Available online at www.eccomasproceedia.org
Eccomas Proceedia COMPDYN (2023) 2677-2686

ECCOMAS
Proceedia

COMPDYN 2023
9th ECCOMAS Thematic Conference on
Computational Methods in Structural Dynamics and Earthquake Engineering
M. Papadrakakis, M. Fragiadakis (eds.)
Athens, Greece, 12-14 June 2023

EMPIRICAL MULTI-VARIABLE TSUNAMI DAMAGE MODELS BASED ON THE 2011 GREAT EAST JAPAN DATASET: ANALYSIS OF THE PERFORMANCES AT DIFFERENT SPATIAL SCALES

Mario Di Bacco[1,2], Anna Rita Scorzini[1], and Anawat Suppasri[3]

[1] Department of Civil, Environmental and Architectural Engineering, University of L'Aquila, L'Aquila, Via G. Gronchi, 18, 67100 L'Aquila, Italy
annarita.scorzini@univaq.it

[2] Department of Civil and Environmental Engineering, University of Florence, Florence, Via S. Marta 3, 50139 Florence, Italy
mario.dibacco@unifi.it

[3] International Research Institute of Disaster Science, Tōhoku University
468-1 Aoba, Aramaki, Aoba-ku, Sendai, Miyagi, Japan
suppasri@irides.tohoku.ac.jp

Abstract

By implementing data-driven models for the 2011 Great East Japan earthquake and tsunami, the present study aims at investigating the effect of the level of spatial aggregation of the data on model's predictive ability and at identifying the possible existence of regional-dependent patterns affecting model's accuracy and feature importance. An extended version of the dataset compiled by the Japanese Ministry of Land, Infrastructure and Transportation (MLIT) after the 2011 event in the Tōhoku region was used to generate sub datasets at different spatial scales, ranging from individual cities of different sizes to clusters at regional and multi-regional levels. The results indicate a high variance in the accuracy for the models trained on the different subsets, with relative hit rates ranging from 0.68 to 0.89 and exhibiting a positive correlation with the cardinality of the sets, as well as some regional patterns in the prediction errors. The cluster-averaged feature importance is observed to be stable for all selections and reflects the results obtained from the models trained on the whole dataset, thus allowing a more informed identification of the most significant influencing factors for tsunami damage modelling.

Keywords: Tsunami damage, Building vulnerability, Feature importance, Machine learning, Tōhoku tsunami.

ISSN:2623-3347 © 2023 The Authors. Published by Eccomas Proceedia.
Peer-review under responsibility of the organizing committee of COMPDYN 2023.
doi: 10.7712/120123.10592.20347

1 INTRODUCTION

Fragility curves, expressing the probability of exceeding a certain damage state as a function of the features describing the hazard source and the vulnerability of the exposed objects, are the standard tools used for tsunami risk assessment. While the hazard component is generally represented by inundation depth and flow velocity (or a combination of the two), building vulnerability depends on several factors, including not only the structural and geometrical properties of the individual buildings, but also the characteristics of the surrounding environment. In particular, as regards the last aspect, coastal topography and the mutual interactions among neighboring buildings have been described to have a relevant influence on the damage mechanisms, thus making vulnerability modelling a rather complex task [1-3].

Although machine learning, data-driven models could be considered as a suitable option for tackling such complexity [4], the first applications of these approaches for empirical tsunami damage modelling are only very recent, as shown in Saengtabtim et al. [5], who developed tree-based algorithms, with only inundation depth and flow velocity as descriptive features, to study tsunami damage data for a portion (18.000 buildings in Ishinomaki city) of the dataset (containing about 250.000 damage observations) published by the Japanese Ministry of Land, Infrastructure and Transportation after the 2011 Great East Japan event in the Tōhoku region (MLIT [6]). Even more recently, Di Bacco et al. [7] integrated the MLIT dataset with additional features by including explanatory variables that are usually neglected in traditional tsunami damage models. These include building-related (i.e., shape factors) or site-related geometric parameters, calculated as a function of building position and orientation with respect to the coastline, and accounting also for the effect of the reciprocal interaction between buildings and/or singular structures, in terms of two possible opposite mechanisms: (i) a "shielding effect" generated by interacting buildings or barriers (e.g., seawalls), which may reduce water impact, and a (ii) "debris generation effect" caused by collapsed buildings that can act as an additional source of hazard.

In addition, by testing the performances of machine learning models trained with different combinations of input features, Di Bacco et al. [7] pointed out the importance of considering these mutual interactions among buildings and they suggested the implementation of more specific local models to avoid a drop in accuracy when transferring models from a region to another. By implementing data-driven ensemble decision trees on subsets of the enhanced MLIT database at different spatial scales, ranging from individual cities of different sizes to clusters at regional and multi-regional levels, this study then aims at investigating:

- the effect of the level of spatial aggregation (i.e., the cardinality of the dataset) on model's predictive ability and
- the possible existence of regional-dependent patterns in model's accuracy on unseen data and in the ranking of the feature importance on tsunami damage.

