Rural Household Landslide Loss Estimation in Peru: An Exploratory Neural-Network Approach

Miguel Fernández-Perez
Pontifical Catholic University of Peru
mfernandezp@pucp.edu.pe

Jorge Vargas-Florez
Pontifical Catholic University of Peru
jorge.vargas@pucp.edu.pe

Noam D.V. Lopez-Villanes
Pontifical Catholic University of Peru
lopez.noam@pucp.edu.pe

Esteli C. Vela-Cacéres
Pontifical Catholic University of Peru
evela@pucp.pe

Martijn Wester
Wageningen University
m.wester99@gmail.com

ABSTRACT

This work-in-progress paper proposes an exploratory Neural-Network framework to estimate household-level economic losses from debris-flow-induced landslides in two rural communities of Carabayllo, Lima, Peru. The study integrates geophysical, geographical, environmental, and sociodemographic information from georeferenced household surveys and geomorphological outputs generated with the LAPSUS model (v6.03). The main challenge is a small, heterogeneous, and partially interdependent dataset (N = 60 households) comprising both discrete and continuous variables. To address this complexity, the framework combines autoencoder-based dimensionality reduction for discrete variables, transformed feature-space construction, unsupervised segmentation, and neural-network regression. Rather than claiming validated predictive performance, the paper presents this architecture as a proof of concept for structurally complex disaster datasets under severe sample constraints. It contributes methodologically by showing how representation learning and segmentation may support loss estimation, and empirically by highlighting the close relationship between social and housing conditions and household disaster losses in vulnerable peri-urban settings.

Keywords

Landslides, exploratory modeling, neural networks, El Niño phenomenon, rural household.

INTRODUCTION

The impacts of natural hazards on poor communities are largely invisible in aggregated economic statistics, as their assets and incomes represent only a small fraction of national wealth (Hallegatte et al., 2018). Nevertheless, these communities are disproportionately affected by such events (Arnold & de Cosmo, 2015; Hallegatte et al., 2017). Post-disaster crisis conditions in rural areas differ significantly from those in urban contexts (Kapucu et al., 2013). In particular, "poor households take longer to recover from disasters and are more likely to face long-term consequences" (Hallegatte & Walsh, 2021, p. 1). Several factors explain the severity of hazard impacts on poor communities. Arnold and de Cosmo (2015) and Hallegatte (2012) argue that these populations are more exposed to natural hazards because they often settle in marginal or unsafe areas. Their vulnerability is further exacerbated by precarious housing conditions and insecure property rights, which discourage investment in risk reduction. In addition, limited post-disaster coping capacity constrains their ability to absorb and recover from losses (Hallegatte et al., 2020). Financial exclusion—such as the absence of bank accounts—and restricted access to information technologies further hinder access to government assistance. However, Kapucu et al. (2010) highlight the effectiveness of local capacities in disaster response, often surpassing formal intergovernmental operations. Community resilience is strengthened by collective organization, institutional linkages with government and non-profit organizations, and social capital, all of which facilitate effective recovery and restoration to pre-disaster conditions (United Nations International Strategy for Disaster Reduction [UNISDR], 2009; Kapucu et al., 2013).

According to Peres et al. (2012) and Braman et al. (2013), disaster recurrence can be predicted based on historical records and past disaster performance, using historical data to characterize vulnerability and exposure conditions. Vargas et al. (2016) highlight that so-called "recurrent disasters," such as El Niño events, frost episodes, and mass movements (called huaicos in Peru), can be particularly estimated through such approaches. In this context, Ferris et al. (2013) define "recurrent disaster" as the repeated occurrence of a single natural hazard within the same geographic region. Recent research further emphasizes that historical data are essential for modelling spatial and temporal patterns of disaster occurrence and severity, thereby informing effective mitigation and response policies (United Nations Office for Disaster Risk Reduction [UNDRR], 2022; Centre for Research on the Epidemiology of Disasters [CRED], 2021). The systematization of past events enables the identification of trends, estimation of future occurrence probabilities, and anticipation of economic, social, and environmental impacts. These needs have led to the development of open-access platforms such as DesInventar and EM-DAT, which compile information on natural and anthropogenic events and their impacts. These databases are widely used by governments and international organizations for disaster risk reduction planning (UNDRR, 2022).

The economic effects of natural disasters are typically mediated by the number of affected individuals, human losses, and asset destruction—indicators recorded in the aforementioned databases. The literature identifies multiple determinants of these economic impacts, including disaster intensity, socioeconomic characteristics, vulnerability-related factors, and institutional variables. To determine which factors, explain economic effects, statistical and econometric models are employed (Noy & Yonson, 2018). These models allow for the control of contextual variables and the estimation of counterfactual scenarios, thereby enabling more accurate measurement of disaster impacts. Time-series, cross-sectional, and panel data models, as well as machine learning techniques applied to historical datasets, have yielded valuable insights for predicting economic losses (Anbarci et al., 2005; Gaiha et al., 2015; Hochrainer-Stigler et al., 2020; Kellenberg & Mobarak, 2008; Zhang et al., 2024; Liu et al., 2024). Algorithms such as Random Forest, Support Vector Machines, and Artificial Neural Networks exhibit strong capabilities for handling large volumes of data and identifying complex patterns those traditional statistical methods may fail to capture. These approaches have been successfully applied to landslide susceptibility mapping and risk estimation in mountainous regions characterized by multiple interacting physical and social variables (Feng et al., 2020; Linardos et al., 2022).

