Supplement of Nat. Hazards Earth Syst. Sci., 23, 2133–2156, 2023
https://doi.org/10.5194/nhess-23-2133-2023-supplement

Natural Hazards and Earth System Sciences
Open Access

Supplement of

Using machine learning algorithms to identify predictors of social vulnerability in the event of a hazard: Istanbul case study

Oya Kalaycıoğlu et al.

Correspondence to: Oya Kalaycıoğlu (oyakalaycioglu@ibu.edu.tr)

S1 Social vulnerability research and construction of SoVI in 2017

The Aim

This document presents the construction of social vulnerability index (SoVI) score, based on the household survey which is conducted as a component of the “Megacity Indicator System for Disaster Risk Management (MegaIST)” project carried out within the Istanbul Metropolitan Municipality (IMM) in 2017. In this project, it is aimed to determine the disaster-related social vulnerability in Istanbul. In this respect, determining the indicators that will represent the concept of social vulnerability, designing the questionnaires for obtaining these data, conducting the household survey, and carrying out statistical factor analysis to develop a “social vulnerability index” score constitute the main parts this document.

Indicators of social vulnerability

In this study conducted in Istanbul within the scope of MegaIST project, Cutter’s factor analytic framework was taken as a reference for constructing a SoVI score for the households. 7 clusters that are thought to be related to “social vulnerability” were used and 53 indicators related to these clusters were chosen based on the extensive literature reviews. The variable clusters determined to measure social vulnerability and the assumptions of the variables in the clusters in relation to having negative impact on social vulnerability are given in Table 1.

Table S1. Variable clusters and variables used in household survey in MegaIST project

Clusters	Variables and their assumptions in relation to having negative impact on SV
1. Socio-demography • Demography • Education • Occupation • Social security	1. Being an individual under the age of 5 living in the household 2. Individuals over the age of 65 living in the household 3. The state of being disabled, chronically ill and in need of care living in the household 4. Being an illiterate and literate person in the household who has not completed school 5. Individuals living in the household with the status of “unpaid family worker”, “daily / casual / seasonal worker”, “unemployed” and “housewife” 6. Being an individual living in the household without social security
2. Duration of living in the city (İstanbul)	1. If living in Istanbul for less than 5 years 2. Living in the same neighbourhood for less than 5 years 3. Have lived in the same house for less than 5 years 4. If the reason for living in the neighbourhood is “close to my workplace”, “inexpensive housing”, “close to relatives”, “modern and spacious housing”

3. Socio-economy	1. If the ownership of the residence is “rent”, “house belongs to a relative”, “lodging” 2. If the house you live in is a “gate keeper’s lodge” or “squatter house” 3. If the residence is “one roomed flat” 4. If the house is heated with “wood” or “electricity” 5. If there is no “internet”, “dishwasher”, “continuous hot water”, “digital broadcasting platform” in the house where they live 6. If there is no house owned other than the house occupied within the provincial borders of Istanbul 7. If there is no shop/land belonging to the family within the provincial borders of Istanbul 8. If there is no house belonging to the household outside the province of Istanbul 9. If there is no land/shop belonging to the household outside the provincial borders of Istanbul 10. If the household does not have savings for emergencies 11. If one of the members of the household has a debt of 2 thousand TL or more (at the time of the survey min. wage was 1400 TL) 12. If one of the household members does not have a driver's license 13. If a household member does not have a vehicle such as a car/minibus/van, 14. If the household's source of income is "income support from public authorities" 15. If the income from all kinds of sources to the household is below 3000 TL per month
4. Access to health care facilities	1. If there is an individual in the household who needs constant care and treatment 2. If there is no support from the state for the care of this member 3. If there is no easily accessible health centre close to the household's home 4. If health problems are delayed “sometimes” or “often” due to financial insufficiencies
5. Social Solidarity	1. If the members of the household do not participate in various activities such as religious, political, trade union, NGO, fellow citizens' associations 2. If the household members do not have a political party membership 3. If the household members do not have a membership to any NGO 4. If the household does not participate in the decision-making processes of the issues concerning the neighbourhood or street they live in 5. If the household “gets a loan from the bank” and/or “has no one to go to in case of economic hardship” 6. If the language spoken by the household members at home is a language other than Turkish
6. Risk perception and attitudes	1. If two or more of the options from the list of preventable risks are not accepted saying that they are unavoidable 2. If the reasons why natural events such as earthquakes turn into disasters are stated as “because of some individuals”, “comes from God”, “unavoidable” 3. If says that “individuals” or “NGOs” are responsible for preventing possible negative consequences that may arise from an earthquake 4. If no precautions are taken against the possibility of earthquake 5. If said no to the questions about did you “Get training” against the possibility of an earthquake, “make a DASK insurance”, “inspected the interior of the house and determined safe places”, “fixed the furniture”, “know the locations of natural gas, electricity and water valves and how to close them”, “know how to behave during an earthquake”, “know the emergency phone numbers”, “identify and know a specific meeting point in the event of a disaster”, “have a searchlight, fire extinguisher at home”, “know the assembly area after the earthquake” 6. If renovations such as wall and column removal have been made to expand the room in the house where you live. 7. If there is no idea whether the house you live in is safe or not. 8. If it is thought that there will be no earthquake in Istanbul in the near future