2 MATERIALS AND METHODS

2.1 The extended MLIT dataset

The present study leveraged the spatial database compiled by the Japanese Ministry of Land, Infrastructure and Transportation (MLIT [6]) for about 250.000 buildings affected by the 2011 Tōhoku tsunami. In its original form, this dataset consisted of a polygon shapefile containing building scale information on observed damage ((Damage) distinguished in 7 classes: from "no damage" (class 1) to "washed away" (class 7)) and on other associated hazard and vulnerability factors, as follows: inundation depth at the building location (h) and other possible concurrent hazard factors (OthHaz) (e.g., earthquake, fire, liquefaction), structural type (BS) (i.e., construction material) and number of floors (NF) of the building, its intended use (Use), and year of construction (Year) [8-11]. Di Bacco et al. [7] enriched this dataset by adding other variables, calculated by means of a geospatial analysis, to account for other possible damage explicative factors, which included:

- Building geometry, in terms of extension and shape: footprint area (FA) and compactness of building shape (DegComp);
- Coast-related parameters, indicating the distance between the building and the shoreline (Distance), the direction of tsunami wave attack (WDir) and a binary factor (CoastType) for distinguishing between plain and ria types of coast over the impacted region;
- Synthetic parameters describing the shielding (ShArea, ShVol, SW) and debris generation (DIArea, DIVol) mechanisms of potentially interacting elements (i.e., buildings and/or seawalls) within defined buffer areas created around each building. Di Bacco et al. [7], while recognizing the importance of including such factors in tsunami damage predictions, demonstrated that the choice between the two proposed buffer geometries (denoted as "large" (L) and "narrow" (N) in the original study) implied only slight differences in the prediction accuracy. Moreover, proposed areal and volumetric parameters were characterized by a high cross-correlation and therefore, for the sake of conciseness, in the present study, we considered only the volumetric ones calculated on "large" buffers (i.e., LShVol, LDIVol and LSW).

2.2 Data handling: spatial clustering and implemented machine-learning algorithm

In this study, the subsets of the extended MLIT database were generated by selecting eight of the most impacted cities (with more than 8.000 damaged buildings in each city) or by dividing the whole dataset into 3, 5, 7, and 9 clusters, identified by using K-means spatial clustering [12] based only on geographical coordinates of the data points (i.e., geographical clustering). Figure 1 shows the identified regions and cities considered in the analysis, with indication (in brackets) of the cardinality of each sub-dataset. Within each clusterization scheme (i.e., maximum number of considered clusters), the different sub-regions are numbered with decreasing cardinality.

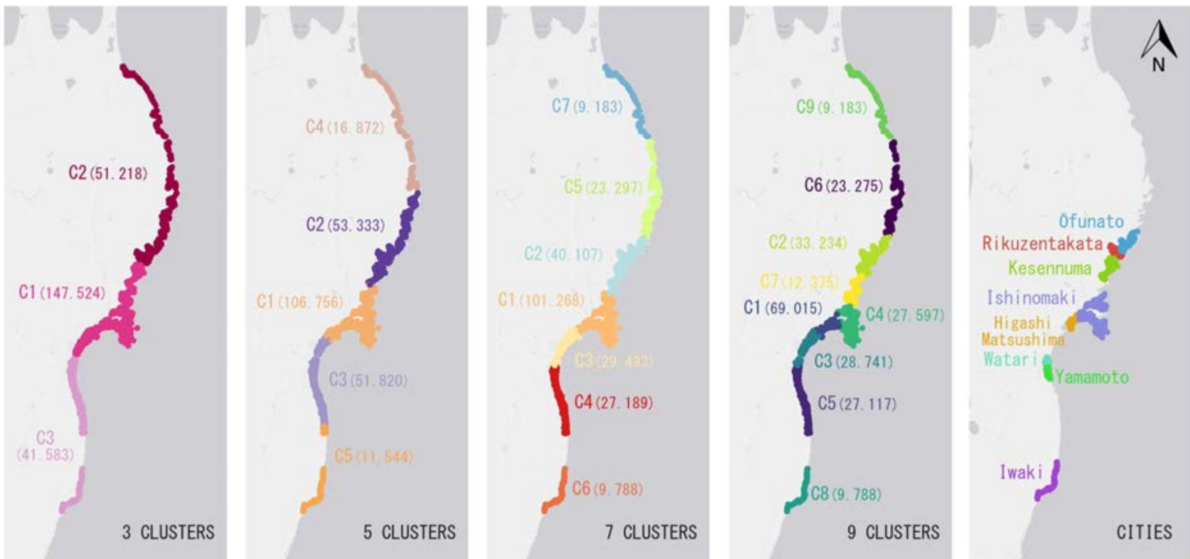

Figure 1: Identified clusters (the cardinality of the subsets is indicated in brackets) and selected cities over the impacted Tōhoku region.

