Victim clustering with k-prototype algorithm for flood evacuation planning

Jin Wang Chang, *Lay Eng* Teoh*, and *Hooi Ling* Khoo

Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, 43000, Kajang, Selangor, Malaysia

Abstract. Global warming intensifies inevitable severe floods, thus necessitating robust evacuation planning to minimize disaster impacts through swift assistance. Recognizing the interconnectedness of demand and supply aspects, effective evacuation planning considers evacuee behavior through victim clustering, which is of utmost importance. Despite previous efforts in modeling victim behavior, there remains a gap in incorporating victim clustering explicitly in flood evacuation planning. Thus, this study aims to adopt k-prototype algorithm, which is capable of handling mixed-type features, to perform victim clustering for probable flood occurrence by considering numerous influential factors including risk perception, compliance level and arrival pattern of victims. The k-prototype clustering was performed via McClain index (for performance assessment) on an illustrative mixed-type dataset (with 10 variables) comprising 498 valid respondents, for the context of Central Region of Malaysia. The findings show that the optimal number of clusters, which ranges from 2 to 5, could be formed effectively for 8 distinct scenarios. Besides, it is noticeable that the probable response to evacuate (somewhat and very likely) is about 52% under uncertainty. Concisely, this study aspires to furnish emergency planners with beneficial insights in implementing effective evacuation strategies to reduce the negative impact of flood occurrence significantly.

1 Introduction

Globally, the number of disasters that require an evacuation is reaching up to 45-75 annually [1]. In particular, floods pose a serious hazard for many countries (e.g., Malaysia, Philippines, Japan) and result in enormous life, property and economic losses. Major harmful floods that occurred in the past include 2006-2007 Southeast Asian floods [2] and 2010 floods in northern Malaysia [3]. And recently, the collapse of two dams in Libya in September 2023 have caused worrying floods in Derna, causing more than 11000 deaths and 10000 missing people [4]. A lake overflow in India in October 2023 (due to heavy downpours) also caused flash floods in Sikkim, causing more than 14 deaths, 102 missing people, and displaced more than 22000 people [5]. The floods also affected 3000 tourists due to the main highways and bridges being washed away by floods. In early 2023, extended periods of heavy downpours for four days (starting 28 February 2023) in Johor, Malaysia have caused worrying floods

*Corresponding author: teohle@utar.edu.my

that were affecting Johor statewide [6], causing at least 4 deaths [7] and the displacement of more than 50,000 residents [8]. It is predicted that more people would be exposed to the disaster on a global scale, with floods affecting an average of 21 million people per year [9]. Certainly, the rising frequency of flood occurrences and its consequent catastrophic damages emphasize the necessity of a feasible properly-coordinated evacuation system [1, 10-11].

Evacuation, as a risk-reduction strategy, is a complex process [12-13] with two major concerns, i.e., demand and capacity (supply) uncertainties. The aspect of demand refers to the identification of the number of victims that are distributed at different places (spatially) and at different times (temporally) while the aspect of supply focuses on the provision of evacuation vehicles and aid to the victim. Victim (evacuee) behavior is crucial in addressing the demand aspect of the evacuation planning. Victim behavior is defined as the victim's willingness to participate in the evacuation process. The modeling of the victim behavior is not straightforward as individual victim will behave differently based on his/her belief, perceived accuracy of flood information, and other personal concerns. Typically, a full compliance of the victim may not always be possible. Unless it is a mandatory evacuation performed by the law enforcement units, it is very likely to have a certain portion of the population not complying with the evacuation orders. This has made the evacuation planning difficult. Most of the existing studies on victim behavior modeling and analysis concentrate on the influential factors and destination choice for evacuation. Besides, a response curve, that is specifically controlled by a particular probability distribution, is used to model the demand of victims [1]. Apparently, evacuation planning by capturing victim behavior is unfortunately not a simple process because it is usually done with inaccurate or partial information. And, it poses a great challenge to foresee a disaster's timing, location, and severity.

Thus, this paper aims to perform the grouping of victims according to various characteristics in order to determine their likelihood of evacuation in various scenarios. This is done by adopting k-prototype algorithm to perform victim clustering for probable flood occurrence by considering numerous influential factors including risk perception, compliance level and arrival pattern of victims. This study is significant as it allows the quantification of probable demand level of victims under uncertainty, thus enabling the evacuation fleet planning to be done under stochastic demand.

