Effectiveness of Investments in Prevention of Geological Disasters

Ugo Fiore, Zelda Marino, Francesca Perla, Mariafortuna Pietroluongo, Salvatore Scognamiglio, and Paolo Zanetti

Abstract Research on geological disasters has made several achievements in monitoring, early warning, and risk assessment. Substantial resources are being invested in prevention projects, but, due to geographical and demographical complexity, incompleteness of data, and small number of samples, a quantitative analysis on the number of geological disasters and the entity of investments in their prevention is a difficult problem. In this work, the relation is studied between the amount of resources invested in prevention and the number of geological disasters in subsequent years. The analysis is performed on historical data, using statistical methods and a LSTM recurrent neural network.

Keywords First keyword · Second keyword · Another keyword

1 Introduction

Geological disasters are ubiquitous and people awareness about them is rising. It is also thought that mutations induced by climate change are going to exacerbate geological disasters in frequency and magnitude [8]. The availability of abundant data and ability to visualize and analyze them to understand and predict disasters is changing humanitarian operations, crisis management, and investments into prevention dramatically [1]. Studies targeting the regional heterogeneity of geological disasters, evaluating risk and the relation with factors such as direct economic losses,

Authors acknowledge the financial support provided by the Research grant of Università Parthenope, DR no. 953, november 28th, 2016.

U. Fiore (✉) · Z. Marino · F. Perla · M. Pietroluongo · S. Scognamiglio · P. Zanetti
Parthenope University, Napoli, Italy
e-mail: ugo.fiore@uniparthenope.it; zelda.marino@uniparthenope.it; francesca.perla@uniparthenope.it; mariafortuna.pietroluongo@uniparthenope.it; salvatore.scognamiglio@uniparthenope.it; paolo.zanetti@uniparthenope.it

I. S. Kotsireas et al. (eds.), *Dynamics of Disasters*, Springer Optimization and Its Applications 169, https://doi.org/10.1007/978-3-030-64973-9_6

frequency, and cost of prevention, attract much attention. Methods from several disciplines have been applied to this problem, ranging from hydrodynamic models [10] to gray systems [13].

Prevention has many facets, including also the assessment of the extent to which hill-slope communities are prepared [3]. In this work, the relation is studied between the number of prevention projects and the number of geological disasters in subsequent years, accounting for the effect of geographical distribution. The analysis is performed on historical data provided by the Chinese Statistics Bureau. A recurrent neural network of the long short-term memory (LSTM) type has been used to forecast the number of geological disasters in China. The accuracy of predictions was measured with and without the inclusion of data relative to the number of projects of prevention.

2 LSTM

In this work, an LSTM neural network has been used. In addition, well-known statistical methods provided a verification. In recent years, neural networks (NNs) are back on the stage. New and more powerful architectures, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have been proposed. Their main applications include pattern recognition, image processing, anomaly detection [4], and function approximation [5].

Among RNNs, the LSTM proposed in [6] seems to be one of the most promising candidates to analyze sequential data. The traditionally RNNs gradually lose their ability to learn from the past due to the problem of the gradient vanishing. LSTM networks overcame this problem because their structure, composed by three "gates," respectively called input, output, and forget gate and the memory cell resulting from their interaction, allow to keep the long-term information and to combine it with short-term memory. This mechanism promotes in-depth data learning and produces excellent results in terms of performance. An extensive description of NNs and back-propagation algorithm can be found in [2] and [9].