For each sub-cluster, a tree-based ensemble model was set up to analyze the predictive accuracy and variable importance on tsunami damage across the different spatial domains. Among existing data-driven methods, Extremely Randomized Trees (or "Extra-Trees") [13] were chosen in this study because on the full extended MLIT dataset, they showed, as demonstrated in Di Bacco et al. [7], better performances compared to multi-layer perceptron networks, random forest and extreme gradient boosting algorithms. The rationale behind the Extra-Trees algorithm is to create a number of unpruned decision trees by randomizing attribute and splitting value selection, while a majority vote is used to combine the predictions of all trees to determine the final prediction, thus ensuring an excellent variance reduction effect.

Similarly to [7], the number of trees was here fixed at 130 and the splitting process was carried out by evaluating the drop in Gini impurity on random subsets containing 60% of the training items. The entries for each cluster dataset (Figure 1), with Damage as target output and all variables described in Section 1 as input features, were split into training and test sets in the proportion 95-5. The relative hit rate (HR), or the proportion of correct predictions over the total, was chosen as the metric for measuring the accuracy of the model; this metric was combined with the calculation of a normalized confusion matrix in order to have a deeper understanding of the error patterns.

Moreover, a permutation feature importance analysis was carried out to gain a deeper understanding on the possible variability of the feature importance on damage predictions across the different spatial clusters. This procedure consisted in assessing the mean decrease in model's accuracy when a single feature value is randomly shuffled [14], thus providing an indication of the drop in the predictive ability of the model when excluding each variable.

3 RESULTS AND DISCUSSION

3.1 Model's accuracy at different levels of spatial aggregation

Figure 2 reports the overall predictive accuracies (in terms of relative hit rate) obtained with the locally trained Extra-Trees algorithms for the different spatial clusters and cities displayed in Figure 1. The results indicate a high variance in the accuracy for the models trained on the different subsets of each selection, with hit rates ranging from 0.68 to 0.89 and exhibiting a positive correlation with the cardinality of the sets. Figure 2 also indicates that the subdivision of the dataset into a smaller number of clusters (i.e., larger size of the sub-datasets) tends to reduce the differences in the accuracies observed among the identified regions within each sub-group (the difference between maximum and minimum relative hit rate gradually increases from 0.06, 0.17, 0.18 and 0.20 when moving from a 3- to 9-clusterization).

Furthermore, a geographical pattern can be recognized for the models providing the worst relative hit rates (<0.75): in particular, these zones appear to be located at the northern (5.C4, 7.C7 and 9.C9) and southern (5.C5, 7.C6, 9.C8, with the last two almost overlapping with the Iwaki subset) extremes of the impacted region (Figures 1 and 2). The small differences (~1%) in the accuracy obtained for the couples 7.C7-9.C9 and 7.C6-9.C8, which contain exactly the same data, can be explained by the intrinsic randomization of the training process and shuffling of the dataset at the base of the implemented algorithm.

The abovementioned results are confirmed by the normalized confusion matrices represented in Figure 3, indicating the relative hit and misclassification rates among the different damage classes, obtained with the Extra-Trees models trained on the different clusters. For the sake of completeness, Figure 3 also shows, as a comparison means, the confusion matrix (displayed in larger size in the legend box) obtained by Di Bacco et al. [7] with the same model trained with the whole extended MLIT dataset (with an associated overall hit rate of 0.84).