This paper is organized as follows. Section 2 discusses the existing studies on victim behavior modeling and machine learning approach. Section 3 presents the methodology used in this paper, from data collection to victim clustering with k-prototype algorithm. Section 4 shows the case study area and describes on how the survey was designed and conducted. Section 5 discusses the descriptive analysis of the dataset and the clustering result. Section 6 concludes this paper and highlights the limitations and suggestions for future works.

2 Literature review

This section reviewed the relevant recent studies for two major aspects, namely victim behavior modeling and machine learning approach.

2.1 Victim behavior modeling

The understanding of evacuation behavior is crucial for disaster management [1, 14] and thus the analysis of influential factors for evacuation decisions turns out to be a key input for better evacuation planning [15]. With regard to the response of an individual to disaster, several influential factors (including disaster awareness, risk perception, evacuation compliance and involvement of family/community) were found to be significant. Schlef et al. [16] and Mahdavian et al. [17] stated that there is a strong relationship between risk perception and

evacuation decision. Besides, disaster warning dissemination was found to affect the victims to some extent [9, 18-19]. However, Liu et al. [20] revealed that improving a local warning system will only be effective if the people at risk can be evacuated to safe shelters. Thus, victim behavior analysis is needed for successful evacuation.

In addition, victim demographic (e.g., gender, age, income and education level) emerges as the contributing factor for evacuation [15, 19, 21]. According to the past studies, age, gender, highest education level and household income were identified as major influential factors affecting the victim behavior [1, 14-15, 21-26]. House ownership, household size and occupation were also identified in some of the past studies above as influential factors to the victim behavior. However, vulnerable groups were neglected in past studies as most of the studies did not highlight pregnancy and physical disability as influential factors.

Lim et al. [15] conducted a post flood event face-to-face survey to analyze flood evacuation decision modeling. And, they applied a multinomial logit model for analysis purposes. Yin et al. [21] adopted a multi-stage sampling techniques to seek for respondents to answer a well-structured questionnaire, then applied a multinomial probit model to analyze the evacuation willingness of households. Besides, Yang et al. [14] adopted a stated preference survey and structural equation modeling approach to model evacuation decision behavior and destination choice. However, the heterogeneity of the surveyed respondents was not considered in the modeling. More recently, Zhao et al. [27] used machine learning method, i.e., random forest approach, to model emergency pre-evacuation behavior. They highlighted that both social and environmental factors could affect the probability of responding.

In order to estimate the demand level for every evacuation zone, the total number of households that are expected to evacuate can be determined by multiplying the population of the evacuation zone with the anticipated participation rate. Subsequently, response or mobilization curve can be used to estimate the proportion of the total evacuation demand for the respective time period [1]. Numerous types of response curve (e.g., sigmoid curve, Rayleigh probability distribution function, exponential distribution, Poisson distribution, logistic regression, Weibull distribution) were proposed by the past studies in demand modeling. Besides, the behavioral decisions of individuals for evacuation were found to follow a binary choice model (to evacuate or not) and the travel preference for evacuation was modeled in the form of route choice or hybrid route choice models. And, simulation was adopted to forecast the dynamic travel behavior of victims [28]. Besides, Bayram [1] highlighted that behavioral analysis before or after a disaster would contribute to the demand forecasting of traffic (including the number of vehicles required for evacuation). In addition, Diakakis [29] revealed that behavior forecasting of flood victims is important in particular to reduce the flood mortality and hence critical behavioral analysis is required.

Amideo et al., [19] revealed that the uncertainty of victim demand, arrival time at pick up location and travel times for evacuation, need to be reliably modeled by using probabilistic analysis, statistics methods and social science studies. However, this aspect has not yet been discussed in the literature for flood evacuation planning.

Although the modeling proposed by the afore-mentioned studies is remarkable to a certain extent, there are several limitations as listed below:

- the adoption of probability distribution to simulate the victim behavior led to the creation of circumstances-specific model which are only applicable for the analysed context [1, 24];
- the heterogeneity of victims (including vulnerable people) is not explicitly taken into account as most of the existing studies focus exclusively on evacuation with private vehicle, rarely taking into account victims with reduced mobility and special need groups [14-15, 21];

- the uncertainty on evacuation travel patterns of victim was not explored explicitly (especially on the victim's arrival time at the designated marshal location and the evacuation's departure time which are extremely variable due to individual preference) [12, 19, 30].