	9. If it is thought that the neighbourhood you live in will not be affected by the earthquake 10. If you do not feel prepared for a possible earthquake 11. If the level of knowledge about the earthquake is “knows nothing” and “knows a little” 12. If the preparatory activities of earthquake-related organizations such as “muhtar”, “local NGOs”, “national Associations and foundations”, “municipalities”, “university”, “governorship”, “governorship”, “army”, “AFAD” are not considered sufficient 13. If the information about the earthquake is obtained from “neighbours” or “close acquaintances”
7. Values	1. If the hazards are seen as the responsibility of the citizen such as building a house that is resistant to earthquakes, fires and floods. 2. If the person is not willing to pay more taxes for a cleaner and safer environment 3. If it is believed that everything is destiny 4. If the persons do not think that state institutions should work and prepare to increase earthquake safety

Social Vulnerability Survey

A large-scale household survey was carried out by the by Istanbul Metropolitan Municipality between the period of 2017 and 2018. The target sample size of the research is 50,420 households. But the realized sample is N=41,093 households in 955 sub-districts/neighbourhoods, with residential occupation covering the whole jurisdiction boundaries of the metropolitan municipality of Istanbul, were included in the study of social vulnerability. The households were randomly selected from the Address Based Population Registration System Database of the Turkish Statistical Institute using the proportionate stratified sampling method. All 955 neighbourhoods within 39 districts of Istanbul were taken as strata, and then households were randomly selected from each neighbourhood. The number of households in each neighbourhood taken is proportional to the neighbourhood population. The study covered all 955 sub-districts with residential occupation and as a result, social vulnerability score for each household, sub-district and district is calculated.

The research is composed of three stages. First stage includes the literature review, determination of relevant indicators to evaluate social vulnerability level, production of household survey to provide data to indicators and validation of household survey through pilot survey. In line with these indicators, survey is designed and carried out in pilot level to validate the efficiency and applicability of the surveys. In second step, validated household survey is applied in city scale. The survey data was obtained via face-to-face interviews with one household member, who is between 18 and 70 years of age and is able to give relevant and accurate information about the household. Immigrant houses are within the scope of survey, and student houses and the houses of domestic and foreign individuals who have the characteristics of student houses are not included in the scope. The survey included questions related to socio-demography, socio-economy, duration lived in an urban environment, access to health services, social solidarity, risk perception, actions taken to reduce risk and cultural beliefs. In third step, the survey data were statistically analyzed and factor analysis was used to select indicators in order to test the representation quality of the indicator themes.

According to the findings of the analysis, it is found that only the indicators related to socio-demography, socio-economy, risk perception/actions and values are validated statistically for analysis.

Data Collection

The database system "mysql database" and "nginx + push engine" has been put on the server. Access to the site is provided in a user-specific and secure manner by providing a user name and password for each user. The "Html5 Geolocation javascript high accuracy" method was used to obtain the coordinates and the standard deviation was determined as ±200 m. These coordinate data were then processed and matched with the interviewed household addresses. In addition, the data provided is both raw. It was prepared with .mbd and .mxd extensions as household and neighborhood based, using the GIS software ArcINFO 10.1, processed as requested by DEZİM.

Basic demographic, health and social security information of all the household members living in the household were asked to the suitable respondent from each household accepting to be interviewed.