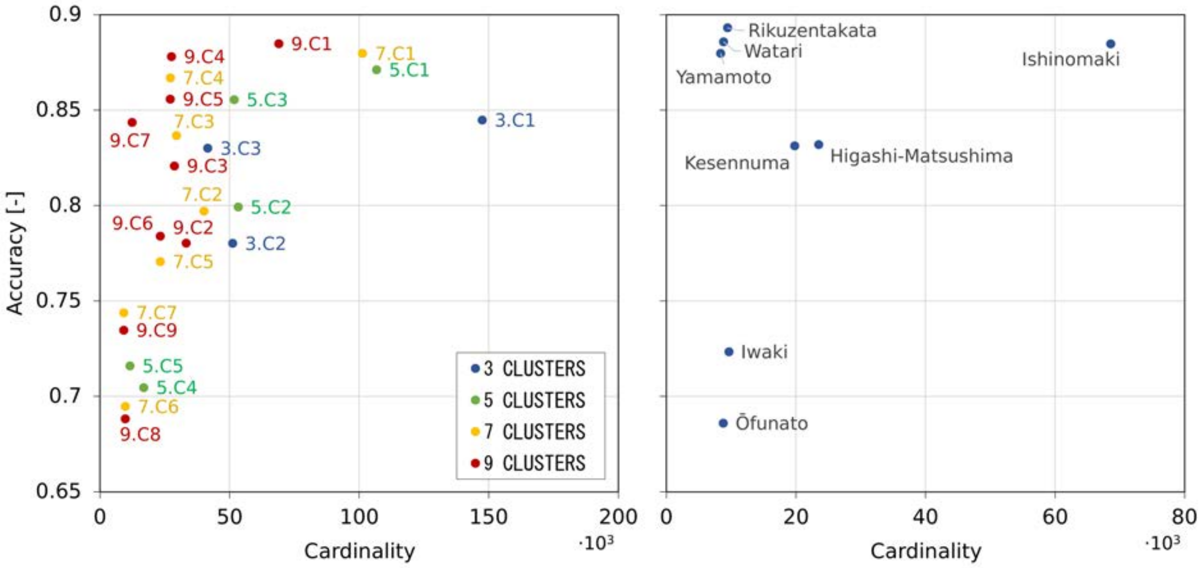

Figure 2: Models' accuracies (i.e., relative hit rate) for the models trained on the different clusters (Figure 1).

The matrices (ordered top-down with decreasing cardinality), while still revealing a dependency of models' performance on the size of the sub-sets, seem again to suggest the existence of local specificities that may significantly affect the predictive accuracy.

These local peculiarities may originate from uneven distributions of the damage classes within the sub-sets and/or from particular local relationships between damage and explicative factors that can make the predictions more complex, thus increasing the probability for misclassification. This clearly appears with a more in depth analysis of Figure 3 in combination with the empirical distributions of damage classes and inundation depth (recognized as the most important damage predictor in [7]) within the sub-sets, reported in Figure 4 for the "Cities" and "9" cluster areas as explicative cases.

For instance, if focusing on the confusion matrices for the discussed sub-sets of Iwaki (C.C4) and the northern 9.C8, it can be seen that most of the prediction error is due to a frequent misclassification of the intermediate damage levels 4 ("major damage"), 5 ("complete damage") and 6 ("collapse"). As clearly visible from the empirical distributions in Figure 4, both of these zones are almost totally covered by shallow water depths, which inherently bring larger uncertainty in damage estimation, because of the significant importance, in such conditions, of local vulnerability and hazard factors that may exacerbate damage variability [15-17]. Watari (C.C6) is also characterized by a large share of low water depths, but the model still provides one of the best accuracies (0.88), with most of the damage levels correctly predicted, if excluding only class 5 (represented in less than 1% of the C.C6 sub-set (Figure 4)), which is always mispredicted as class 6. The reason for these different behaviors can be attributed to other additional input features ("Distance" in the specific case of Watari) that can locally play an important role in reducing uncertainty at the scale of individual clusters.

Particular is the case of the model for Rikuzentakata (C.C5), which exhibits the best overall hit rate (0.89, Figure 2), while the related confusion matrix in Figure 2 depicts a very scattered pattern, with correct guesses almost only for "Washed away" (Class 7) buildings. Such a possible result is, however, well explained by the damage distribution for this city (Figure 4), where buildings in Class 7 practically represent the exclusive reported cases, thus making the weight of the recognized scatter on other classes negligible on overall accuracy.