2.2 Machine learning approach

Machine learning methods are widely used in data analysis and grouping of population for various purposes such as flood evacuation planning, disease-related research, school quality analysis and so on. Lin et al. [31] applied ISO-Maximum clustering algorithm, which is a combination of ISODATA clustering algorithm and maximum likelihood algorithm, to cluster flash flood risk. Xu et al. [32] proposed an OT-K-means++ algorithm for victim grouping, and discovered that OT-Kmeans++ outperforms K-means and K-means++ in term of reducing time cost of clustering and number of iterations. However, both ISODATA clustering algorithm and OT-K-means++ algorithm only work for numerical variables and cannot cater for categorical variables. Sreejith and Sinimole [33] applied various machine learning methods including Random Forest, Naive Bayes, k-Nearest Neighbour and Support Vector Machines in modeling evacuation preparation time for flood, and discovered that Random Forest performed the best among all machine learning methods adopted in their studies. These machine learning methods are supervised learning in which it can only be used to datasets with a responding variable.

Among all machine learning methods, k-prototype clustering is commonly used for grouping people in various scenarios. Li et al. [34] pioneered the use of k-prototype in disease-related research by adopting k-prototype in grouping of vulnerable populations for malaria. Sulastri et al. [35] applied k-prototype in clustering quality of schools using the student admission data. For transportation related research, Soria et al. [36] adopted k-prototype in clustering ridesourcing trip data to examine the emerging patterns of mobility. As k-prototype clustering can cater for mixed-type dataset, using k-prototype for clustering prevents the dataset from any destroy. Converting categorical data to binary data destroys the original structure of data and results in the meaningless binary features after conversion, and k-prototypes can avoid such destroy [37]. K-prototype clustering also does not require any assumption on data distribution [35], which makes it user-friendly to be applied in various scenarios.

There are various performance indices available to determine the optimal number of clusters in k-prototype as well as validating the performance of k-prototype clustering. Aschenbruck and Szepannek [38] studied and compared 8 performance indices (Cindex, McClain Index, Ptbiserial Index, Gamma Index, Gplus Index, Tau Index, Dunn Index and Silhouette Index). They discovered that Silhouette Index and McClain Index have short computational time, causing them to be used more frequently in various fields. The 8 performance indices above were also adopted by Choi et al. [39] for performance validation of their proposed novel model-based clustering method, ''regClustMD".

McClain Index has been applied in various fields for performance validation. For example, Kumar et al. [40] applied McClain Index and Silhouette Index to determine optimal number of clusters for k-prototype clustering of crop dataset for precision farming. Šulc et al. [41] used Dunn Index, McClain Index and Silhouette Index to evaluate the performance of Gower dissimilarity coefficient. They discovered that McClain Index performs well and is relatively consistent when being applied to different combinations of number of numerical and categorical variables, and performs better when all variables are categorical. Sulastri et al. [35] also adopted McClain Index in determining optimal number of clusters for school clustering using k-prototype. They stated that McClain Index is useful in identifying the diversity within the cluster and diversity between the clusters, thus ensuring the optimal

number of clusters has smallest diversity within the cluster and greatest diversity between the clusters.

Other than the 8 performance indices, there are still other validation methods in examining the performance of k-prototype clustering. Soria et al. [36] used scree plot in determining number of clusters for ridesourcing trip data clustering using k-prototype, while Ji et al. [42] adopted Rand Index to validate the performance of k-prototype in grouping three datasets, zoo, heart diseases and credit approval.

Remarkably, the literature mentioned above is reasonably relevant to the clustering context. However, no documented study has constructed explicit clustering in solving flood evacuation problems, especially victim behavior analysis that is highly governed by numerous mixed-type features. In contrast to the existing studies, this paper possesses the contributions in the following aspects:

- numerous influential factors including risk perception, compliance level and arrival pattern of victims are considered in the victim grouping, which address the uncertainty in evacuation demand during the evacuation planning;
- various flood scenarios are included in the grouping of victims, which allows the clustering result to be applicable for different scenarios under uncertainty;
- the use of k-prototype clustering algorithm in grouping of victims prevents unnecessary conversion of data types, which protects the dataset from any destruction;
- the proposed approach is relatively flexible to be modified for other applications and the resultant findings provide useful insights for disaster management, particularly in reducing the harmful impacts of flood occurrence.