Descriptive findings

The profile of the households in the survey whose social vulnerability was measured in terms of earthquake-induced disasters are summarized as follows:

- Average age 34.94
- Average household size is 3.4 people
- Average years of education 8.8 years
- Most of them are either retired from a paid job or working in paid employment
- Well established resident in Istanbul in terms of duration of living in the city
- They are able to access health services
- High dependent population (children and elderly)
- Disabled dependent member is low
- More than half owns the house they live in
- Low ownership outside of Istanbul
- More than half of them have a monthly income of less than 3000 TL
- Debt rates are low
- High accumulation capacity
- More than half single-income households

- Almost non-existent civil society participation.
- 1 out of 10 people is a member of a political party.
- There is high solidarity among family members.
- It exhibits a pattern of solidarity that is dependent on bank loans in coping with economic difficulties outside the family, where the solidarity of relatives and acquaintances is low.
- Neighbor recognition and strong social relations
- Low risk perception
- Low confidence for earthquake preparedness of various government Institutions and NGOs
- Media is the most important source of information.
- They consider state institutions primarily responsible for taking precautions against earthquakes.
- The tendency to externalize risk is high.
- It can be said that their desire to be a stakeholder for a safer environment is weak.

Construction of SoVI with principal component analysis

53 variables presented in Table 1 were used to construct SoVI for Istanbul. Since variables did not have a homogenous unit of measurement to make them comparable, they were normalised, transformed to achieve normality when necessary, and then standardised to obtain a mean of 0 and a deviation of 1. Factor analysis, particularly principal component analysis (PCA) was applied by using the SPSS version 25.0. PCA is commonly used to reduce the number of variables to a smaller number of components, with the first component explaining the most variance. By grouping the variables that are most correlated, PCA creates components, known as principal components. No rotation of the data was performed. Only components with Eigenvalue >1 were retained and Kaiser-Meyer- Olkin (KMO) of sampling adequacy was applied. As a result of the PCA, 4 components were extracted that satisfies eigenvalue >1. Variables with high factor loadings (>0.30) within each cluster were kept for the further analysis. In Table 2, variable clusters that can be grouped into these components and their factor loads are given.

Table S2. Components extracted with PCA and their factor loads.

Variable clusters	Factor loading
Demographic variables	0.67
Socio-economic variables	0.75
Risk perception and attitudes	0.35
Cultural norms and values	0.42

The variables in each cluster were summed up based on their sign (+ or -) to obtain cluster scores. Then, cluster scores were converted into standardized Z-scores. All standardized cluster scores were summed up to obtain total scores. In the final stage, the social vulnerability index scores of households were obtained by converting the total scores to a T score with a mean of 50 and a standard deviation (SD) of 10. The distribution graph of the SoVI scores presented in the Fig. 1 shows that SoVI scores have a symmetric distribution which is close to Normal.

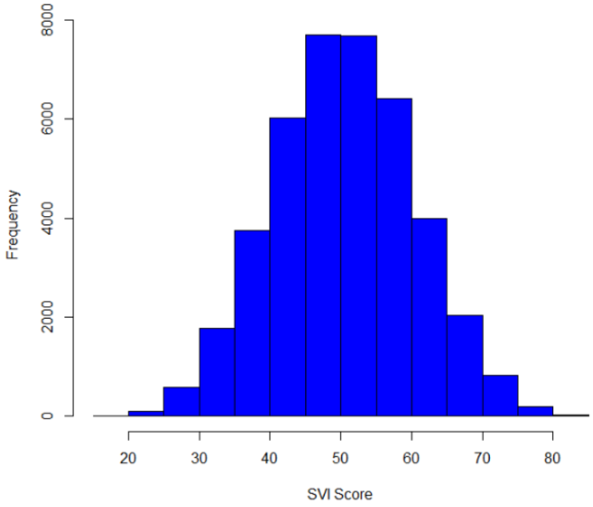

Figure S1. Histogram of SoVI Scores

At the final stage, SoVI scores were classified into four categories based on Z score: Low vulnerability (< -1 SD), Low-moderate Vulnerability (-1 to 0 SD), High-moderate vulnerability (0 to 1 SD), and High Vulnerability (> 1 SD). From Table 3, it can be seen that the prevalence of those who are moderately vulnerable is the most, which corresponds to 67.7%. The households who are the most socially vulnerable corresponds to 17.2% of the households.

Table S3. Distribution of households according to their SV status.