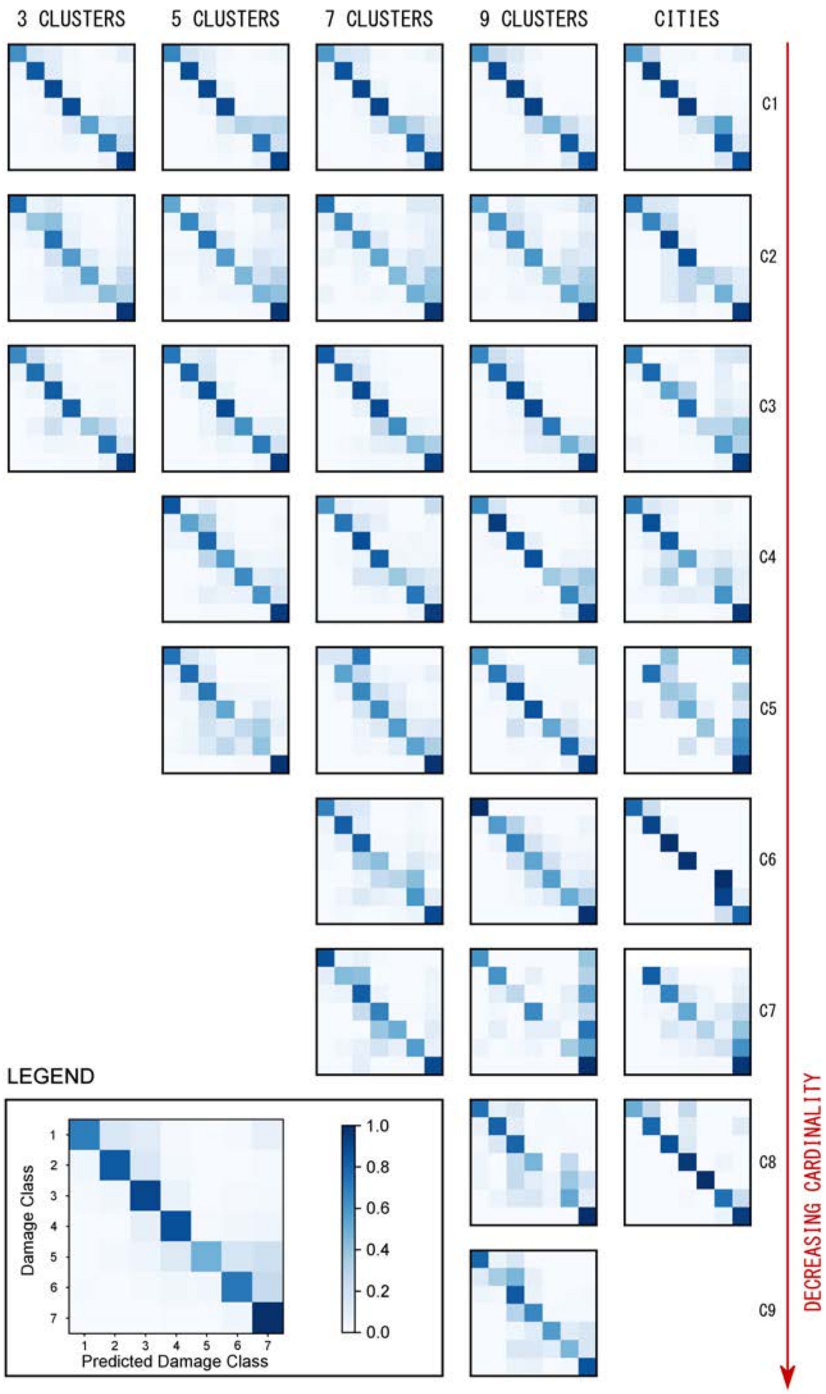

Figure 3: Normalized confusion matrices for the models trained on the different clusters (Figure 1). The matrices related to the "Cities" group are displayed in the following order (top-down): Ishinomaki (C.C1), Higashi-Matsushima (C.C2), Kesennuma (C.C3), Iwaki (C.C4), Rikuzentakata (C.C5), Watari (C.C6), Ōfunato (C.C7) and Yamamoto (C.C8).

In general, also in datasets with high cardinality, the presence of unbalanced shares of data within the different classes is mainly the obvious cause for most of the misclassification error, as evident, for instance, in Ishinomaki (C.C1), 9.C1 and 9.C4 models (Figures 3 and 4), as wells as for the whole extended MLIT-based model [7].