3 Methodology

As displayed in Fig. 1, this study aims to perform necessary grouping for various victim categories (including vulnerable people). In view of the fact that victims would behave differently, various influential elements including the personal characteristics, risk perception (i.e., the perceived risk level of victims according to varying water levels of floods) [16-17], compliance level of victims on evacuation commands [1] and also travel patterns (i.e., the arrival of victims at the evacuation pick-up locations) [19] would be captured explicitly in performing the grouping of victims.

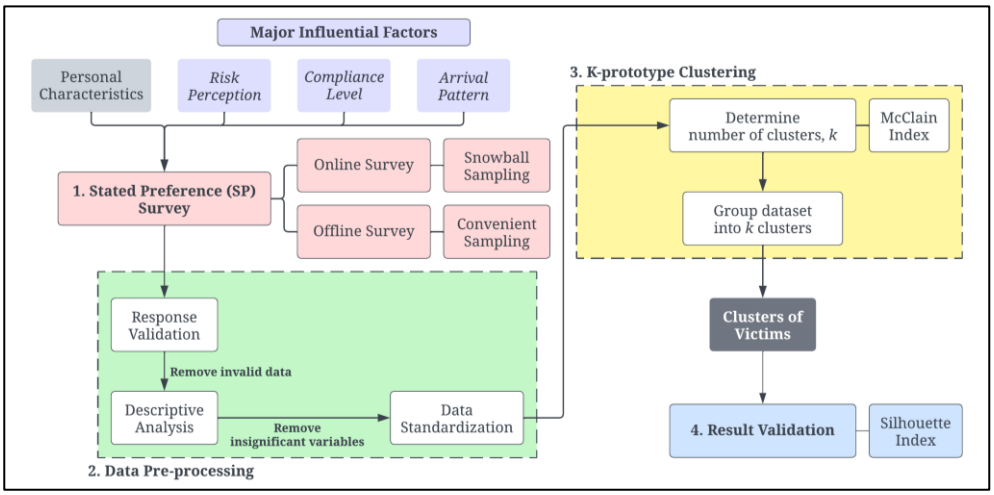

Fig. 1. Overall methodology for victim clustering.

3.1 Survey for data collection

For data collection purposes, a stated preference (SP) survey questionnaire can be designed prior to the survey, with the aim to perform grouping of victims according to the respondents' demographics and major influential factors. SP survey is chosen instead of revealed preference (RP) survey because SP survey captures responses from respondents who have and have not experienced flood, while RP survey only captures responses from respondents who have experienced flood. SP survey enables the capture of responses that was unable to be captured in RP survey, allowing the analyzed result to be applicable to the population [43]. SP survey also has the ability to analyze respondents' reaction to future events while RP survey only allow the analysis of existing options. Bočkarjova et al. [44], who performed the valuation of flood risk in Netherlands, stated that SP survey ensures that the valuation of risk obtained can be objective.

The survey can be conducted via both online and offline surveys. For online survey, snowball sampling method can be used whereby the survey can be distributed to family and friends, and spread out through their social circle. Survey link can also be sent through e-mail and be posted in social media to seek for respondents. For offline survey, convenience sampling method can be used whereby the survey can be done in convenient locations such as shopping malls, bus terminals and restaurants whereby people are asked to fill in the survey questionnaire. As both snowball sampling and convenience sampling are non-probabilistic sampling methods, the cost associated is low, allowing the survey to be conducted in resource-limited condition [45].

After the survey is conducted, the offline survey data needs to be digitalized to be combined with the online survey data. Then, response validation needs to be carried out to remove invalid responses from the responses collected. Invalid responses include responses from respondents below age 15, foreign respondents who filled in foreign states and cities, and incomplete responses. The valid responses then can be used for grouping of victim behavior by using k-prototype clustering algorithm.

