CONSTRUCTING EARTHQUAKE DISASTER-EXPOSURE LIKELIHOOD INDEX USING SHAPLEY-VALUE REGRESSION APPROACH[*]

Rahma Anisa[1‡], Bagus Sartono[2], Pika Silvianti[3] and Aam Alamudi[4]

[1]Department of Statistics, IPB University, Indonesia, rahmaanisa@apps.ipb.ac.id
[2]Department of Statistics, IPB University, Indonesia, bagusco@ apps.ipb.ac.id
[3]Department of Statistics, IPB University, Indonesia, pikasilvianti@gmail.com
[4]Department of Statistics, IPB University, Indonesia, aamalamudi@gmail.com
[x‡]corresponding author

Abstract

Indonesia is very prone to earthquake disaster because it is located in the Pacific ring of fire. Therefore, a reference level of earthquake disaster exposure likelihood events in Indonesia is needed in order to increase people's awareness about the risks. This study aims to determine the index that describes the risk of possible future earthquake disaster. As initial research, this study is focus on earthquake disasters in Java region, as it has the largest population in Indonesia. Several indicators that are related to the severity of earthquake disaster impact, were used in this study. The weights of each indicators were determined by considering its shapley-value, thus all indicators gave equal contribution to the proposed index. The results showed that shapley-value approach can be utilized to construct index with equal contribution of each indicators. In general, the resulted index had similar pattern with the number of damaged houses in each districts.

Keywords: earthquake index, java, shapley-value.

1. Introduction

Indonesia is an archipelago country with large potential of natural resources as well as many kinds of industry. In addition, Indonesia also has large number of population, it is the top fourth country with the highest population in the world (U.S. Census Bureau, 2018). However, because it is located in the ring of fire, this country is very

---

prone to natural disaster (Parwanto and Oyama, 2014). This includes extreme rainfall, drought, landslide, flood, forest fires, typhoon, volcanic eruption, tsunami, and earthquake. National Agency for Disaster Management (BNPB) (2016) explained that during 2005-2015 there were 11,648 number of hydro meteorological disaster incidence, and 3,810 number of geological disaster incidence. Therefore, a reference level of disaster exposure likelihood events in Indonesia is needed, and it will be useful for the government as well as the industry and the people to prevent more losses.

Currently there are several index which have been developed in order to map the risk of disasters, such as the world risk index (Welle and Birkmann, 2015). Other research was conducted as well to obtained similar result (Peduzzi *et al.*, 2009). However, these approaches were developed at country level. It has been known that Indonesia, as the largest archipelago country, has varying condition in each region at lower subpopulation level. Therefore, local disaster risk index is preferred in order to describe the likelihood of disaster exposure in the regions of Indonesia. Furthermore, since Indonesia is located on the meeting point of major tectonic plates, this makes it very vulnerable to earthquake disaster. Therefore, this research will focus on earthquake disaster exposure.

BNPB (2016) calculated the index of earthquake disaster hazard based on ground amplification factor and peak ground acceleration. A different approach will be studied in this research by combining the information of historical earthquake around the observed regions and its geographical condition. As initial research, this study will be limited to Java region only, considering that Java is the center of economy in Indonesia which has high population density compared to other region.

This research aims to construct an index of earthquake disaster-exposure likelihood which consider the historical data and geographic conditions at district level in Java, Indonesia.

2. Existing Disaster Index
2.1 World Risk Index (WRI)

According to Birkmann and Welle (2016), the world risk index is constructed using 28 indicators which included in four components, that is exposure to natural hazards, susceptibility, coping capacities, and adaptive capacities. Those are selected by considering United Nation (UN) Sustainable Development Goals. Indicators which describe the exposure are earthquakes, cyclones, floods, droughts, and sea level rise. At the final stage, they combined the four components in order to obtain the index which is valid for country level. This index can be useful for identifying countries which considered to have extreme condition regarding the risk of disaster and their capacity to deal with it.

2.2 Disaster Risk Index (DRI)

Peduzzi *et al.* (2009) developed disaster risk index at global level as part of United Nations Development Programme (UNDP) program. They calculated physical exposure of typical hazards (tropical cyclones, droughts, earthquakes, and floods) to construct the index. Multiplicative approach of the physical exposure and vulnerability was used to developed the risk index by using parametric modeling for each hazard type. Model calibration also performed for each hazard type in order to validate the calculated physical exposure. This was conducted using the number of estimated killed by the

hazards, which considered as a variable describing the damage severity of the hazards. The index for each country was obtained by aggregated the model for all hazard type. Finally, further analysis to determine high risk countries were performed by exploring the geographical distribution of the resulted index. Peduzzi *et al.* (2009) also mentioned that the index could not describe any information about future disaster risk.