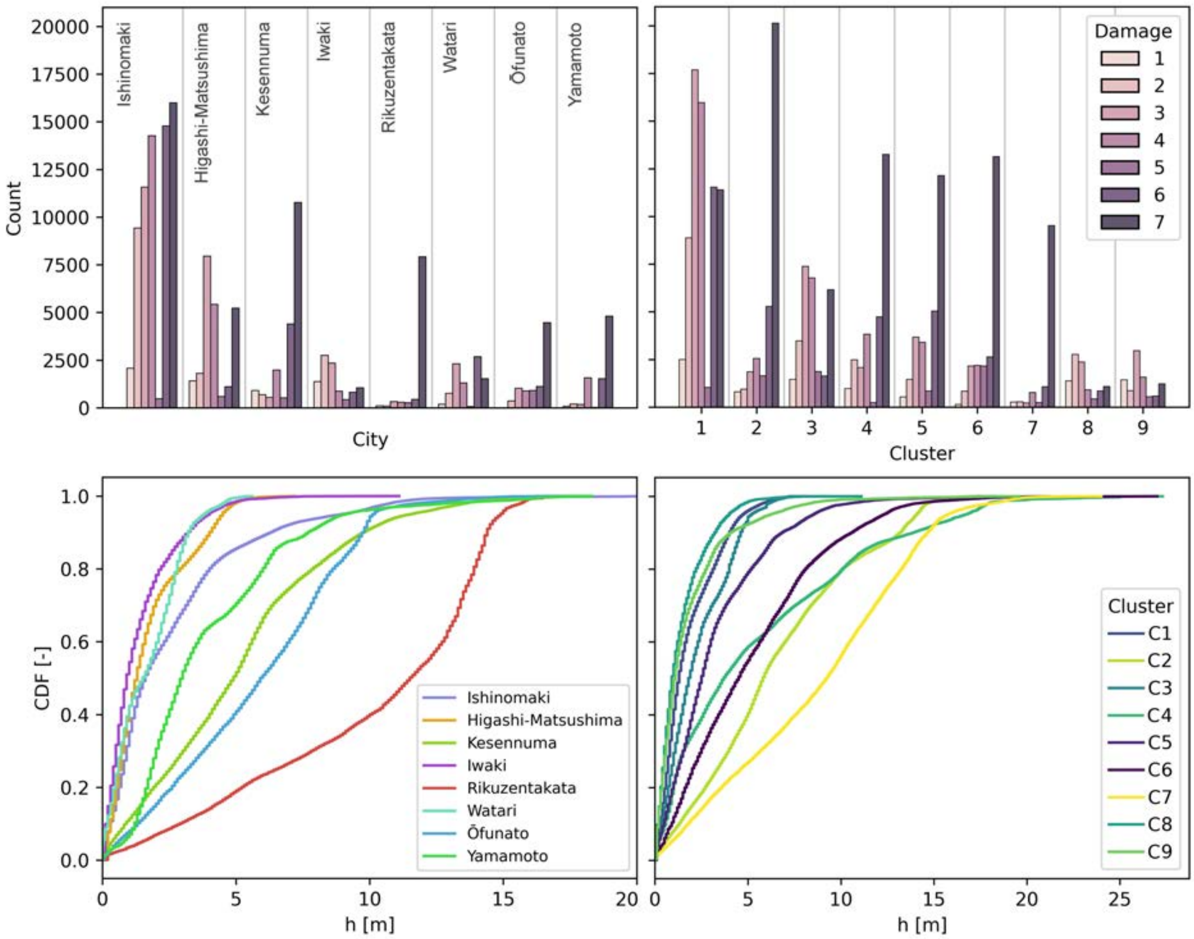

Figure 4: Empirical distributions for Damage (top) and Inundation depth (bottom) data for the "City" (left column) and "9" clusters (right column).

3.2 Regional variability of the feature importance at different levels of spatial aggregation

Figure 5 shows the results of the cluster-averaged feature importance and variability in terms of mean decrease accuracy obtained when the individual input features are randomly shuffled. It appears that the importance of the different predictors is stable for all selections and in good agreement with the results obtained from the models trained on the whole extended MLIT dataset [7].

In detail, out of the obvious inundation depth (highly correlated with the distance from the shoreline, which explains the share of the importance in Figure 5), other parameters recognized as important in tsunami damage modelling, almost irrespective of the clusterization method and geographical location of the impacted area, are the following: the direction of tsunami wave attack (WDir) and the indicators proposed by Di Bacco et al. [7] for accounting for the shielding and debris generation effects created by interacting buildings (LShVol and LDIVol).

Typical predictors considered in tsunami damage modelling, as the structural type (BS) and the number of floors of the building (NF), appear only after the mentioned ones, with a more pronounced local importance variability in smaller cluster datasets (e.g., for clusterization into 7 and 9 areas). Other ancillary building characteristics (related to its geometrical/shape features as well as its intended use and year of construction) have, instead, only a marginal impact on the predictive accuracy and, therefore, they may even be neglected in damage modelling.

Figure 5 also confirms that information on coastal morphology (CoastType) is another key factor for tsunami damage predictions in Japan, because of the amplifying effects occurring in ria coasts [8, 9, 11]. The importance of this variable is only negligible for models developed for cluster areas characterized by the same (or with a large share of the same) coastal type: for example, the large variability in the mean decrease accuracy for CoastType observed in the "5 clusters" plot (Figure 5) is the result of a limited effect of this variable on clusters 5.C2, 5.C3 and 5.C5, as opposed to the more heterogeneous cluster 5.C4.