Existing research on disaster-related economic losses has predominantly relied on aggregated datasets at national, regional, or provincial levels, using statistical and machine learning techniques to model macro-level impacts and hazard susceptibility. While these approaches have demonstrated strong predictive capabilities, they largely overlook the micro-level heterogeneity of vulnerability and loss processes at the household scale, particularly in rural and peri-urban contexts. In parallel, landslide studies have focused primarily on susceptibility mapping based on physical and geomorphological factors, with limited integration of socioeconomic and resilience-related variables into loss estimation frameworks. Moreover, most machine learning applications in disaster contexts assume relatively large and statistically representative datasets, leaving a methodological gap in handling small, heterogeneous, and structurally interdependent datasets derived from fieldwork. This study addresses these gaps by proposing an exploratory pipeline that integrates geomorphological process indicators with household-level vulnerability data, explicitly designed for small-sample conditions and aimed at organizing complex socio-environmental information for loss estimation rather than claiming validated predictive performance. Rather than extending existing large-scale predictive approaches, this study focuses on structuring and understanding complex micro-level data under severe sample constraints.

VULNERABLE COMMUNITY SELECTION

As a first step, the study identified the community most vulnerable to landslides. According to information retrieved from Peru's national open disaster-impact database, the National Information System for Response and Rehabilitation (SINPAD; https://sinpad2.indeci.gob.pe/sinpad2/faces/public/listSinpadEnviadosPubli.xhtml), managed by the National Institute of Civil Defense (INDECI), one of the populations most severely affected by landslides during the 2017 Coastal El Niño event was the rural community of Huatocay, located in the district of Carabayllo on the left bank of the Chillón River. This information was complemented with additional reports from the Geological, Mining and Metallurgical Institute (INGEMMET) and the National Emergency Operations Center (COEN), particularly Emergency Report No. 323-20/03/2017/COEN, which indicates that intense rainfall also severely affected Río Seco, located on the right bank of the Chillón River, causing fatalities and significant damage to housing (see Figure 1). Based on this evidence, the research team established as a starting point a comprehensive landslide-susceptibility assessment of these peri-urban areas. Fieldwork activities included coordination with local authorities and residents to obtain support in identifying suitable sites for analysis.

Figure 1. Huatocay and Rio Seco location.

CHARACTERIZATION VULNERABILITIES

In this research, the framework of LAPSUS (LandscApe ProceS modelling at mUlti-dimension and Scales) is used. A master's student in geology from Wageningen University (WUR) was recruited to conduct the geomorphological characterization. The model can predict landscape evolution due to landslides on coarse temporal and spatial scales with limited data requirements (De Sy et al., 2013). The LAPSUS-LS (v6.03) model gives the following outputs: (1) A critical rainfall map classified according to the minimum steady state rainfall needed to trigger landslides, (2) If the input threshold is passed: a map of eroded material in the catchment and deposition of material and (3) a log file with the total amount of soil displaced by landslides and the amount of soil deposited inside the catchment. The difference between the total amount of soil displaced and the amount of deposited in the area represents the amount of soil that was deposited outside the modelled catchment (Rossi et al., 2017). This work estimates of the erosion and sedimentation assessment associated with landslide processes under prospective El Niño scenarios.

To determine the socioeconomical characteristics of community vulnerabilities, a structured questionnaire was developed based on a comprehensive review of key literature on vulnerability and resilience. The primary references included Vargas-Florez and Lauras (2021), who examine resilience and vulnerability measurement frameworks, and Hallegatte et al. (2017), who analyze the socioeconomic impacts of disasters on poor communities. Additionally, the work of Popovici et al. (2013) and the methodological guidelines issued by the Centro Nacional de Planeamiento Estratégico (CEPLAN, 2019) provided a solid foundation for developing indicators and survey questions aimed at capturing both vulnerability and resilience among at-risk populations. These studies, together with the conceptual recommendations of Matyas and Pelling (2012), supported the operationalization of vulnerability into five key dimensions: social, infrastructural/environmental, economic, social safety networks, and institutional governance.

The dependent variable, termed "economic impact on vulnerable populations", constituted the central axis of the questionnaire. Independent variables were organized into three broad areas: vulnerability assessment, resilience assessment, and socioeconomic impact assessment. These areas were further operationalized into dimensions, evaluation criteria, indicators, and specific survey questions, following the literature review and subsequent adaptation to local community conditions. The vulnerability dimension focused on sociodemographic characteristics. The resilience dimension included access to resources, structural or functional vulnerability, population vulnerability, and humanitarian response capacity. The impact dimension addressed effects on life and assets, economic losses, and productive impacts.

