 the areas mapped as potentially flooded and to 0 the other (marginal) areas. Similarly, the second and the third models exploit $MH_F$ (medium flood hazard) and the $LH_F$ (low flood hazard) zones, respectively. We apply the three calibrated LR models to produce three raster maps showing the probability that each grid cell in the DEM is inundated, for short ($HH_{prob}$), medium ($MH_{prob}$) and long ($LH_{prob}$) return periods. We used a threshold to classify the LR probability maps into binary maps showing potentially inundated (1) and marginal (0) areas. We tested various thresholds, and compared the corresponding maps with the original dependent variables. The optimal threshold maximizes at the same time the True Positive Rate (TPR) and the True Negative Rate (TNR) i.e., the capability of the binary map to predict the spatial distribution of the potentially inundated and the marginal areas. Lastly, we applied the optimal threshold to the entire catchment to obtain the modeled, statistically based binary FHMs ($HH_{mod}$, $MH_{mod}$, $LH_{mod}$).

## 2 RESULTS AND CONCLUSIONS

We determined the extent of the flood prone areas for six RBAs (Figure 2), and assessed the performance of the LR models constructing Receiver Operating Characteristic (ROC) curves (Table 1). Inspection of Table1 reveals that the Area Under the ROC Curve (AUC) values are large or very large, indicating from good to excellent performances of the LR models. We note that the differences in performance of the LR models for the different hazard levels are limited. Interestingly, with the exception of RBA 3, where a significant improvement of the AUC was obtained using EU-DEM, the performances of the classification models is not affected significantly by the DEM.

Our results about the TPR and TNR (Table 1) indicates that the models were able to predict between 71% and 95% true flooded and marginal areas (average 85%).

For the Serchio RBA (7), Figure 3 shows a comparison between the FHMs used for the calibration, and the maps obtained applying the calibrated LR models using the EU-DEM. The inserts reveal the general agreement between the maps, and the capability of the LR-derived map to extend the mapping of the flood hazard zones to the entire study area.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1937v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2016, publ: 8 Apr 2016