Social vulnerability status	**Number of households**	**%**
Low SoVI	6,207	15.1
Low-moderate SoVI	13,723	33.4
High-moderate SoVI	14,111	34.3
High SoVI	7,052	17.2
Sum	**41,093**	**100.0**

S2 Possible data sources

Table S4. Data sources from which the predictors of ML models can be available

Themes	**Variable**	**The sources that data could be obtained upon permission**
Socio-Demographic	Household size	Address Based Population Registration System
	Age of HH members	Address Based Population Registration System
	Number of women in the HH	Address Based Population Registration System
	Number of men in the HH	Address Based Population Registration System
	Number of <5 year olds in the HH	Address Based Population Registration System
	Number of >65 years of age in the HH	Address Based Population Registration System
	Education of HH members	Address Based Population Registration System
	Social security of HH members	Social Security Institution
Health	Individuals with health insurance	Social Security Institution
	Number of HH members with disability	Ministry of Health, Ministry of Family and Social Services
	Hospital/Health center within close proximity to the house of residence	Municipalities, Ministry of Health
Socio-Economic	Income earners in the HH	Social Security Institution
	Regular salary income earners in the HH	Social Security Institution
	Pension income earners in the HH	Social Security Institution
	Any income earner from property rent in the HH	National Taxation Department
	Households getting income support from public authorities	Social Security Institution, Ministry of Family and Social Services, Municipalities
	HH members with job insecurity	Municipalities/The Ministry of Environment and Urbanization/Social Security Institution
	Owning house of residence	Land Registry and Cadastre
	Type of the house of residence	Municipalities/The Ministry of Environment and Urbanization
	Natural gas heating in the house of residence	Municipalities
	Own a house other than house of residence in Istanbul	Land Registry and Cadastre
	Own land in Istanbul	Land Registry and Cadastre
	Own house out of Istanbul	Land Registry and Cadastre
	Own land out of Istanbul	Land Registry and Cadastre
	Have a registered saving	National Taxation Department / Banks
	Having a debt to banks	National Taxation Department / Banks

S3 Machine Learning Methods

The machine-learning community divides learning problems into various categories: the two most relevant to statistics are those of supervised learning and unsupervised learning. Supervised learning adopts an algorithm to learn the mapping function from the input variable(s) to the output variable(s), whereas unsupervised learning uses for problems in which there is no knowledge to define a suitable output variable. In this research, we defined the output variable as a binary class which corresponds to high social vulnerability, and we aimed to find the best performing classification algorithm for predicting the social vulnerability of the households using input variables. Therefore, we used supervised machine learning algorithms which suit well with the aim of our study.

Logistic regression technique and six supervised machine learning algorithms were compared in our study in terms of their predictive performances. We note that these algorithms have different tuning parameters. For different tuning parameter alternatives, the choice of the optimal tuning parameter was determined by the largest area under the curve (AUC) value of the receiver operating characteristic (ROC) curve using the automated grid search available in R **caret** package (Kuhn, 2008).

The methods that we used for the study are explained below:

1. Logistic regression (LR) is a type of statistical regression model where the response variable is binary, which usually represents presence or absence of a condition, or occurrence or non-occurrence of an event. It predicts the probability of occurrence of by fitting data using a logit function (Hosmer et al., 2013). Logistic regression is also involved in the machine learning literature because the predicted model is based on a binary classification which differentiate between the two classes. For the application of LR, we used **glm** method with the **binomial** family option within the **caret** package.

2. Classification and regression trees (CART) is a decision tree algorithm that are widely used for classification and regression learning tasks in several disciplines (Breiman et al., 1984). CART tree algorithms are typically asking binary questions; in other words, if-else statements that can be used for predicting results based on input data. A classification tree is the result of asking an ordered sequence of questions, and the type of question asked at each step in the sequence depends upon the answers to the previous questions of the sequence. The sequence terminates when the class label prediction is made. CART is not only easy to implement and interpret, but also computationally less expensive. On the other hand, the evaluation metric results may not be successful as the other supervised learning algorithms, because it is only based upon a single decision tree task. For the application of CART, **rpart** method is used in the **caret** package.

3. Random Forest (RF) is another supervised learning algorithm that randomly creates and merges multiple decision trees into one forest, which can be also defined as the multiple versions of CART (Breiman, 2001). As the RF algorithm averages of the outcomes from many different CARTs, it has a lower chance of the variance compared to CART. However, it is computationally expensive learning algorithm due to vast number of trees. **caret** package allows users to adopt different forms of random forest functions, and the simple version of random forest function, **rf** is used for this research. The only tuning parameter was **ntree**, which is the number of trees in the forest.