Formally, let $\{\mathbf{x}_t, y_t\}$ a sequence of training examples, with $\mathbf{x}_t \in \mathcal{X}$ and $y_t \in \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^d$ is the input space and $\mathcal{Y} \subset \mathbb{R}$ the output space. According to the universal approximation theorem [7], a feed-forward network can approximate any continuous functions on compact subsets; our aim is to approximate the function g that links the elements $\mathcal{Y}$ and $\mathcal{X}$:

$$y_i = g(\mathbf{x}_i) \tag{1}$$

By denoting $\mathbf{i}$ the input gate, $\mathbf{o}$ the output gate, $\mathbf{f}$ the forget gate, $\mathbf{z}$ and the intermediate state, the general mode of operation of a recurrent network with LSTM architecture with $n \in \mathbb{N}$ units can be described by the following set of equations:

$$\mathbf{i}_t = \sigma(W_i \mathbf{x}_t + U_i \mathbf{h}_{t-1} + \mathbf{b}_i) \tag{2}$$

$$\mathbf{o}_t = \sigma(W_o\mathbf{x}_t + U_o\mathbf{h}_{t-1} + \mathbf{b}_o) \tag{3}$$

$$\mathbf{f}_t = \sigma(W_f\mathbf{x}_t + U_f\mathbf{h}_{t-1} + \mathbf{b}_f) \tag{4}$$

$$\mathbf{z}_t = \tanh(W_z\mathbf{x}_t + U_z\mathbf{h}_{t-1} + \mathbf{b}_z) \tag{5}$$

$$\mathbf{c}_t = \mathbf{c}_{t-1} \odot \mathbf{f}_t + \mathbf{i}_t \odot \mathbf{z}_t \tag{6}$$

$$\mathbf{h}_t = \tanh(\mathbf{c}_t) \odot \mathbf{o}_t \tag{7}$$

where $\odot$ denotes the Hadamard product, σ the logistic sigmoid function, and, for $k = i, o, f, z$, $W_k \in \mathbb{R}^{d\times n}$ and $U_k \in \mathbb{R}^{n\times n}$ are weight matrices and $\mathbf{b}_k \in \mathbb{R}^n$ bias vectors. Weight matrices and bias vectors are adjusted to reflect the characteristics of training data through a learning algorithm. According to equation (7), the output of the LSTM network is computed by the long-term filtered information $\mathbf{c}_t$. The information in the memory cell $\mathbf{c}_t$ is updated according to equation (6). Equations (2), (3), and (4) describe how the three gates work and how they control the long-term memory. The first one filters the flow of information entering the memory cell, while the second one directs the information directly to the output step. In addition, the forget gate determines which information must be retained or removed from the long-term memory.

3 Empirical Evidence

The experiments were aimed to analyze the effectiveness of prevention projects in relation to the number of geological disasters in China. In particular, we investigate the gain in terms of accuracy, in the prediction of the number of disasters, when the information concerning the projects of preventions is included.

The data source is the China Statistics Bureau (http://www.stat.gov.cn/) that provides the number of disasters occurred and prevention projects carried out in the country in the years between 2004 and 2017, distinguishing by region (Fig. 1). After collecting the data, a first round of data processing was carried out. The data source does not contain the observations for 2014 which, for this reason, was excluded from the analysis. Other data were missing. Since there are only 13 years of observations, the regions for which more than three values were missing were excluded. For regions with one or two missing data, the missing values were imputed with the average value between the previous and the following observation. After this preprocessing, the number of regions included in the study, initially equal to 29, was reduced to 23. In addition, the number of disasters has been transformed into a logarithmic scale. This transformation has been also confirmed by visual inspection of the log-log scatterplots of geological disasters and projects of prevention (the year before) by region, shown in Fig. 2. A log-log chart has been chosen to enhance readability. However, in the evaluation, the number of prevention projects has not been transformed in a logarithmic scale, because (i) the range of values is not as