The questionnaire comprised 46 questions and was developed through a comparative process with prior studies, particularly those of Matyas and Pelling (2012) and Popovici et al. (2013), whose resilience and vulnerability measurement approaches served as methodological benchmarks. Common indicators and elements were identified and subsequently adapted to the local context following preliminary field visits aimed at gaining deeper insight into the case studies. A non-probabilistic sampling strategy was adopted to take the questionnaire. They were administered until reaching a saturation point, determined according to the internal variation observed across the identified zones in each case. Prior to data collection, interviews were conducted with municipal authorities and key informants to identify initial access points and define the specific zones for each case study. A virtual questionnaire was designed to facilitate data collection, available at: https://form.jotform.com/240626834772058. A detailed survey administration guide, including the operational plan, was also developed in order to track and audit research procedure. To ensure comprehensive variable characterization, multiple question formats were employed, including categorical, numerical, and ordinal items, allowing the collection of both quantitative and qualitative data. Categorical questions facilitated classification of responses; numerical questions enabled precise measurement of damage magnitude and income losses; and ordinal questions were essential for assessing perceptions of risk and response capacity. This mixed-method questionnaire design enhanced analytical depth and supports subsequent comparisons across households and communities.

APPLIED METHODOLOGY

This work-in-progress study explores whether a structured neural networks workflow can effectively organize and model a small, heterogeneous, and partially interdependent household dataset for loss estimation. The proposed approach follows a five-stage pipeline: (1) fieldwork for vulnerability and resilience characterization linked to economic loss; (2) representation learning for discrete variables; (3) construction of a transformed feature space; (4) exploratory segmentation; and (5) nonlinear regression under small-sample constraints. The final stage is explicitly framed as exploratory rather than operational forecasting. A formal validation stage is planned as part of future work.

Figura 2. Five-stage research methodology

COMMUNITIES FIELDWORK

After the desk work, where the literature review, survey design, collection of secondary scientific information and information from key actors in the study area were carried out. A three-phase fieldwork was run:

- In the first phase, coordination meetings were conducted with local authorities, including representatives from rural communities and the Municipality of Carabayllo. These efforts were supported by the PUCP Crisis and Disaster Management Research Group.
- In the second phase, processed imagery with pre-existing hazard conditions—such as cliffs, unstable slopes, soil types, and ravines—to identify susceptibility patterns and populations potentially exposed to rainfall-triggered landslides, including those associated with El Niño events. To ensure data consistency across Digital Elevation Models (DEMs), TanDEM-X datasets were adopted as the primary source (12-meter resolution).
- In the third phase, sociodemographic data characterization of communities facing hazard exposure. This component aimed to build upon existing research on rural community resilience and social capital.

Fieldwork was conducted across multiple sessions between May and June 2024.

DATASET

Source of Information

The data used in this research were obtained from surveys and the landslides simulation system, LAPSUS LS. Each observation corresponds to a household georeferenced through latitude and longitude coordinates.

Sociodemographic, structural, economic, and environmental information was collected, along with the monetary estimation of the loss incurred by the household in the event of a landslide. The final modeling dataset consists of 24 predictor variables (XV1–XV24) and one response variable (YR).

Predictor Variables

The variables were organized into four analytical dimensions: (1) Sociodemographic variables, (2) Housing conditions and basic services, (3) Economic and labor conditions and (4) Environmental variables.

Table 1. Sociodemographic Variables

Variable	Description	Type
XV1	Gender of the head of household	Binary
XV2-XV3	Age of the respondent (categorical)	Binary
XV4	Total number of household members	Integer
XV5	Number of household members aged 65 years or older	Integer
XV6	Number of children under 10 years of age	Integer
XV7	Number of household members with access to health insurance	Integer
XV8	Number of household members with chronic illnesses	Integer
XV9	Number of household members with technical or university education	Integer
XV10	Number of household members who are illiterate	Integer
XV1	Gender of the head of household	Binary

Table 2. Housing Conditions and Basic Services

Variable	Description	Type
XV11-XV12	Main construction material of the dwelling walls (categorical)	Binary
XV13-XV15	Access to basic services (categorical)	Binary

Table 3. Economic and Labor Conditions

Variable	Description	Type
XV16	Approximate monthly household income	Continuous
XV17	Number of household members with formal employment	Integer
XV18	Number of household members with informal employment	Integer

Table 4. Environmental Variables

Variable	Description	Type
XV19	Cumulative erosion	Continuous
XV20	Mean erosion	Continuous
XV21	Standard deviation of erosion	Continuous
XV22	Cumulative sedimentation	Continuous
XV23	Mean sedimentation	Continuous
XV24	Standard deviation of sedimentation	Continuous

Response Variable

As discussed in previous sections, disaster impact may be operationalized through multiple quantitative indicators capturing different dimensions of disruption. Infrastructure impact can be measured through metrics such as the percentage of damaged dwellings or the total volume of debris generated. Economic impact may be assessed through reductions in productive activity, loss of sales, or income contraction. Human impact includes indicators such as the number of fatalities and livestock losses. Logistical impact may be reflected in the reduction of circulating vehicles or the percentage loss of storage capacity, among other operational performance measures.