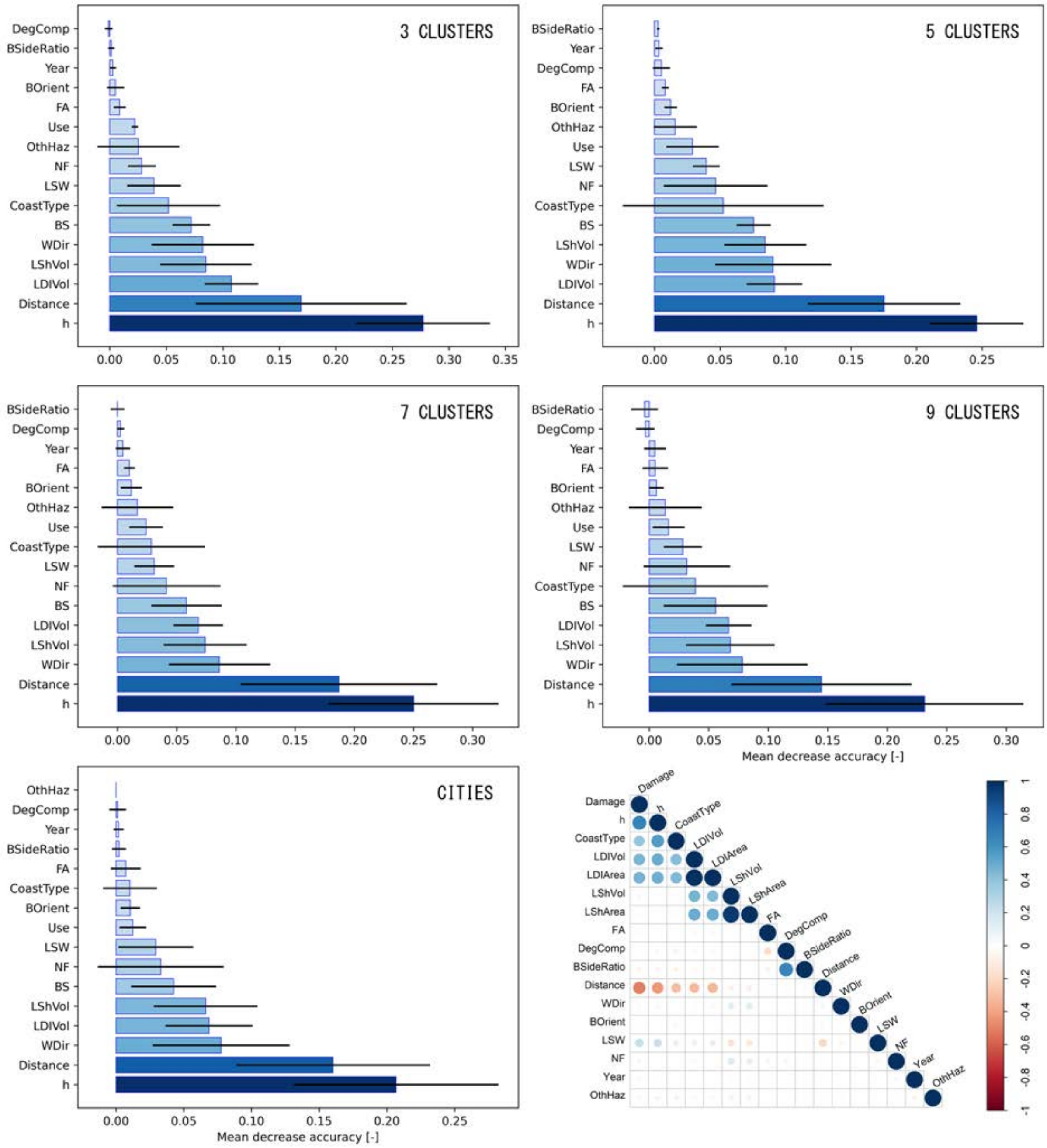

Figure 5: Results of the feature importance analysis for the different clusterization methods. The bars indicates the cluster-averaged mean decrease accuracy and the error bars the corresponding (±) standard deviation. A correlation matrix is also included (bottom right) to highlight the relationships between the variables.

Although with a generally smaller weight, similar considerations apply to OthHaz (the parameter accounting for possible concurrent hazard factors), which only in some specific cases, as in 3.C3 and 7.C4, may become significant for improving damage estimation accuracy. These findings should be interpreted while considering the relatively small portion of buildings characterized by the presence of an additional concurrent hazard factor. Specifically, these buildings accounted for only 2% of the whole MLIT dataset, with the vast majority of them (78%) concentrated within cluster 3.C3. A more in-depth analysis of this subset revealed inland flooding as the most frequent additional hazard factor, which worsened the situation for the buildings located in internal areas and in close proximity to watercourses. The nature of this concurrent hazard factor, which itself contributes to an increase in water depth, also implies that the actual importance of the OthHaz parameter may be partly masked within other features.

4 CONCLUSIONS

By implementing Extra-Trees models based on observed damage data for the 2011 earthquake and tsunami in the Tōhoku region (Japan), this study investigated the effect of data spatial aggregation on damage predictive ability of the models and analyzed the possible existence of regional-dependent features that may affect model accuracy.

The results showed high variance in the accuracy of models trained on data classified in spatial clusters with different cardinality. The worst performances were observed in the northern and southern extreme impacted areas and they could be attributed to the hazard features experienced in such regions, inundated by relatively shallow water depths, which emphasize the importance of local vulnerability and additional hazard factors that increase damage variability and then the probability of damage prediction errors. Such observation also explains the results found by Di Bacco et al. [7] in their analysis for the spatial transferability of the model developed with the whole extended MLIT dataset, which exhibited low accuracies for the cities of Iwaizumi and Hashikami (hit rates in the range ~0.5-0.6), both located in northern clusters, or in the more central, but still affected by shallow waters, Rifu and Matsushima.