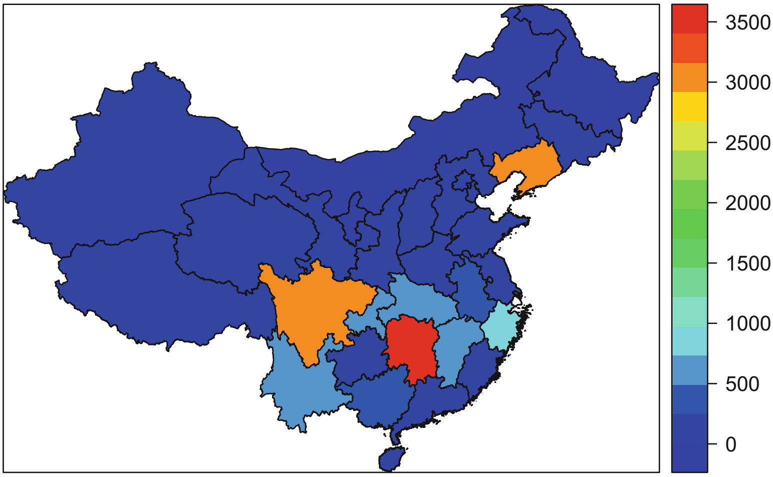

Fig. 1 Geographical distribution of geological disasters in China for the period 2004–2017

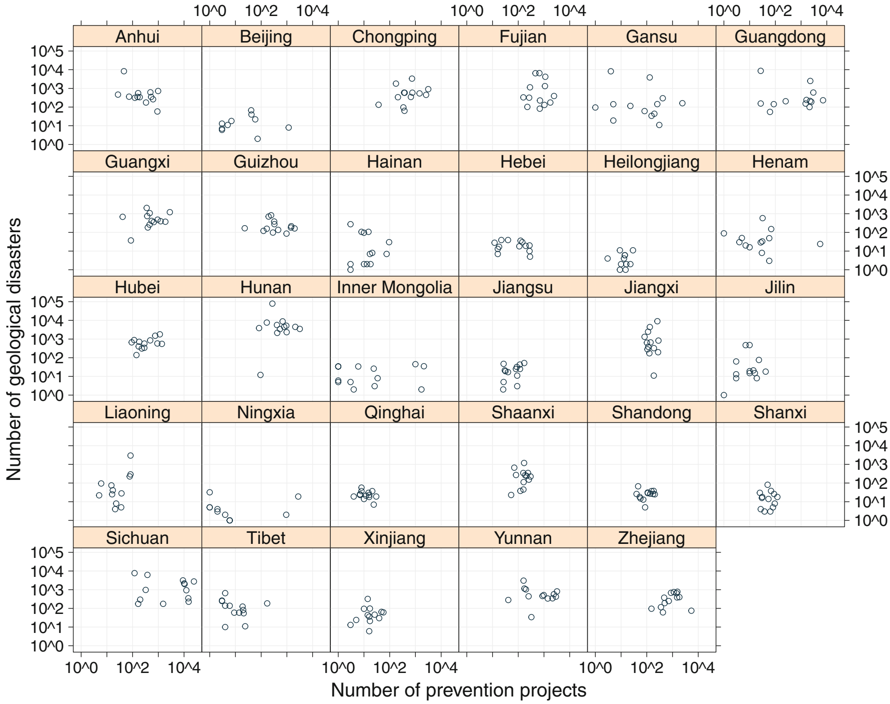

Fig. 2 Geological disasters vs. prevention projects, by region

extreme as it is the case for the number of geological disasters and (ii) the association evidenced by models is stronger when the number of prevention projects is in the original linear scale.

In order to validate the existence of a connection between the number of geological disasters and the time-lagged number of prevention projects, and to identify appropriate values for the lag, we applied statistical analysis. In the following, results of such analyses are briefly reported and then the outcomes of LSTM experiments are shown.

3.1 VAR

Since it is reasonable to think that the number of geological disasters influences and is influenced by the number of prevention projects, the vector autoregressive (VAR) [11] framework seems appealing for the analysis. In VAR models, all variables are treated symmetrically and they influence each other equally. A VAR model with lag one is specified as follows:

$$\mathbf{x}_t = \mathbf{b}_0 + \mathbf{B}_1 \mathbf{x}_{t-1} + \mathbf{u}_t \tag{8}$$

where $\mathbf{x}_t$ is the vector of variables at time t, the components of $\mathbf{u}_t$ are (possibly simultaneously correlated) white noise processes, and the coefficients $\mathbf{B}_1$ capture the influence of the lagged (at lag one) variable to itself and $\mathbf{b}_0$ is an intercept.