3.2 K-prototype clustering

Before performing k-prototype clustering, data standardization can be performed for all numerical variables in the dataset to ensure they are unitless. Data standardization can be performed using Z-score standardization [46], in which the mean is subtracted from the numerical data and divided by their standard deviation. This can be represented by Equation (1):

$$Z(x_i) = \frac{x_i - \bar{x}}{\sigma} \tag{1}$$

where $Z(x_i)$ represents the standardized value, x_i represents the original value, $\bar{x}$ is the mean of the variable while σ is the standard deviation of the variable. Data standardization leads to a more accurate clustering result [46].

After performing data standardization, k-prototype clustering can be performed for the grouping of victim behavior. According to Huang [47], k-prototype algorithm is a clustering algorithm that can cluster both numerical and categorical data. The objective function is to minimize the total distance between every observation in the sample and the prototype observation, which is outlined in Equation (2) below:

$$\sum_{k=1}^{q} \sum_{i=1}^{n} w_{i,k}\, d_{MT}(X_i, c_k) \tag{2}$$

where q represents the number of clusters, n represents the number of objects, X_i represents the observation in the sample while c_k represents the prototype observation of cluster k. For the entries in binary partition matrix $[w]_{i=1,\dots,n;k=1,\dots,q}$, $w_{i,k} \in \{0,1\}$ and $\sum_{k=1}^{q} w_{i,k} = 1$.

From Equation (2), $d_{MT}(X_i, c_k)$ is the intuitive distance between X_i and c_k which can be represented by Equation (3):

$$d_{MT}(X,Y) = \sum_{j=1}^{l} (x_j - y_j)^2 + \lambda \sum_{j=l+1}^{m} \delta(x_j, y_j) \tag{3}$$

where l represents number of numerical variables, m represents total number of variables, λ is the weighing parameter that is greater than 0, and $\delta(x_j, y_j)$ is simple matching whereby the value is 0 when $x_j = y_j$ and 1 when $x_j \neq y_j$ [47-48].

The intuitive distance $d_{MT}(X_i, c_k)$ corresponds to weighted sum of Euclidean distance between two points in the metric space and simple matching distance for categorical variables (i.e., the count of mismatches). Trade-off between both terms can be controlled by the parameter λ which has to be specified in advance as well as the number of clusters. The impact of categorical variables increases when λ increases. The parameter λ is obtained from average standard deviation of all numerical variables with the range between $\frac{1}{3}\sigma$ and $\frac{p}{3}\sigma$, while the commonly used λ is $\frac{p}{6}\sigma$ (p is the sum of all variables while σ is the standard deviation of all numerical variables) [35, 47].

In general, k-prototype algorithm is performed according to the steps below [48]:

1. Determine the first k prototypes (centroids of the clusters).
2. For each observation:
   (a) Calculate $d_{MT}(X_i, c_k)$, the distance or similarity of data points on the data set against the prototype.
   (b) Assign observations to its closest prototype according to $d_{MT}(X_i, c_k)$.
   (c) Update cluster prototypes by cluster-specific means/modes for all variables.
   (d) Re-group all observations on the new prototypes.
3. Repeat step 2 until the centroids do not change or have converged, or the maximum number of iterations has been reached.

3.2.1 McClain Index

To determine the optimal number of clusters, k, McClain Index is used accordingly. McClain Index, also known as McClain-Rao Index, measures the ratio of the within-cluster and between-cluster distances [41].

According to McClain and Rao [49], the steps for computing McClain Index is as below:

1. Calculate N_W, the number of pairs of distinct points in every cluster, and S_W, the sum of within-cluster distances. The mean within-cluster distance, $\bar{S}_W$, is then calculated by dividing S_W by N_W as shown in Equation (4):
$$\bar{S}_W = \frac{S_W}{N_W} \tag{4}$$
2. Calculate N_B, the number of pairs constituted of points which do not belong to the same cluster, and S_B, the sum of between-cluster distances. The mean between-cluster distance, $\bar{S}_B$, is then calculated by dividing S_B by N_B as shown in Equation (5):
$$\bar{S}_B = \frac{S_B}{N_B} \tag{5}$$

3. The McClain Index, $v_{McClain}$, can be calculated by dividing the mean within-cluster distance by the mean between-cluster distance. This step can be represented by Equation (6):

$$v_{McClain} = \frac{\bar{S}_W}{\bar{S}_B} = \frac{S_W/N_W}{S_B/N_B} \tag{6}$$

The range of McClain Index is between 0 and ∞, and a McClain Index closer to 0 means that the mean within-cluster distance is extremely lower than the mean between-cluster distance, indicating that all objects in the dataset are appropriately clustered. Therefore, a smaller McClain Index indicates better clustering results and is preferred [39, 49].