4. Support Vector Machine (SVM), which is developed by Vapnik (1995), distinctly classifies the data points with a hyperplane in a P-dimensional space (P: the number of variables). Essentially, a hyper plane defines a decision boundary and separates the observations in the predictor space into the two classes. The main objective behind defining a such hyperplane is deciding on the trade-off between maximizing the marginal distances for both classes and minimizing the number of misclassified variables (Tharwat and Gabel, 2019). In our study, **svmRadial**, which is the method of support vector machines with radial basis function kernel, is used in the **caret** package, as it provided larger AUC compared to using linear or polynomial kernel functions. In this setting, cost (constant of the regularization term in the Lagrange formulation) and sigma (standard deviation that is used in the kernel function) are the two tuning parameters are adapted with different value selections.

5. Naive Bayes (NB) calculates the probability of a sample unit to be a member of a certain class, based on the Bayes theorem (Friedman et al., 1997). It constructs a Bayesian probabilistic model that assigns a posterior class probability to a sample unit based and uses these probabilities to assign a sample unit to a class. It assumes that the input variables to be used in the model are conditionally independent given the classes. Therefore, if there exists a dependence between the input variables the classification performance may be negatively affected. In this research, we used **nb** method within the **caret** package for training NB. Kernel density estimation is incorporated with the Naive Bayes model; therefore, different tuning parameters were involved. Such tuning parameters are **fL** (Factor for Laplace correction) and **adjust** (Bandwidth Adjustment).

6. K-nearest neighbour (KNN) is the simplest classification algorithm that makes classification by assigning the new sample units to the class most observed among their K nearest neighbours based on a similarity measure (e.g., dissimilarity functions) (Cover and Hart, 1967). For the same dataset, the classification results may differ for different values of *K*. The **knn** method in the **caret** package is used for this algorithm. Only tuning parameter that needs to be considered is the value of *K*. The choice of K is dependent on the balance between overfitting and underfitting (Zahng, 2016). Therefore, we adapted the KNN model with different values of *K* using a range of values.

7. Artificial Neural Network (ANN), which is capable of learning any non-linear function, is created by imitating the functioning of the human brain and transferring it to the computer environment, is first proposed by McCulloch and Pitts (1946). The mechanism operates with artificial neurons that form input and output neurons and a hidden layer(s), which is frequently used for the data set that cannot be separated linearly. When there are multiple layers between the input and output layers, ANN is as a deep neural network (DNN) (Schmidhuber, 2015). Although computationally expensive, it is successful to detect complex non-linear relationships between variables. We used sigmoid/logistic function as the activation function, which is commonly used to add non-linearity to an ANN model. Additionally, the *multi-layer* perceptron choice was employed for ANN model by adapting **nnet** method in **caret**, which contains two tuning parameters: number of neurons in a hidden layer and decay parameter that controls initial weights for input neurons.

References

Breiman, L.: Random forests, Machine Learning, 45: 5–32, 2001.

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J.: Classification and regression trees, Wadsworth, Belmont, 1984.

Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13: 21-27, 1967.

Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian Network Classifiers, Machine Learning, 29: 131–163, 1997.

Hosmer, D. W., Lemeshow, S. W., Sturdivant, R.X.: Applied logistic regression, 3rd ed., Wiley, New Jersey, 2013.

Kuhn, M.: Building Predictive Models in R Using the caret Package, J. Stat. Sotfw., 28, https://doi.org/10.18637/jss.v028.i05, 2008.

Lantz, B.: Machine learning with R 2nd ed., Packt Publishing, Birmingham, 2015.

McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity, Bull. Math. Biophys., 5, 115-133, 1943.

Schmidhuber, J.: Deep learning in neural networks: An overview. Neural networks, 61:85–117, 2015.

Tharwat, A., Gabel T.: Parameters optimization of support vector machines for imbalanced data using social ski driver algorithm, Neural Computing and Applications, 32: 6925-6938, 2019.

Vapnik, V.: The Nature of Statistical Learning Theory, Springer, New York, 1995.

Zhang, Z.: Introduction to machine learning: k-nearest neighbors, Annals of Translational Medicine, 4: 2018, 2016.