These indicators collectively represent both direct and indirect consequences of hazard events. However, for the purposes of the present study, the selected outcome variable corresponds to the direct household-level economic loss associated with the most recent major landslide event triggered by the El Niño phenomenon in the study area. This variable provides a measurable and comparable proxy of disaster impact at the microeconomic level, allowing for the modeling of vulnerability and loss estimation within a nonlinear predictive framework.

Table 5. Response Variable

Variable	Description	Type
YR	Approximate reported economic loss (Peruvian currency)	Continuous

It was constructed a nonlinear regression model capable of estimating the economic loss reported by households (YR) from a dataset of (N = 60) observations containing heterogeneous sociodemographic, structural, economic, and environmental variables. The dataset presents two structural characteristics that condition the modeling process:

- A high proportion of discrete variables (binary and integer), many derived from intra-household counts (e.g., number of individuals with a given characteristic).
- Structural dependency among variables, as several originate from the same informational core. For example, the total number of household members is naturally related to the number of older adults, children, insured individuals, or persons with chronic illnesses. Similarly, categorical housing and service variables are derived from the same questionnaire module.

This interdependence implies that the predictors are not fully independent, potentially generating informational redundancy, implicit multicollinearity, and model overparameterization if all original variables are directly used. During an initial exploratory phase, a neural network was trained using the 24 original predictor variables. However, the model exhibited instability during training and poor predictive performance, suggesting that the input space required prior transformation to reduce structural complexity while preserving relevant information.

The exploratory analysis indicates that the transformed input space is not homogeneous and that segmentation may be analytically useful. The clustering step identifies two internally more coherent subsets of households, which is consistent with the field evidence that Huatocay and Rio Seco differ in risk perception, social priorities, and territorial dynamics.

Dimensionality Reduction via Autoencoders

This stage consisted of transforming the subset of discrete variables (binary and integer) using an autoencoder model, with the objective of reducing the effective dimensionality of the input space while preserving its informational structure. From a theoretical standpoint, an autoencoder is an unsupervised neural network that learns a compressed representation of the data. Let $x \in \mathbb{R}^n$ denote the input vector. The model is composed of two parameterized functions:

$$z = f_\theta(x)$$
$$\hat{x} = g_\phi(z)$$

where:

- f_θ is the encoder that maps the original vector into a latent representation $z \in \mathbb{R}^m$, with (m < n);
- g_ϕ is the decoder that reconstructs an approximation $\hat{x}$ of the original vector;
- θ and ϕ represent the parameters learned during training.

The model is trained by minimizing the reconstruction error defined through the mean squared error:

$$\mathcal{L}(\theta, \phi) = \frac{1}{N}\sum_{i=1}^{N} \|x_i - \hat{x}_i\|^2$$

When the latent dimension (m) is smaller than the original dimension (n), the model is forced to capture the most relevant structural dependencies among variables, eliminating redundancy and noise.

In the present study, 17 discrete variables (XV1–XV15, XV17, and XV18) were selected, forming a matrix $X \in \mathbb{R}^{60\times17}$. These variables include sociodemographic, structural, and labor information characterized by potential internal relationships (e.g., household-derived counts). Continuous variables (monthly income and environmental indicators) were excluded at this stage, as their metric scale provides direct quantitative information relevant for subsequent regression modeling.

A deep autoencoder architecture with multiple fully connected (dense) layers was implemented for both encoder and decoder components. The encoder progressively reduced the 17-dimensional input to an 8-dimensional latent representation through intermediate layers of 100, 80, 60, 40, and 20 neurons. The decoder mirrored this structure in reverse order to reconstruct the original dimensionality. Parameter optimization was performed using the Adam (Adaptive Moment Estimation) is a first-order, gradient-based optimization algorithm, with training extending up to 10000 epochs (an epoch is one complete pass through the entire training dataset). This configuration allows the model to capture higher-order nonlinear relationships among discrete variables. However, given the limited sample size, these representations should be interpreted as exploratory rather than definitive.

Variable Accuracy

The model was evaluated by comparing the original and reconstructed values. Since the variables are discrete, the reconstructed values were rounded before calculating the accuracy. The overall reconstruction accuracy obtained was 95.49%. The accuracy per variable is presented in the following table:

Table 6. Variables Accuracy

Variable	Type
XV1	85.00%
XV2	96.67%
XV3	96.67%
XV4	100.00%
XV5	95.00%
XV6	100.00%
XV7	98.33%
XV8	100.00%
XV9	100.00%
XV10	86.67%
XV11	100.00%
XV12	100.00%
XV13	91.67%
XV14	88.33%
XV15	98.33%
XV17	88.33%
XV18	98.33%

These results indicate that compressing 17 variables into 8 latent dimensions to try preserves the structural information of the original dataset. The encoder produced the latent representation:

$$Z = (z_1, z_2, \dots, z_8)$$

These eight latent variables replaced the 17 original discrete variables. Subsequently, they were concatenated with the continuous variables (monthly income and environmental indicators), forming the new input space used in the segmentation and regression stages. This transformation reduced effective dimensionality, mitigated structural redundancy, and improved predictive stability.