3.2.2 Silhouette Index

To validate the clustering result, Silhouette Index is used. Silhouette Index, also known as Average Silhouette Width, measures how similar an object is to its own cluster compared to other clusters [38].

According to Rousseeuw [50], the steps for computing Silhouette Index is as below:

1. For object X_i in cluster C_A,
   (a) Calculate $a(X_i)$, the mean distance between object X_i and all other objects in the same cluster C_A. This step can be represented by Equation (7):
   (b)

$$a(X_i) = \frac{1}{n_A - 1} \sum_{j \in C_A, i \neq j} d(X_i, X_j) \tag{7}$$

   (c) Choose another cluster, C_S. Calculate the mean dissimilarity, which is the mean distance between object X_i and all other objects in cluster C_S. Repeat the step for all other clusters and choose $b(X_i)$, the smallest mean dissimilarity. This step can be represented by Equation (8):

$$b(X_i) = \min_{S \neq A} \frac{1}{n_S} \sum_{j \in C_S} d(X_i, X_j) \tag{8}$$

   (d) Calculate $s(X_i)$, the silhouette value of object X_i using Equation (9):

$$s(X_i) = \frac{b(X_i) - a(X_i)}{\max(a(X_i), b(X_i))} \tag{9}$$

2. Repeat the steps for all objects in the dataset. The Silhouette Index, $v_{Silhouette}$, can be calculated by finding the average of all $s(X_i)$ obtained. This step can be represented by Equation (10):

$$v_{Silhouette} = \frac{1}{n} \sum_{i=1}^{n} s(X_i) = \frac{1}{n} \sum_{i=1}^{n} \frac{b(X_i) - a(X_i)}{\max(a(X_i), b(X_i))} \tag{10}$$

The range of Silhouette Index is between -1 and 1, and a Silhouette Index closer to 1 indicates that all objects in the dataset are appropriately clustered. Therefore, a larger Silhouette Index indicates better clustering results and is preferred [38, 50].

4 An illustrative case study

The Central Region of Malaysia, which includes Selangor, Wilayah Persekutuan Kuala Lumpur and Wilayah Persekutuan Putrajaya, which has been affected by floods frequently, was selected for analysis purposes. Recently, heavy downpours on 17 December 2023 (due to the northeast monsoon) have caused worrying floods that were affecting four states (Kelantan, Terengganu, Perak and Selangor), causing the displacement of more than 6500 residents [51]. Another heavy downpour in November 2023 have also caused flash floods in Selangor, Perak and Pahang, displacing nearly 1000 residents although no casualties were reported [52]. Therefore, Central Region of Malaysia has been chosen as the study area as displayed in Fig. 2.

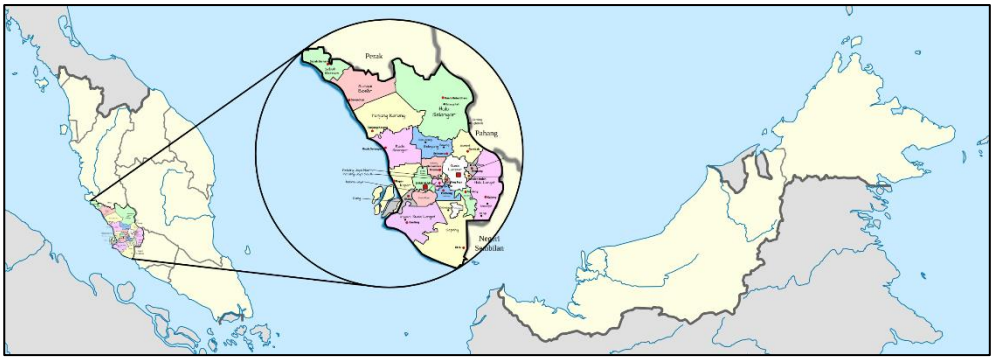

Fig. 2. The study area of Central Region of Malaysia.