2.3   Indonesian Risk Index

This index is provided by National Agency for Disaster Management (BNPB). BNPB (2016) calculated risk index including three components: hazard, vulnerability, and capacity. Hence, the risk would be able to describe the potential losses in the form of casualties, property damage, and environmental damage. There are 10 types of hazard included in the constructed index: (1) earthquake, (2) tsunami, (3) volcanic eruption, (4) flood, (5) landslide, (6) drought, (7) forest and land fire, (8) extreme weather, (9) extreme wave and abrasion, and (10) flash flood. As part of their study, the risk index has been developed for each type of hazard at province level using the following formula:

$$R_{ij} = \left[H_{ij} \cdot V_{ij} \cdot \left(1 - C_{ij} \right) \right]^{\frac{1}{3}}; \ i = 1,2,\dots,10 \text{ and } j = 1,2,\dots,34$$

for R_{ij} denotes risk, H_{ij} denotes hazard, V_{ij} denotes vulnerability and C_{ij} denotes capacity for the i-th hazard type and j-th province.

Particularly for earthquake, BNPB (2016) determined the hazard (H) based on the analysis of ground motion intensity. This value was produced by combining the value of ground amplification factor and peak ground acceleration (PGA) at the bedrock level.

3. Shapley-Value Regression

According to Mishra (2016a) the basic idea of shapley-value is to retrieve a measures of contribution of each independent variables towards a regression model. This is performed by calculating their influence to the determinant coefficient (R^2) of the model. Complete algorithm to calculate this was explained in Mishra (2016a).

Mishra (2016b) explained that a composite index can be constructed by utilizing shapley-value of each indicators. The idea was to develop an index using indicators with equal contribution. According to Mishra (2016b), this can be done by finding appropriate weights that minimize Euclidean norm of the Shapley-value of each indicators towards the proposed index.

Grömping (2006) described that shapley-value (SV) can be calculated using the following formula:

$$SV(x_i) = \frac{1}{p!} \sum_{S \subseteq \{x_1,\dots,\ x_p\} \setminus \{x_i\}} n(S)! \ (p - n(S) - 1)! \ [R^2(\{x_i\} \cup S_i(r)) - R^2(S_i(r))] \quad (1)$$

for p denotes the number of independent variables, S denotes the set of all independent variables, $S_i(r)$ denotes the set of independent variables before x_i is entered into the model in the order of $r = (r_1,\dots,r_p)$. In other words, $R^2(\{x_i\} \cup S_i(r))$ is the R^2 of the model after x_i is included, while $R^2(S_i(r))$ is the R^2 of the model without x_i.

4. Methodology

4.1   Data Sources

The proposed index will be constructed using the following indicators:
(1)   Historical earthquake data
(2)   Historical data of earthquake impact (property damage)
(3)   Average distance of each districts to the Indo-Australian plate
(4)   Average distance of each districts to nearest mountain

   Three years period data set were used in this research, that is in the period of 2015-2018. The data were downloaded from the following sources (Table 1).

<div align="center">Table 1: Data Sources.</div>

No.	Sources	URL
1	National Agency for Disaster Management (BNPB)	http://bnpb.cloud/dibi/tabel2
2	Statistics Indonesia (BPS)	https://bps.go.id/
3	Geospatial Information Agency (BIG)	http://tanahair.indonesia.go.id/portal-web
4	United States Geological Survey (USGS)	https://earthquake.usgs.gov/earthquakes/

Data pre-processing is required towards all indicators in order to obtain useful information with regard to spatial and distance information of each district.

4.2   Research Procedure

The proposed index of disaster exposure likelihood will be constructed using selected indicators and through the following steps:

(1)   Data Preprocessing
   The original historical earthquake data set in this research were the record of earthquake epicenters, this includes the location, magnitude, and the depth. However, the observed region is the districts in Java region. Therefore, in this research, we calculated the distance of each village in Java to the nearest epicenters of major earthquake in the observed period. Further, we aggregated the data into district level by calculating the average value of each village. Using the same idea, the average distance to the nearest mountain and Indo-Australian plate of each district can be obtained.

(2)   Exploratory Data Analysis
   Data exploration using graph and cross-tabulation were used in this research in order to investigate the association between selected indicators and the historical data of earthquake impact during the observed period.

(3)   Standardization of the indicator value
Standardization of data needs to be done to equalize the measurement scale of data coming from various sources. When there are variables measured on different scales, the variables need to be standardized, prior to the sum or aggregation of the data (Ebisudani and Tokai, 2017). The proposed technique is range-equalization where each indicator (X) will be transformed into intervals 0 to 1. The formula to be used is:

$$S_{ij} = \frac{X_{ij} - \min(X_i)}{\max(X_i) - \min(X_i)}$$

for $i = 1, 2, \dots p,$, $j = 1, 2, \dots. n$, while p denotes the number of indicators, n denotes the number of districts, and the value of $\max(X_i)$ and $\min(X_i)$ were determined in the process of data exploration.