Exploratory Analysis and Segmentation of the Transformed Space

Once the new continuous input space was constructed, an exploratory analysis was conducted to examine the relationship between the transformed explanatory variables and the response variable (YR).

Scatter plots were generated to visualize the distribution of observations within the feature space and to identify potential internal structures, as illustrated in the corresponding Figure 2.

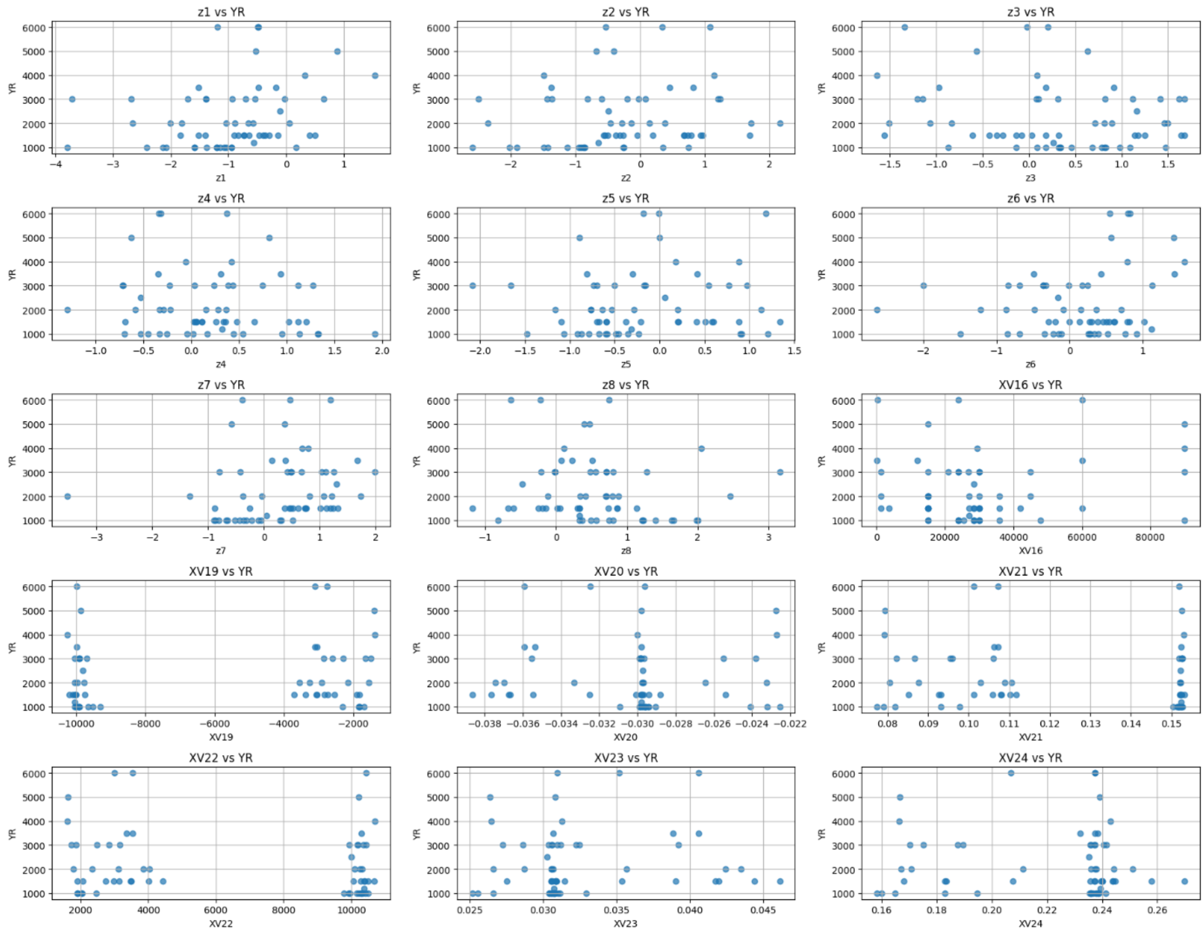

Figure 3. Dispersion Between Input Variables and YR

Visual inspection of the scatter plots indicates that the observations are not homogeneously distributed within the feature space. In particular, two distinct concentrations of records can be identified. These groupings exhibit relatively compact behavior within each subset and a noticeable separation between them. The presence of these internal structures suggests that the phenomenon under study cannot be adequately described by a single global model capturing a uniform functional relationship across the entire sample. Under this context, it was deemed appropriate to apply an segmentation technique to partition the dataset into more homogeneous subsets prior to regression modeling. To this end, the K-means algorithm was implemented over the transformed input space.

The objective of the algorithm is to minimize within-cluster variability by assigning each observation to the nearest centroid in terms of Euclidean distance, solving the following optimization problem:

$$min \sum_{j=1}^{k} \sum_{x_i \in C_j} \left\| x_i - \mu_j \right\|^2$$

Where (C_j) denotes the set of observations assigned to cluster (j), and (μ_j) represents its corresponding centroid.

Based on the graphical evidence observed previously, (k = 2) clusters were selected. However, the choice of k = 2 remains exploratory and has not been validated through formal clustering validation metrics. The resulting segmentation produced two balanced groups, each consisting of 30 observations.