An SP survey was properly designed to collect responses for the grouping of victims. The survey is divided into two sections: Section 1 is about respondents' demographics while Section 2 is about their response of flood evacuation on different scenarios. In Section 1, various personal information (i.e., gender, age, highest education level, occupation, household size, daily transport mode etc.) are included in the survey as these factors were identified as influential factors to victim behaviour in the past studies [15, 21, 24]. The respondents' demographics captured consist of 4 numerical variables and 8 categorical variables. The numerical variables include age, household size, number of times of experiencing flood, and number of times of experiencing flood evacuation. The categorical variables are: gender, pregnant, highest education level, occupation/employment status, residence type, residence state/federal territory, daily transport mode, and special care (disability).

Based on the three major influential factors (i.e., risk perception [16-17], compliance level [1], arrival pattern [19]) with two conditions for each factor, eight hypothetical scenarios are presented to the respondents to choose their tendency to evacuate with the evacuation vehicle arranged by the disaster planner. The respondents can choose their tendency level from four scales: very likely, somewhat likely, not very likely and very unlikely. The response of victim behavior can be identified for 8 scenarios, which are shown in Table 1.

Table 1. Risk perception, compliance level and arrival pattern of all scenarios.

Scenario	Risk Perception	Compliance Level	Arrival Pattern
1	Low/medium water level	Tend to obey evacuation commands	Punctual/arrive earlier at the pick-up location
2	Low/medium water level	Tend to obey evacuation commands	Late arrival at the pick-up location/no-show

3	Low/medium water level	Reluctant to obey evacuation commands	Punctual/arrive earlier at the pick-up location
4	Low/medium water level	Reluctant to obey evacuation commands	Late arrival at the pick-up location/no-show
5	High water level	Tend to obey evacuation commands	Punctual/arrive earlier at the pick-up location
6	High water level	Tend to obey evacuation commands	Late arrival at the pick-up location/no-show
7	High water level	Reluctant to obey evacuation commands	Punctual/arrive earlier at the pick-up location
8	High water level	Reluctant to obey evacuation commands	Late arrival at the pick-up location/no-show

The survey was conducted in November and December 2022 via online and offline mode. For online survey, snowball sampling method was applied in which the survey was distributed to family and friends through social media, and was spread out through their social circle. The survey was also sent via e-mail to seek for respondents. For offline survey, convenient sampling method was applied whereby the survey was distributed in the shopping malls and bus terminals to seek for respondents. Throughout the survey period, 4 student assistants were involved in conducting both online and offline surveys.

5 Results and discussion

5.1 Findings from survey

A total of 1011 responses were collected from the SP survey, 429 from online survey and 582 from offline survey. After response validation, a total of 870 responses are valid and were used for grouping of victim behavior, indicating the survey completion rate of 86%. The valid responses were filtered and 498 responses from Central Region (as shown in Table 2) were used for the victim clustering.

As shown in Table 2, it could be seen that 42% of the respondents are male and 58% are female. About 72% of the respondents are 30 years old and below, and 53% of the respondents are students, implying that most of the respondents of the survey are adolescents and young adults. Among the respondents, 73% have the household size between 4 to 6. As the study focuses on the Central Region of Malaysia, 33% of the respondents live in apartments or condominiums, and 43% of the respondents live in terraced houses of double storey and above. Only 21% of the respondents use public transport as their daily transport mode, while the remaining 78% use private vehicle as their daily transport mode. Out of the 498 respondents, only 4% (21 respondents) require special care.

Although floods always occur in Central Region of Malaysia, only 32% of the respondents have experienced floods of various severity. Only 13% of the respondents have experienced flood evacuation, which implies that 19% of the respondents did not evacuate during the floods experienced before. Therefore, it is important to know their willingness to evacuate (with the evacuation vehicles from disaster planner's team) in various scenarios.

Table 3 shows the responses of the respondents on the 8 hypothetical scenarios (as described in Table 1). In scenarios 1 and 5, the percentage of responses to evacuate ("very likely" and "somewhat likely") is high. Both scenarios 1 and 5 share the same characteristics: the victims tend to obey evacuation commands and arrive early/on time at the pick-up locations. The respondents stated that the likelihood to evacuate is high in this case for both low and high water level. On the other hand, in scenarios 4 and 8 in which the victims are reluctant to obey evacuation commands and arrive late at the pick-up locations or no show-

up, the percentage of responses to not evacuate ("not very likely" and "very unlikely") is high. In scenarios 2, 3, 6 and 7, the percentage of responses for "somewhat likely" and "not very likely" are high, implying that the likelihood of evacuation is moderate. This means that the uncertainty in evacuation demand for these scenarios is higher, which necessitates the victim clustering process.