(4)   Determining the weights (w_i) for each indicator
Determination of the weights will be conducted objectively using data driven approach. In this research, we will use Shapley-Value Regression technique (Mishra, 2016b). Shapley-Value is one of the methods to calculate contribution of independent variables towards the observed variable. In this research, calculation of the weights is conducted iteratively until the aforementioned properties of shapley-value is satisfied, using the following procedure:
   i.   define initial weights (α_0), starting with the lowest weights.
   ii.   calculate $Z_j = \frac{\sum_i \alpha_i S_{ij}}{\sum_i \alpha_i}$ using the pre-determined weights for j-th district, $j = 1, \dots, n$.
   iii.   generate $e \sim \text{normal}(0, 0.01)$ with the size that equal to the number of districts
   iv.   calculate $Z_j^{*} = Z_j + e_j$
   v.   regress all indicators towards Z_j^{*}
   vi.   calculate shapley-value of each indicators based on regression model which developed in the previous step
   vii.   repeat the procedure in step (ii) – (v) for all possible value of weights
   viii.   evaluate the weights, use one which has equal shapley-value

(5)   Constructing the proposed index
Our proposed index (Z) will be constructed for j-th district using the following weighted average approximation formula for j-th district:

$$Z_j = \frac{\sum_i \alpha_i S_{ij}}{\sum_i \alpha_i} \qquad (2)$$

where α_i denotes weight of the i-th indicator, and $\sum_i \alpha_i = 1$. This index will be calculated in six months period at district level. By using the calculation of shapley-value in step (4), here we are constructing index with weights such that all indicators will have equal contribution.

(6)   Index Validation

There are several methods which can be used to validate the proposed index. In this research, we will use four methods, they are:

(i)   Correlation between exposure potential index and frequency of actual occurrence. A valid index should have a high correlation.

(ii)   Spatial autocorrelation between regions. The adjacent districts should have great relevance. Moran's Index will be used for this validation process.
  a) According to Bivand *et al.* (2015), Moran's Index (I_M) can be calculated by the following formula:

$$I = \frac{n}{\sum_{k=1}^{n} \sum_{l=1}^{n} w_{kl}} \cdot \frac{\sum_{k=1}^{n} \sum_{l=1}^{n} w_{kl} (Z_k - \bar{Z})(Z_l - \bar{Z})}{\sum_{k=1}^{n} (Z_k - \bar{Z})^2}$$

  for n denotes the number of districts, k and l denotes the order of districts, w_{ij} denotes spatial weight from k-th disctrict to l-th district, and Z denotes the proposed index which obtained using equation (2).
  b) Hypothesis testing of Moran's Index can performed by setting null hypothesis (H_0) as no autocorrelation, and otherwise for the alternative hypothesis. Bivand and Wong (2018) explained that the hypothesis testing can be performed using Z-statistics.

5.   Result and Discussion
5.1   Exploratory Data Analysis

Data exploration using graphs was performed to investigate data distribution as well as association between the selected indicators and the variable which considerably able to describe the impact of earthquake disaster. Figure 1 describes that earthquake tend to occurred in the location that close to Indo-Australian plate. As it get closer to the south of Java (near to the plate), the number of earthquake epicenter was higher. In other words, the historical data shows that distance to Indo-Australian plate has negative association with the likelihood of earthquake incidence.
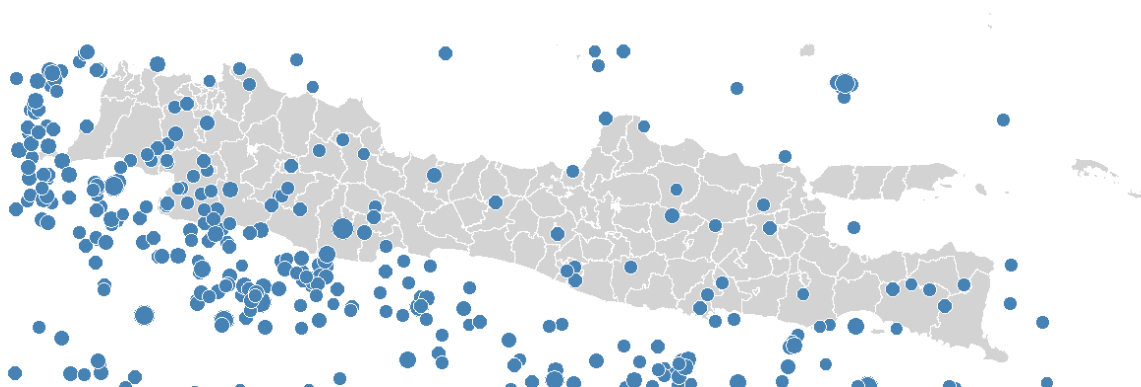