Regression Modeling Using Neural Networks

After segmenting the dataset into two clusters, independent regression models were trained for each group using artificial neural networks. Artificial neural networks can be interpreted as nonlinear function approximators. In a regression setting, the objective is to estimate the relationship between an input vector of explanatory variables and a continuous output variable. Let ($x \in \mathbb{R}^n$) denote the input vector, and let ($y \in \mathbb{R}$) denote the observed output variable. The neural network produces an estimate ($\hat{y}$) given by:

$$\hat{y} = F(x, W, b)$$

Where (W) and (b) represent the weights and biases adjusted during training. The function (F) arises from the composition of successive linear transformations and nonlinear activation functions across the network layers. This structure allows the model to approximate the functional relationship among explanatory variables, preliminary capturing interactions and combined effects that cannot be adequately represented by traditional linear models. This capability is particularly relevant in the present study, given the structural heterogeneity and nonlinear dependencies inherent in the dataset. However, a validation-stage is pending to be run in an extension of this work. For each cluster, a fully connected multilayer neural network was executed. The input layer dimension corresponded to the total number of variables in the transformed feature space (14 variables). From this initial layer, information was propagated to a first hidden layer consisting of 64 neurons with ReLU activation, followed by a second hidden layer of 32 neurons, also using ReLU activation. This progressive reduction in the number of neurons allowed the model to approach preliminary to understand nonlinear interactions among variables while controlling representational complexity. Finally, an output layer composed of a single neuron with linear activation was incorporated, appropriate for estimating the continuous response variable (YR).

Model training was conducted by minimizing the mean squared error (MSE) as the loss function, while mean absolute error (MAE) was used as a complementary evaluation metric. Parameter optimization was performed using the Adam algorithm, with training extending up to a maximum of 10,000 epochs. This configuration allows the model to approximate the functional relationship the model to approach the functional relationship between explanatory variables and economic loss within each cluster, achieving near-zero error levels at the final training iterations. The evolution of the training process for each cluster is presented in the following tables:

Table 7. Training Evolution – Cluster 1

Epoch	MSE	MAE
500	515,344.5938	487.8517
1000	166,734.4375	251.9855
1500	7,109.3423	49.1457
2000	18.0698	2.1298
2500	0.1944	0.3476
3000	0.0023	0.0310
4000	0.0001	0.0047
10000	≈ 0.0000	≈ 0.0000

Table 8. Training Evolution – Cluster 2

Epoch	MSE	MAE
500	802,037.1875	698.4172
1000	368,365.9375	467.8779
1500	185,152.0000	310.7194
2000	51,126.7656	144.9456
3000	523.7507	11.3459
3500	0.9239	0.4472
4000	0.0000	0.0010
10000	≈ 0.0000	≈ 0.0000

In both cases, the training error decreases consistently throughout the optimization process, reaching near-zero values in later iterations. While this behavior indicates that the neural networks are capable of fitting the observed data, it should not be interpreted as evidence of generalization performance. Given the small sample size (30 observations per cluster) and the complexity of the model, the results are likely indicative of overfitting. Therefore, the findings should be interpreted strictly as evidence of methodological feasibility rather than predictive accuracy. In this revised interpretation, the results suggest that the network has sufficient flexibility to closely fit the observed sample; however, this should not be considered evidence of predictive capability.

CONCLUSION

Conducting the research in rural communities was a great challenge for the team. One of the main limitations to carry this study was the low rate of fully completed questionnaires. Several respondents omitted questions due to time constraints or recall difficulties. Telephone-administered surveys frequently resulted in missing key data, while self-administered questionnaires were often partially completed. Limited internet connectivity—particularly in Huatocay and parts of Río Seco—further restricted the effective implementation of online surveys. This required conducting multiple follow-up field visits in order to complete the missing information. Community interests differed notably. Huatocay residents expressed greater concern regarding risk management and environmental hazards such as landslides, whereas Río Seco residents emphasized social issues, including land trafficking.

This work-in-progress paper presents an exploratory framework for organizing and modeling household-level landslide-loss data in a vulnerable peri-urban context. The study integrates survey-based vulnerability information and LAPSUS-derived environmental indicators into a five-stage pipeline that combines vulnerability and resilience characterization at the community level with representation learning, transformed feature construction, clustering, and nonlinear regression. Its main contribution is methodological.

At the empirical level, the study reinforces the relevance of household social conditions in understanding disaster impacts. The data structure suggests that losses are not explained by geomorphological exposure alone, but by the interaction between physical hazard processes and pre-existing social vulnerability. Yet the present study does not claim validated forecasting performance. Because the sample is small, non-probabilistic, and limited to a single event context, the current results should be interpreted as exploratory and context-bound. The near-zero training errors observed in the original models do not demonstrate generalization and may reflect overfitting.