The choice of the maximum number of lags to be included is an important step in the analysis. Procedure VARselect in the var package [11] provides the best values for the maximum lag in accordance with four information criteria: Akaike information criterion (AIC), Hannan-Quinn (HQ), Schwarz criterion (SC) – more commonly known as Bayesian information criterion – and final prediction error (FPE). On the basis of that procedure, the suggested maximum lag was one, and this value has been used in the following.

VAR models were fitted for the pair (number of geological disasters, number of prevention projects) considering each region separately. Significant effects at the 5% level were found for each region. The portmanteau test did not permit to reject the null hypothesis that autocorrelations are significantly different from what would be expected from a white noise process.

3.2 Dynamical Models

We are interested in studying variation among regions and over time. Dynamical models [12] allow to estimate the size of current and future reactions of a variable **Y** to a change in another variable **X**. A dynamical model was fit for the number of geological disasters versus the number of prevention projects in the previous year. The maximum lag was kept at one, as in the previous subsection. Effects were found to be highly significant ($p < 10^{-6}$). Analysis of the residuals showed that the residuals were not autocorrelated. When the region was also included, the effects were found to be highly significant for almost all regions. Again, residuals showed no autocorrelation.

3.3 LSTM

After data processing, two neural network models were calibrated: in the first one, the desired quantity y_{t+1} is the number of disasters in the next year; the predictor $\mathbf{x}_t$ contains the number of disasters in the previous 3 years, and the region and year of interest. In the second model, the number of projects of prevention in the previous year is also included in the predictors. Hereinafter, the two models will be referred to as the "basic" and the "advanced" model.

The data are split in two parts: a training set, including the data examples before 2014, and a testing set including the observation after 2014. About the architecture, the network includes a LSTM component with 64 units to process the multivariate time series concerning the number of disasters and an additional feed-forward layer with 64 units, which further jointly processes the partial output with information relating to the region and year of interest. It is important to note that in the second model, the information concerning the projects of preventions is included in this last stage.

The results seem to show that, although the dependence between the number of geological disasters and project of prevention is very small, the use of the information concerning the project of prevention in a supervised learning model with LSTM architecture can be useful. Indeed, by introducing this additional information, the predictive power of the model increases. Table 1 shows the performances, in terms of mean absolute error (MAE), in the testing set for both models. The third column reports the ratio between the errors in the basic and advanced models. It can be seen as a gain factor in performance. It is clear that the prediction of the "advanced model" overperforms the other one in most regions. In particular, a reduction in MAE can be observed in about 70% of regions. The mean gain factor, overall, is 1.56. This evidence seems to be confirmed also by Fig. 3 that shows the residuals in the testing set for both models. In fact the residues of the advanced model seem to be closer to 0 than the others.

4 Conclusions

An LSTM recurrent neural network, a powerful and versatile tool in the analysis of time series data, can provide interesting insights when it is also coupled with a module able to handle grouping information. In this work, an LSTM has been used to assess the relationship between the number of geological disasters and the

Table 1 Mean absolute error of the two models and gain factor, by region

Region	MAE Basic model	MAE Advanced model	Gain factor
Anhui	0.9405	1.1958	0.79
Chongping	1.3894	1.2183	1.14
Fujian	1.1781	0.9725	1.21
Gansu	1.9216	0.8120	2.37
Guangdong	0.7364	0.2860	2.57
Guangxi	0.8260	0.5254	1.57
Guizhou	0.7932	0.3046	2.60
Hebei	1.0762	0.7563	1.42
Henam	0.7054	0.7905	0.89
Hubei	0.6063	0.9151	0.66
Hunan	0.3630	0.3158	1.15
Jiangsu	1.0412	1.2059	0.86
Jiangxi	0.7794	0.6102	1.28
Jilin	0.7370	0.7271	1.01
Qinghai	0.7978	0.4100	1.95
Shaanxi	1.6862	1.1783	1.43
Shandong	0.9115	0.6442	1.41
Shanxi	1.4275	1.6065	0.89
Sichuan	2.9303	2.2625	1.30
Tibet	0.9344	0.4699	1.99
Xinjiang	0.6832	1.0179	0.67
Yunnan	0.8496	0.1487	5.71
Zhejiang	0.7584	0.6898	1.10