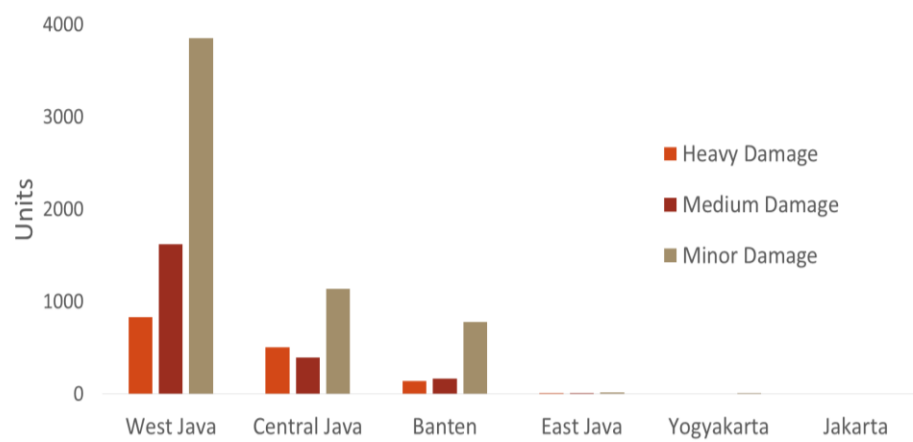Figure 1: Earthquake epicenters around Java region in 2015-2018.

Figure 2: Distribution of the number of damaged houses due to
earthquake disaster in Java (2015-2018).

As described before, in order to determine whether the selected indicators can be used to construct the index, we have to choose a variable that capable to define the damage severity of earthquake incidence. In this study, the number of damaged houses was selected as the reference variable. Figure 2 shows that West Java had the highest number of damaged houses due to earthquake in the last three years, followed by Central Java and Banten.
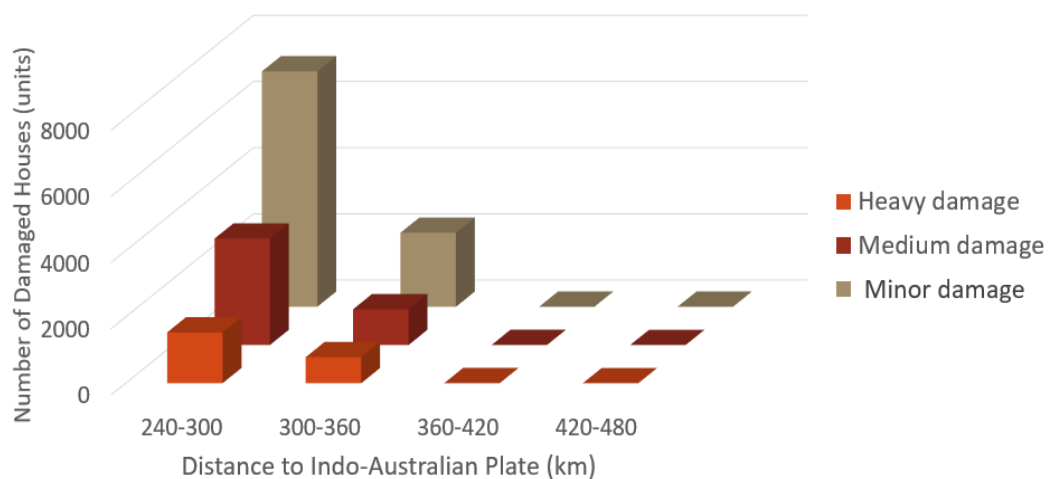
Figure 3: Number of damaged houses at different value of average
distance to Indo-Australian plate

In order to explore the association between distance to Indo-Australian plate and the number of damaged houses, the bar plot of both variables were created. Figure 3 describes that the further the distance, the higher the damage. This confirmed the aforementioned association between both variables.
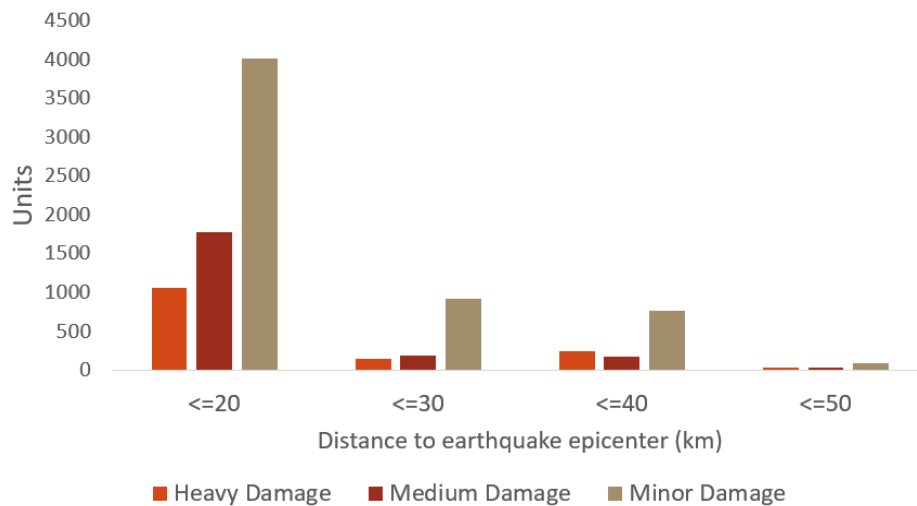
Figure 4: Number of damaged houses at different value of average
distance to the nearest epicenter