From a theoretical perspective, the model contributes to the vulnerability–resilience understanding by empirically demonstrating that social determinants play a central role in shaping disaster impact magnitude. The characterization of causal variables associated with landslide-induced losses seems enables the identification of factors whose variation increases or mitigates economic impact. Variables such as housing construction quality, access to education, family planning, and household income improvements show a clear inverse relationship with loss magnitude. While geomorphological and physical exposure conditions remain relevant, the results suggest that improvements in social conditions may counterbalance physical and geographical vulnerability. This finding reinforces the argument that resilience is not solely determined by environmental constraints but is strongly mediated by social and economic capacities.

FUTURE WORK

Future work should prioritize methodological validation before any operational application. While neural networks approaches show potential for capturing nonlinear dependencies and structural interrelations, their robustness must be systematically assessed. Complementary econometric techniques should be incorporated to evaluate coefficient stability, test causal assumptions, and compare predictive and explanatory performance.

On the empirical side, the dataset should be expanded spatially and temporally. The current outcome variable is based on losses associated with a recent major event, which limits the temporal scope of the analysis. Repeated observations across events and over time would make it possible to model recovery dynamics, compare event severities, and distinguish structural vulnerability from temporary post-disaster conditions. The sample size (N = 60) and the subdivision into two clusters (30 observations per cluster) limit the generalization capacity of the trained models. Future research should incorporate external validation datasets to avoid overfitting, cross-validation techniques, or regularization strategies to further assess model generalization.

Additionally, expanding the survey base across multiple localities would improve statistical robustness and external validity. It is also recommended that data collection be conducted periodically. Unlike vulnerability, which is often treated as a static condition, resilience represents a dynamic social construct that evolves over time in response to institutional, economic, and social transformations. Longitudinal data would allow for the modeling of temporal dynamics and causal trajectories rather than static estimations. There is also an imperative need to design shorter and more efficient data collection instruments, as well as to adopt alternative tools to ensure the offline capture of geospatial data.

ACKNOWLEDGMENTS

This research was funded by Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica (CONCYTEC) through the PROCIENCIA program under the "Applied Research Projects 2023-02" call, in accordance with Contract No. PE501082559-2023.

REFERENCES

Anbarci, N., Escaleras, M., & Register, C. A. (2005). Earthquake fatalities: The interaction of nature and political economy. *Journal of Public Economics, 89*(9-10), 1907–1933. https://doi.org/10.1016/j.jpubeco.2004.08.002

Arnold, M., & de Cosmo, S. (2015). *Building social resilience: Protecting and empowering those most at risk*. Global Facility for Disaster Reduction and Recovery, World Bank. https://www.gfdrr.org/sites/default/files/publication/Building-Social-Resilience-Protecting-and-Empowering-Those-Most-at-Risk.pdf

Braman, L. M., van Aalst, M. K., Mason, S. J., Suarez, P., Ait-Chellouche, Y., & Tall, A. (2013). Climate forecasts in disaster management: Red Cross flood operations in West Africa, 2008. *Disasters, 37*(1), 144–164. https://doi.org/10.1111/j.1467-7717.2012.01297.x

Centre for Research on the Epidemiology of Disasters. (2021). *2021 disasters in numbers: Extreme events defining our lives*. https://www.cred.be/sites/default/files/2021_EMDAT_report.pdf

Centro Nacional de Planeamiento Estratégico. (2019). *Vulnerabilidad de las personas en el territorio: Desde una perspectiva socioeconómica*. https://www.ceplan.gob.pe/documentos_/vulnerabilidad-de-las-personas-en-el-territorio/

De Sy, V., Schoorl, J. M., Keesstra, S. D., Jones, K. E., & Claessens, L. (2013). Landslide model performance in a high resolution small-scale landscape. *Geomorphology*, 190, 73–81. https://doi.org/10.1016/j.geomorph.2013.02.012

Feng, Q., Liu, J., Gong, J., & Liu, M. (2020). Landslide susceptibility assessment using random forest combined with grid search and cross-validation: A case study in the Three Gorges Reservoir area, China. *Geomorphology*, 358, 107124. https://doi.org/10.1016/j.geomorph.2020.107124

Ferris, E., Petz, D., & Stark, C. (2013). *The year of recurring disasters: A review of natural disasters in 2012*. Brookings Institution. https://www.brookings.edu/articles/the-year-of-recurring-disasters-a-review-of-natural-disasters-in-2012/

Gaiha, R., Hill, K., Thapa, G., & Kulkarni, V. S. (2013). *Have natural disasters become deadlier?* (Working Paper No. 181). Brooks World Poverty Institute, University of Manchester. https://doi.org/10.2139/ssrn.2209714

Hallegatte, S. (2012). *An exploration of the link between development, economic growth, and natural risk* (Policy Research Working Paper No. 6216). World Bank. https://openknowledge.worldbank.org/server/api/core/bitstreams/a85e448c-7439-5505-94e0-093fd8693018/content

Hallegatte, S., & Walsh, B. (2021). *Natural disasters, poverty, and inequality: New metrics for fairer policies*. World Bank. https://openknowledge.worldbank.org/server/api/core/bitstreams/05c3ef50-de39-515b-b8fe-41d96e163f02/content