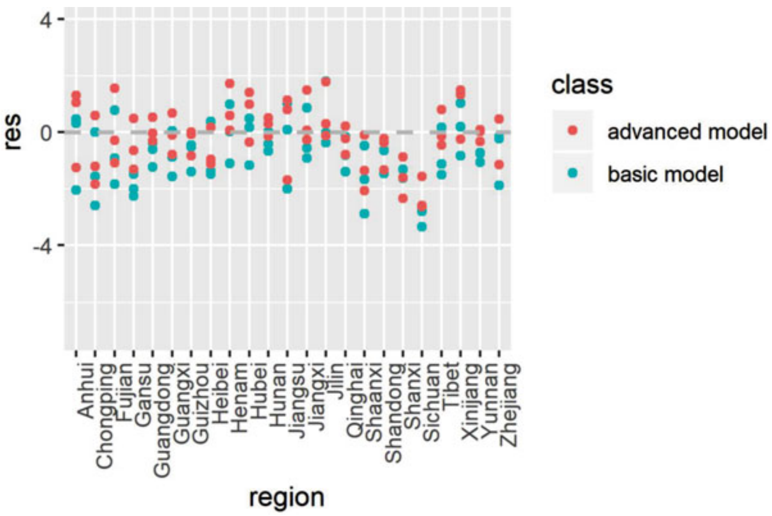

Fig. 3 Residuals of the two models in the testing set

number of prevention projects in previous years. Although discussing the results as related to the effectiveness of prevention measures goes far beyond the scope of this work, results show, in accordance with the findings of statistical methods, that it is possible to identify evidence of a connection between the prevention projects and a reduction in the number of disasters.

References

1. Akter, S., Wamba, S.F.: Big data and disaster management: a systematic review and agenda for future research. Annals of Operations Research pp. 1–21
2. Alpaydin, E.: Introduction to machine learning. MIT press (2014)
3. Chen, S.C., Ferng, J.W., Wang, Y.T., Wu, T.Y., Wang, J.J.: Assessment of disaster resilience capacity of hillslope communities with high risk for geological hazards. Engineering geology **98**(3-4), 86–101 (2008)
4. Fiore, U., De Santis, A., Perla, F., Zanetti, P., Palmieri, F.: Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. Information Sciences (2017)
5. Fiore, U., Marino, Z., Passalacqua, L., Perla, F., Scognamiglio, S., Zanetti, P.: Tuning a deep learning network for solvency ii: Preliminary results. In: Mathematical and Statistical Methods for Actuarial Sciences and Finance, pp. 351–355. Springer (2018)
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
7. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural networks **2**(5), 359–366 (1989)
8. Huang, Y., Cheng, H.: The impact of climate change on coastal geological disasters in southeastern china. Natural hazards **65**(1), 377–390 (2013)
9. James, G., Witten, D., Hastie, T., Tibshirani, R.: An introduction to statistical learning, vol. 112. Springer (2013)
10. Li, X., Yao, J., Li, Y., Zhang, Q., Xu, C.Y.: A modeling study of the influences of yangtze river and local catchment on the development of floods in poyang lake, china. Hydrology Research **47**(S1), 102–119 (2016)
11. Pfaff, B.: VAR, SVAR and SVEC models: Implementation within R package vars. Journal of Statistical Software **27**(4) (2008), http://www.jstatsoft.org/v27/i04/
12. Zeileis, A.: dynlm: Dynamic Linear Regression (2019), https://CRAN.R-project.org/package=dynlm, r package version 0.3-6
13. Zhang, L., Li, G.j., Zhou, Z.g., WANG, Q.: Grey clustering method-based zoning assessment of regional geological disaster. J. Nat. Disasters **18**, 164–168 (2009)