Similar to previous exploration, the graph on Figure 4 and Figure 5 show that the distance to earthquake epicenter and nearest mountain also have negative association to the number of damaged houses. Other than that, it may also be affected by the corresponding earthquake magnitude. Without further investigation, it is assured that the larger the magnitude, the damage severity of the earthquake would be higher as well. Therefore, earthquake magnitude will also be added into the index calculation.
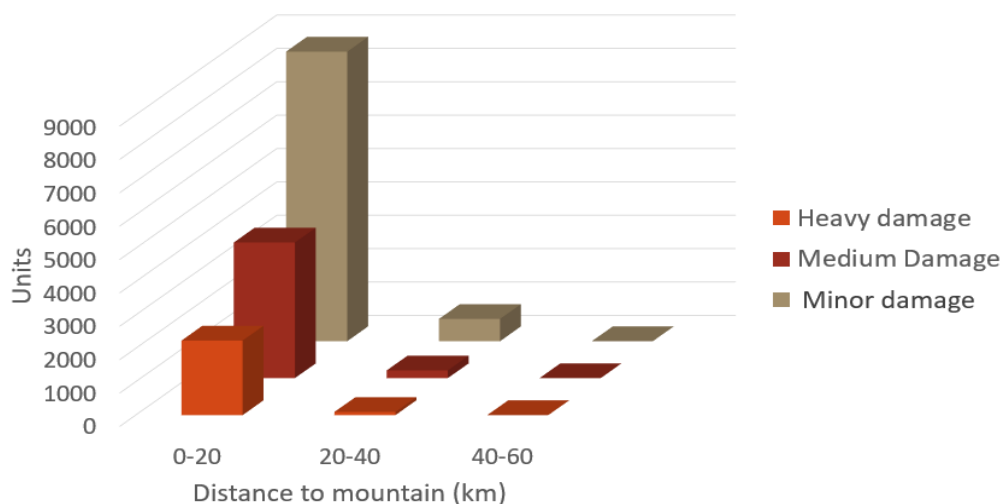

Figure 5: Number of damaged houses at different value of average
distance to the nearest mountain

5.2   Proposed Index

As described in the previous section, the weights were calculated iteratively until certain values which balancing the shapley-value is reached. The following table shows the indicators with corresponding weights and shapley-value.

Table 2: Weights and Shapley-Value of each indicators.

No.	Indicators	Weights	Shapley-Value
1	magnitude·(distance to epicenter)$^{-1}$	0.31	0.33
2	(distance to Indo-Australian plate)$^{-1}$	0.36	0.33
3	(distance to nearest mountain)$^{-1}$	0.33	0.33

Based on the boxplot provided in Figure 6, index of districts in West Java tend to be higher than district in other provinces. It is followed by West Java Province. This is in line with the distribution of the number damaged houses which had already been explained in previous subsection. However, Jakarta province shows different order. This region had no damaged houses in the last three years, but the proposed index were slightly higher than Yogyakarta. This information can be interpreted that this region might has a relatively higher chance to be exposed by earthquake disaster compared to Yogyakarta and Banten, although this incident has not yet occurred.
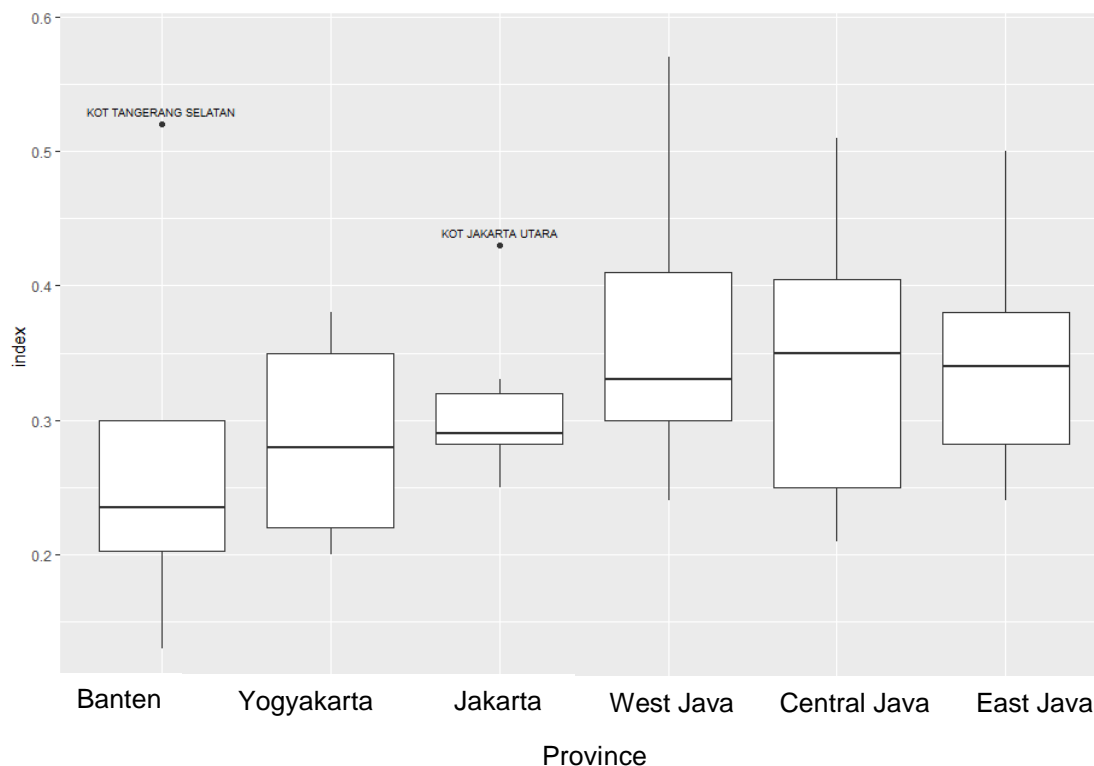