Hallegatte, S., Fay, M., & Barbier, E. (2018). Poverty and climate change: Introduction. *Environment and Development Economics, 23*(3), 217–233. https://doi.org/10.1017/S1355770X18000141

Hallegatte, S., Vogt-Schilb, A., Bangalore, M., & Rozenberg, J. (2017). *Unbreakable: Building the resilience of the poor in the face of natural disasters*. World Bank. https://doi.org/10.1596/978-1-4648-1003-9

Hallegatte, S., Vogt-Schilb, A., Rozenberg, J., Bangalore, M., & Beaudet, C. (2020). From poverty to disaster and back: A review of the literature. *Economics of Disasters and Climate Change, 4*(1), 223–247.

https://doi.org/10.1007/s41885-020-00060-5

Hochrainer-Stigler, S., Laurien, F., Velev, S., Keating, A., & Mechler, R. (2020). Standardized disaster and climate resilience grading: A global scale empirical analysis of community flood resilience. *Journal of Environmental Management, 276*, 111332. https://doi.org/10.1016/j.jenvman.2020.111332

Kapucu, N., Arslan, T., & Collins, M. L. (2010). Examining intergovernmental and interorganizational response to catastrophic disasters: Toward a network-centered approach. *Administration & Society, 42*(2), 222–247. https://doi.org/10.1177/0095399710362517

Kapucu, N., Hawkins, C. V., & Rivera, F. I. (2013). Disaster preparedness and resilience for rural communities. *Risk, Hazards & Crisis in Public Policy, 4*(4), 215–233. https://doi.org/10.1002/rhc3.12043

Kellenberg, D. K., & Mobarak, A. M. (2008). Does rising income increase or decrease damage risk from natural disasters? *Journal of Urban Economics, 63*(3), 788–802. https://doi.org/10.1016/j.jue.2007.05.003

Linardos, V., Mavrouli, O., Papathanassiou, G., & Bathrellos, G. D. (2022). Machine learning techniques for landslide susceptibility mapping: A review. *Earth-Science Reviews, 232*, 104097. https://doi.org/10.1016/j.earscirev.2022.104097

Liu, C.-F., Huang, L., Yin, K., Brody, S., & Mostafavi, A. (2024). *FloodDamageCast: Building flood damage nowcasting with machine learning and data augmentation*. arXiv. https://doi.org/10.48550/arXiv.2405.14232

Matyas, D., & Pelling, M. (2012). *Disaster vulnerability and resilience: Theory, modelling and prospective*. Government Office for Science.

Noy, I., & Yonson, R. (2018). Economic vulnerability and resilience to natural hazards: A survey of concepts and measurements. *Sustainability, 10*(8), 2850. https://doi.org/10.3390/su10082850

Peres, E. Q., Brito Junior, I. de, Leiras, A., & Yoshizaki, H. T. Y. (2012). Humanitarian logistics and disaster relief research: Trends, applications, and future research directions. In *Proceedings of the 4th International Conference on Information Systems, Logistics and Supply Chain* (ILS 2012), Quebec City, Canada.

Popovici, E.-A., Costache, A., Bălteanu, D., Dogaru, D., & Sima, M. (2013). Vulnerability assessment of rural communities to floods in the western part of Romania (Banat Plain). In *Proceedings of the 13th International Multidisciplinary Scientific GeoConference SGEM 2013* (Vol. 1, p. 1161).

Qiu, C., Su, L., Pasuto, A., Bossi, G., & Geng, X. (2024). Economic risk assessment of future debris flows by machine learning method. *International Journal of Disaster Risk Science, 15*(1), 149–164. https://doi.org/10.1007/s13753-024-00545-x

Rossi, L. M. W., Rapidel, B., Roupsard, O., Villatoro-Sánchez, M., Mao, Z., Nespoulous, J., & Stokes, A. (2017). Sensitivity of the landslide model LAPSUS_LS to vegetation and soil parameters. *Ecological Engineering, 109*, 249–255. https://doi.org/10.1016/j.ecoleng.2017.08.010

United Nations International Strategy for Disaster Reduction. (2009). *UNISDR terminology on disaster risk reduction*. United Nations. https://www.unisdr.org/files/7817_UNISDRTerminologyEnglish.pdf

United Nations Office for Disaster Risk Reduction. (2022). *Global assessment report on disaster risk reduction 2022: Our world at risk: Transforming governance for a resilient future*. United Nations. https://www.undrr.org/gar2022-our-world-risk

Vargas, J., Rojas, J. E., Inga, A. J., Mantilla, W. A., Añasco, H. J., Basurto, M. F., Campos, R. J., Sanchez, J. A., & Checa, P. I. (2016). Towards reliable recurrent disaster forecasting methods: Peruvian earthquake case. In *Proceedings of the 13th International Conference on Information Systems for Crisis Response and Management*, Brazil. http://idl.iscram.org/files/jorgevargas/2016/1354_JorgeVargas_etal2016.pdf

Zhang, M., Wang, G., Li, X., & Du, C. (2024). Economic risk assessment of future debris flows by machine learning method. *International Journal of Disaster Risk Science*, 15, 149-164. https://link.springer.com/article/10.1007/s13753-024-00545-x