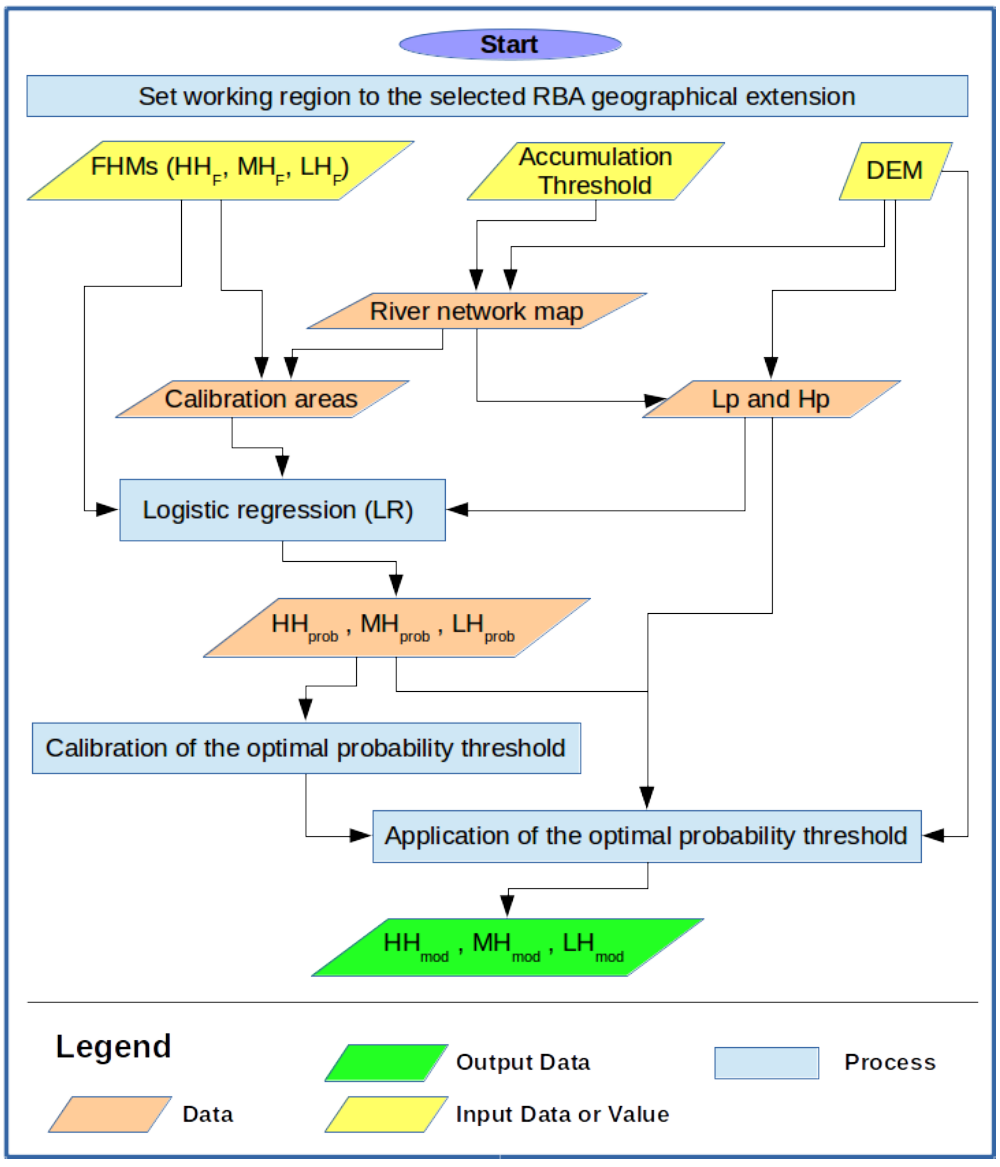


**Figure 1.** Logical framework and algorithm adopted for the production of Logistic Regression (LR) classification models for flood hazard zonation, and the preparation of the modeled, statistically based binary FHMs for selected areas in Italy.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1937v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2016, publ: 8 Apr 2016