Figure 6: Distribution of the proposed earthquake disaster index

The proposed index can be utilized to map the likelihood of earthquake disaster at district level in Java. Figure 7 suggests that there might be a spatial autocorrelation among districts in Java region. This is indicated by the fact that adjacent districts tend to have similar colors.
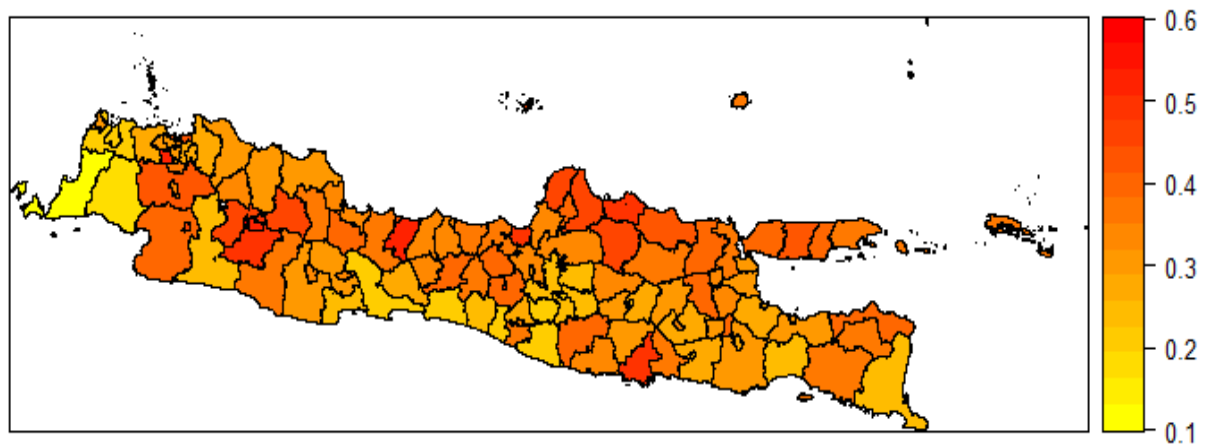
Figure 7: Map of the proposed index

5.3   Validation

To validate the value of proposed index, correlation between the number of total damaged houses and the index was calculated. If the zero value were omitted, the correlation between both variables was 0.094. More specifically, Figure 8 showed the pattern of its correlation. It seems that the index could not yet described all the variance of the total number of damaged houses data. However, if the calculation was performed separately for each damage categories, the correlations became relatively higher (Table 3). In addition, Figure 9 informed that despite of its low correlation, the index still capable of following the pattern of the number of damaged houses, as the reference variable for impact of earthquake disaster incidence.
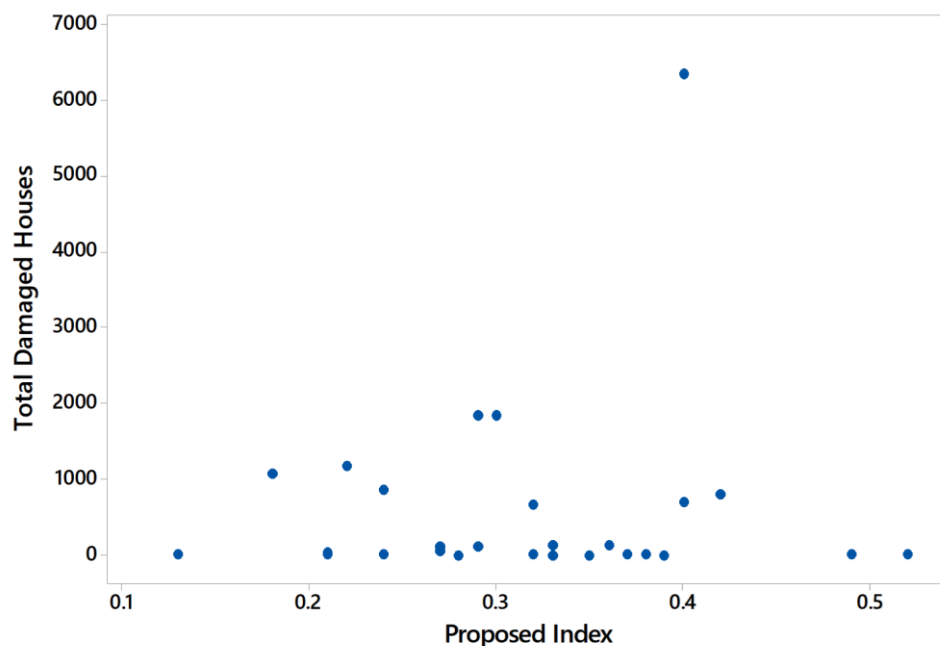