Furthermore, the division of the dataset into clusters with a larger sample size was observed to reduce the differences in the accuracies detected among the identified regions within each sub-group. The importance of the different predictors appeared to be stable for all selections and in good agreement with the results obtained for total MLIT dataset, thus allowing the definition of a ranking for the most useful predictors for tsunami damage modelling, which include factors for taking into account the mechanisms of shielding and debris generation created by interacting buildings in inundated areas.

REFERENCES

[1] C. Tarbotton, F. Dall'Osso, D. Dominey-Howes, J. Goff. The use of empirical vulnerability functions to assess the response of buildings to tsunami impact: comparative review and summary of best practice. *Earth-Science Reviews*, **142**, 120-134, 2015.

[2] I. Charvet, A. Suppasri, H. Kimura, D. Sugawara, F. Imamura, A multivariate generalized linear tsunami fragility model for Kesennuma City based on maximum flow depths, velocities and debris impact, with evaluation of predictive accuracy. *Natural Hazards*, **79**(3), 2073-2099, 2015.

[3] I. Charvet, J. Macabuag, T. Rossetto, Estimating tsunami-induced building damage through fragility functions: critical review and research needs. *Frontiers in Built Environment*, **3**, 36, 1-22, 2017.

[4] K. Gibert, J. Izquierdo, M. Sànchez-Marrè, S.H. Hamilton, I. Rodríguez-Roda, G. Holmes, Which method to use? An assessment of data mining methods in Environmental Data Science. *Environmental Modelling & Software*, **110**, 3-27, 2018.

[5] K. Saengtabtim, N. Leelawat, J. Tang, W. Treeranurat, N. Wisittiwong, A. Suppasri, K. Pakoksung, F. Imamura, N. Takahashi, I. Charvet, Predictive analysis of the building damage from the 2011 Great East Japan Tsunami using decision tree classification related algorithms. *IEEE Access*, **9**, 31065-31077, 2021.

[6] Ministry of Land, Infrastructure, and Transportation of Japan (MLIT). Survey of Tsunami Damage condition, http://www.mlit.go.jp/toshi/toshi-hukkou-arkaibu.html, 2012.

[7] M. Di Bacco, P. Rotello, A. Suppasri, A.R. Scorzini, Leveraging data driven approaches for enhanced tsunami damage modelling: insights from the 2011 Great East Japan event. *Environmental Modelling & Software*, **160**, 105604, 2023.

[8] A. Suppasri, E. Mas, I. Charvet, R. Gunasekera, K. Imai, Y. Fukutani, Y. Abe, F. Imamura, Building damage characteristics based on surveyed data and fragility curves of the 2011 Great East Japan tsunami. *Natural Hazards*, **66**(2), 319-341, 2013.

[9] A. Suppasri, I. Charvet, K. Imai, F. Imamura, Fragility curves based on data from the 2011 Tohoku-Oki tsunami in Ishinomaki city, with discussion of parameters influencing building damage. *Earthquake Spectra*, **31**(2), 841-868, 2015.

[10] I. Charvet, A. Suppasri, F. Imamura, Empirical fragility analysis of building damage caused by the 2011 Great East Japan tsunami in Ishinomaki city using ordinal regression, and influence of key geographical features. *Stochastic Environmental Research and Risk Assessment*, **28**(7), 1853-1867, 2014.

[11] N. Leelawat, A. Suppasri, I. Charvet, F. Imamura, Building damage from the 2011 Great East Japan tsunami: quantitative assessment of influential factors. *Natural Hazards*, **73**(2), 449-471, 2014.

[12] D. Steinley, K-means clustering: a half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, **59**(1), 1-34, 2006.

[13] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees. *Machine learning*, **63**(1), 3-42, 2006.

[14] L. Breiman, Random forests. *Machine learning*, **45**, 5-32, 2001.

[15] B. Jongman, H. Kreibich, H. Apel, J.I. Barredo, P.D. Bates, L. Feyen, A. Gericke, J. Neal, J.C.J.H. Aerts, P.J. Ward, Comparative flood damage model assessment: towards a European approach. *Natural Hazards and Earth System Sciences*, **12**(12), 3733-3752, 2012.

[16] M. Amadio, A.R. Scorzini, F. Carisi, A.H. Essenfelder, A. Domeneghetti, J. Mysiak, A. Castellarin, Testing empirical and synthetic flood damage models: the case of Italy. *Natural Hazards and Earth System Sciences*, **19**(3), 661-678, 2019.

[17] A.R. Scorzini, B. Dewals, D. Rodriguez Castro, P. Archambeau, P., D. Molinari, INSYDE-BE: adaptation of the INSYDE model to the Walloon region (Belgium). *Natural Hazards and Earth System Sciences*, **22**(5), 1743-1761, 2022.