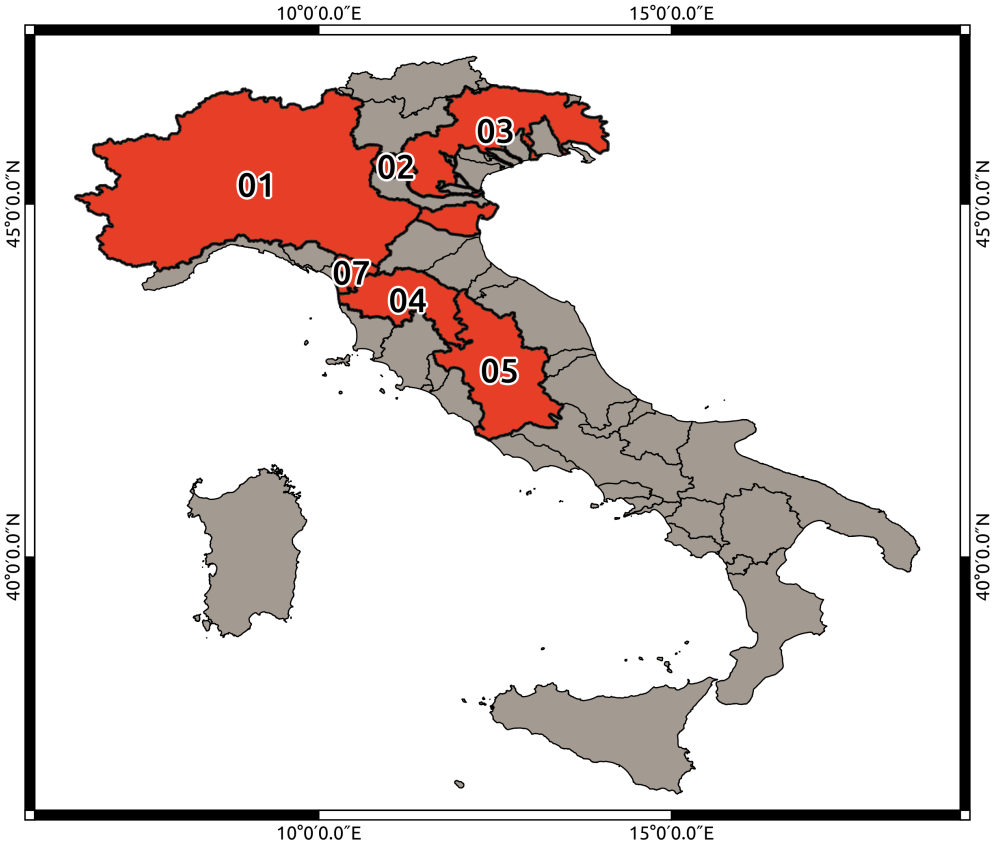


**Figure 2.** Boundaries of the Italian River Basin Authorities (RBAs). Red shows catchments considered in this study.

| RBA | HAZARD LEVEL | AUC | | TPR=TNR | |
|---|---|---|---|---|---|
| | | TINITALY | EU-DEM | TINITALY | EU-DEM |
| 1 | High ($HH_F$) | 0.906 | 0.918 | 0.83 | 0.84 |
| | Medium ($MH_F$) | 0.933 | 0.933 | 0.89 | 0.88 |
| | Low ($LH_F$) | 0.957 | 0.953 | 0.89 | 0.89 |
| 2 | High ($HH_F$) | 0.875 | 0.872 | 0.79 | 0.79 |
| | Medium ($MH_F$) | 0.884 | 0.888 | 0.80 | 0.80 |
| | Low ($LH_F$) | 0.890 | 0.885 | 0.80 | 0.80 |
| 3 | High ($HH_F$) | 0.789 | 0.851 | 0.71 | 0.77 |
| | Medium ($MH_F$) | 0.820 | 0.862 | 0.73 | 0.77 |
| | Low ($LH_F$) | 0.802 | 0.855 | 0.73 | 0.77 |
| 4 | High ($HH_F$) | 0.932 | 0.926 | 0.85 | 0.85 |
| | Medium ($MH_F$) | 0.956 | 0.951 | 0.91 | 0.91 |
| | Low ($LH_F$) | 0.981 | 0.981 | 0.91 | 0.91 |
| 5 | High ($HH_F$) | 0.948 | 0.954 | 0.88 | 0.88 |
| | Medium ($MH_F$) | 0.951 | 0.964 | 0.89 | 0.90 |
| | Low ($LH_F$) | 0.967 | 0.967 | 0.89 | 0.90 |
| 7 | High ($HH_F$) | 0.947 | 0.965 | 0.88 | 0.91 |
| | Medium ($MH_F$) | 0.972 | 0.978 | 0.94 | 0.95 |
| | Low ($LH_F$) | 0.992 | 0.987 | 0.94 | 0.95 |

**Table 1.** AUC ROC obtained from the calibration of the logistic regression (LR) models for six river catchments using two different DEMs, TINITALY and EU-DEM. Values of the TPR = TNR obtained from the optimization of the probability threshold. RBA: River Basin Authorities, AUC: Area Under Curve, TPR: True Positive Rates, TNR: True Negative Rates.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1937v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2016, publ: 8 Apr 2016