Figure 8: Scatterplot of total (non-zero) damaged houses versus the resulted index

Table 3: Correlation between index and non-zero damaged houses

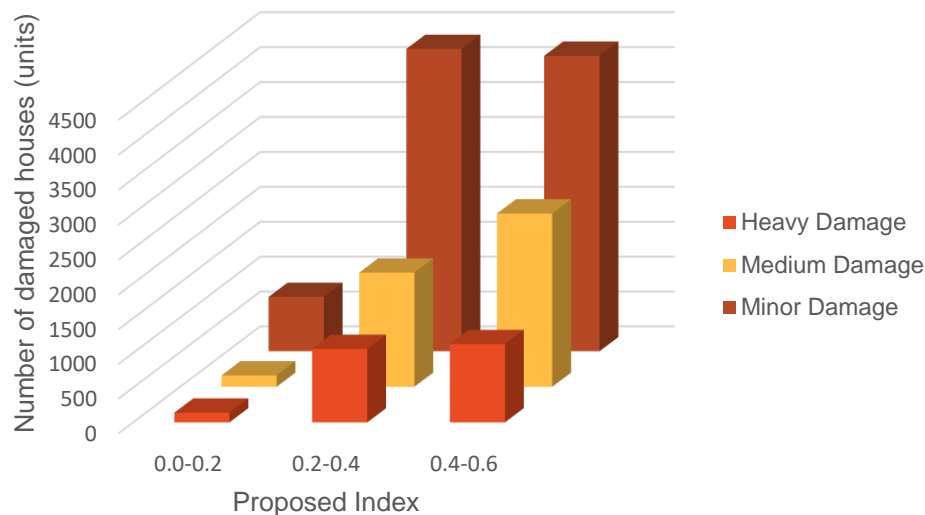| Damage Category	Correlation with the index
Heavy damage	0.074
Medium damage	0.309
Minor damage	0.122

Figure 9: Number of damaged houses at different value of proposed index

Spatial autocorrelation of the resulted index was explored as well. In this study, spatial weights was determined based on k-nearest neighbors approach. Syed (2014) explained that KNN method tend to perform better at larger value of k. By considering previous finding by Syed (2014), we used k=5. The resulted Moran's Index was 0.3482, with p-value less than 5%. Hence, the null hypothesis of no spatial autocorrelation could be rejected at significance level 0.05. This means that there was spatial autocorrelation of the proposed earthquake likelihood index value between districts (Table 4).

Table 4: Test for Spatial Autocorrelation

Moran's I Statistics	Expectation	P-value
0.3482	-0.0008	2.34×10^{-11}

Moran scatterplot in Figure 9 describes positive spatial autocorrelation between districts in Java region. This is in line with the investigation using map in Figure 7. Furthermore, Moran scatterplot also shows several districts which had high influence, they are: Kabupaten Cianjur, Kabupaten Garut, Kabupaten Purwakarta, Kota Cimahi, Kabupaten Wonogiri, Kabupaten Bantul, Kabupaten Pandeglang, Kota Cilegon, and Kota Tangerang Selatan.
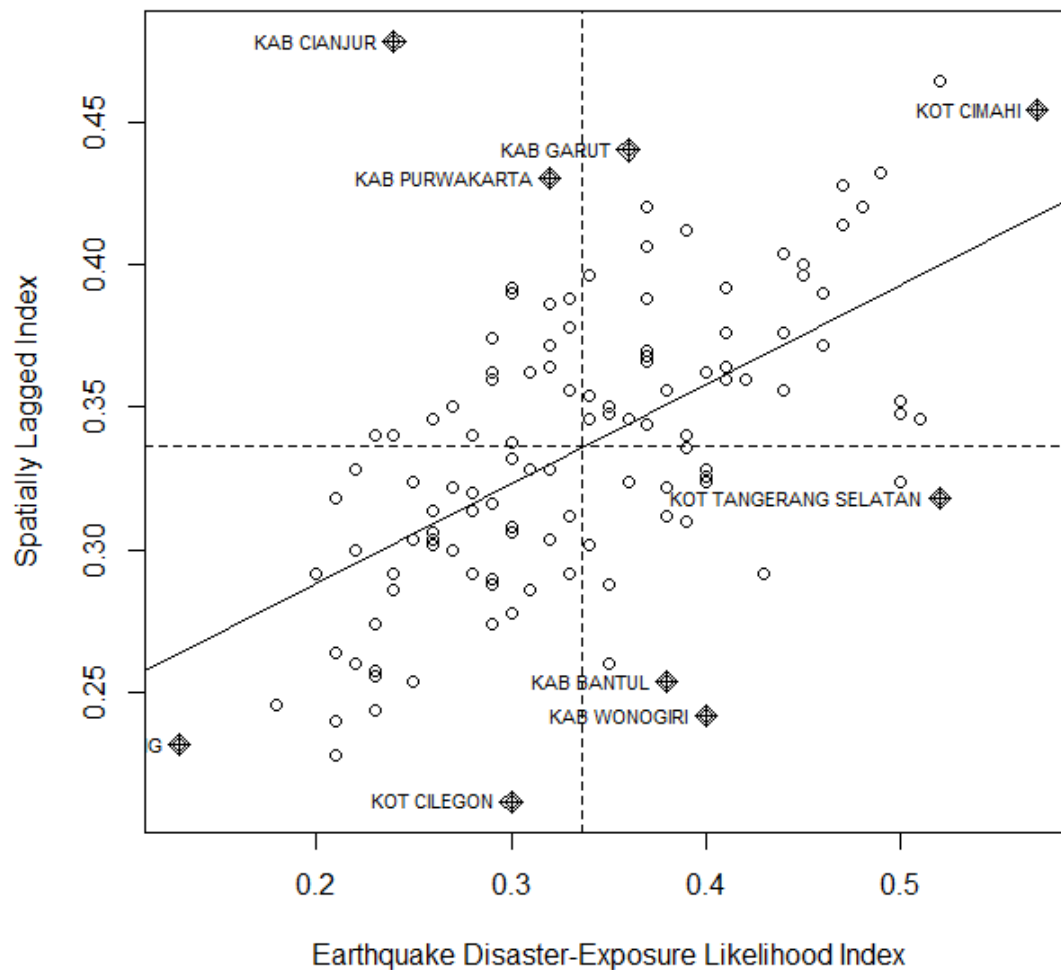
Figure 10: Moran scatterplot for the proposed index

6. Conclusion

The shapley-value can be utilized to construct index with equal contribution of each indicators. The proposed index was calculated using indicators which describe distance to earthquake epicenter, magnitude of the nearest epicenter, distance to Indo-Australian plate, and distance to the nearest mountain. Overall, the proposed index had similar pattern of the number of damaged houses due to earthquake disaster. However, the index could not yet describe the pattern of damages due to earthquake very well. Hence, further development, such as calibration, is required in order to obtain more powerful index to describe the actual earthquake risk.

References

Birkmann, J., & Welle, T. (2016). The WorldRiskIndex 2016: Reveals the necessity for regional cooperation in vulnerability reduction. *Journal of Extreme Events*, *3*(02), 1650005.

Bivand, R., Altman, M., Anselin, L., Assunção, R., Berke, O., Bernat, A., & Blanchet, G. (2015). Package 'spdep'. https://cran.r-project.org/web/packages/spdep/spdep.pdf [December 6th, 2018].

Bivand, R. S., & Wong, D. W. (2018). Comparing implementations of global and local indicators of spatial association. *TEST*, *27*(3), 716-748.

[BNPB] National Agency for Disaster Management. (2016). Indonesian Risk Disaster. BNPB, Jakarta.

Ebisudani, M., & Tokai, A. (2017). The Application of Composite Indicators to Disaster Resilience: A Case Study in Osaka Prefecture, Japan. *Journal of Sustainable Development, 10(1), 81*.

Grömping, U. (2006). Relative importance for linear regression in R: the package relaimpo. *Journal of statistical software, 17(1), 1-27*.

Mishra, S. K. (2016a). Shapley value regression and the resolution of multicollinearity. *Journal of Economics Bibliography, 3(3).* DOI: 10.2139/ssrn.2797224.

Mishra, S. K. (2016b). A note on construction of a composite index by optimization of Shapley value shares of the constituent variables. *Turkish Economic Review, 3(3), 466-472.* DOI: 10.2139/ssrn.2803798.

Parwanto, N. B., & Oyama, T. (2014). A statistical analysis and comparison of historical earthquake and tsunami disasters in Japan and Indonesia. *International Journal of Disaster Risk Reduction, 7, 122-141*.

Peduzzi, P., Dao, H., Herold, C., & Mouton, F. (2009). Assessing global exposure and vulnerability towards natural hazards: the Disaster Risk Index. *Natural Hazards and Earth System Sciences, 9(4), 1149-1159*.

Syed, M. E. (2014). *Attribute weighting in k-nearest neighbor classification*. University of Tampere, Finland. (Master's thesis).

[U.S. Census Bureau]. 2018. Top 10 Most Populous Countries. https://www.census.gov/popclock/world [July 5th, 2018]

Welle, T., & Birkmann, J. (2015). The world risk index–an approach to assess risk and vulnerability on a global scale. *Journal of Extreme Events, 2(01), 1550003*.