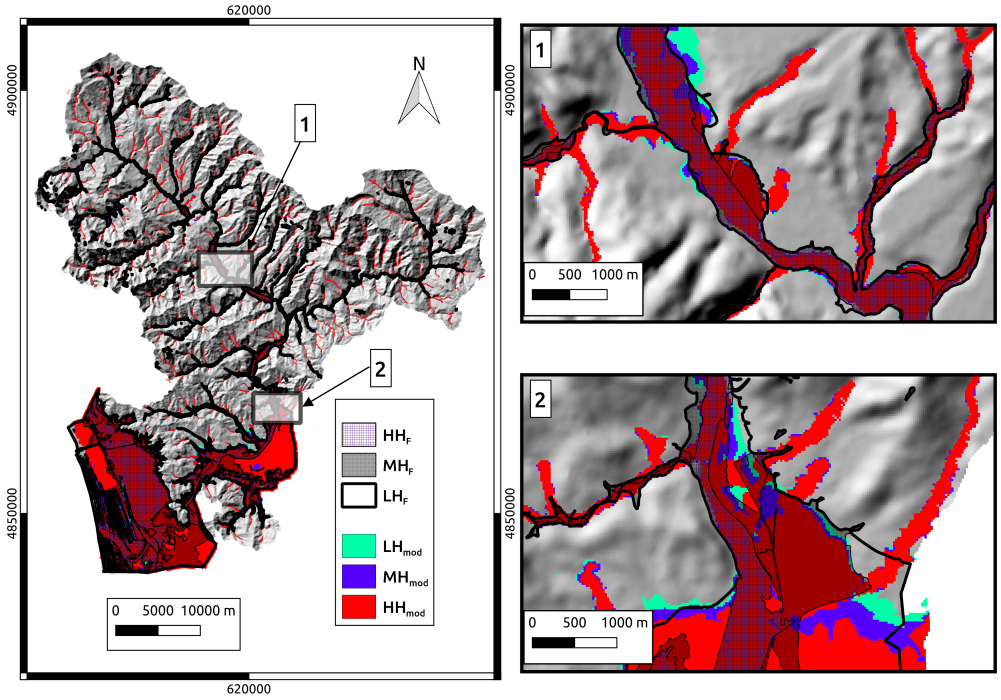


**Figure 3.** Comparison of FHMS produced by River Basin Authorities and prepared in this study using a statistically-based zonation.

## 3 ACKNOWLEDGMENTS

Work partially supported by a grant of the Fondazione Assicurazioni Generali, Trieste. M. Donnini was partially supported by this grant.

## REFERENCES

De Risi, R., Jalayer, F., and De Paola, F. (2015). Meso-scale hazard zoning of potentially flood prone areas. *Journal of Hydrology*, 527:316–325.

Degiorgis, M., Gnecco, G., Gorni, S., Roth, G., Sanguineti, M., and Taramasso, A. C. (2012). Classifiers for the detection of flood-prone areas using remote sensed elevation data. *Journal of hydrology*, 470:302–315.

GRASS Development Team (2016). *Geographic Resources Analysis Support System (GRASS GIS) Software, Version 7.0*. Open Source Geospatial Foundation.

ISPRA (2015). *Dissesto idrogeologico in Italia: pericolosità e indicatori di rischio - Rapporto 2015*.

Manfreda, S., Samela, C., Gioia, A., Consoli, G. G., Iacobellis, V., Giuzio, L., Cantisani, A., and Sole, A. (2015). Flood-prone areas assessment using linear binary classifiers based on flood maps obtained from 1d and 2d hydraulic models. *Natural Hazards*, 79(2):735–754.

Metz, M., Mitasova, H., and Harmon, R. (2011). Efficient extraction of drainage networks from massive, radar-based elevation models with least cost path search. *Hydrology and Earth System Sciences*, 15(2):667–678.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rossi, M., Guzzetti, F., Reichenbach, P., Mondini, A. C., and Peruccacci, S. (2010). Optimal landslide susceptibility zonation based on multiple forecasts. *Geomorphology*, 114(3):129–142.

Tarquini, S., Isola, I., Favalli, M., Mazzarini, F., Bisson, M., Pareschi, M. T., and Boschi, E. (2007). Tinitaly/01: a new triangular irregular network of italy. *Annals of Geophysics*.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1937v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2016, publ: 8 Apr 2016