5.2 K-prototype clustering results

Before performing the k-prototype clustering with the aid of R programming, 3 variables were removed from the dataset. "Pregnant" and "Special Care (Disability)" were removed as they are insignificant for the clustering (because most of the responses belong to one of the categories). "Residence State" was removed as this study focuses on Central Region of Malaysia and it is unnecessary to group the victims again based on their Residence State.

Table 2. Characteristics of 498 respondents.

Variable	Category	No. of Respondents	Percentage
Gender	Male	208	42%
	Female	290	58%
Pregnant	Yes	5	1%
	No	493	99%
Age	≤ 18	76	15%
	19 – 30	283	57%
	≥ 31	139	28%
Highest Education Level	SRP/PMR/PT3	35	7%
	SPM/O-Level/UEC	89	18%
	Pre-U/Matriculation/STPM/ A-Level/Foundation	90	18%
	Certificate/Diploma	64	13%
	Degree/Advanced Diploma	190	38%
	Master/PhD	24	5%
	Others	6	1%
Occupation	Student	262	53%
	Government Servant	8	2%
	Businessperson	38	8%
	Admin/Executive	75	15%
	Professional	30	6%
	Manager/Top Management	20	4%
	Unemployed	30	6%
	Retired	9	2%
	Others	26	5%
Household Size	1 – 3	96	19%
	4 – 6	366	73%
	≥ 7	36	7%
Residence Type	Apartment/Condominium	166	33%
	Terraced House (1 storey)	78	16%
	Terraced House (2 storey or above)	216	43%
	Semi-detached House/Detached House	32	6%
	Others	6	1%
Residence State	Selangor	313	63%
	Kuala Lumpur	184	37%
	Putrajaya	1	0%
	Private Vehicle	394	79%

Daily Transport Mode	Public Transport	104	21%
Special Care (Disability)	Yes	21	4%
	No	477	96%
Flood Experience	0	337	68%
	≥ 1	161	32%
Flood Evacuation Experience	0	433	87%
	≥ 1	65	13%

Table 3. Responses to hypothetical scenarios (498 respondents).

Hypothetical Scenario	Category	No. of Respondents	Percentage
Scenario 1	Very Likely	145	29%
	Somewhat Likely	210	42 %
	Not Very Likely	106	21%
	Very Unlikely	37	7%
Scenario 2	Very Likely	47	9%
	Somewhat Likely	200	40%
	Not Very Likely	170	34%
	Very Unlikely	81	16%
Scenario 3	Very Likely	40	8%
	Somewhat Likely	185	37%
	Not Very Likely	179	36%
	Very Unlikely	94	19%
Scenario 4	Very Likely	47	9%
	Somewhat Likely	124	25%
	Not Very Likely	145	29%
	Very Unlikely	182	37%
Scenario 5	Very Likely	287	58%
	Somewhat Likely	128	26%
	Not Very Likely	55	11%
	Very Unlikely	28	6%
Scenario 6	Very Likely	91	18%
	Somewhat Likely	178	36%
	Not Very Likely	157	32%
	Very Unlikely	72	14%
Scenario 7	Very Likely	65	13%
	Somewhat Likely	139	28%
	Not Very Likely	158	32%
	Very Unlikely	136	27%
Scenario 8	Very Likely	58	12%
	Somewhat Likely	96	19%
	Not Very Likely	104	21%
	Very Unlikely	240	48%

5.2.1 Optimal clusters of victims

Tables 4 and 5 show the cluster prototypes of all 8 scenarios (with "Daily Transport Mode" removed as all cluster prototypes identified "private vehicle" as the daily transport mode). The trade-off between numerical and categorical variables, λ, ranges from 1.64 to 1.70, indicating that the impact of categorical variables in the clustering process is 64% to 70% higher than the impact of numerical variables.

As lower McClain Index indicates a better clustering result, $k = 4$ is determined as the optimal number of clusters for scenarios 1, 2, 3, 5, 7 and 8. Scenario 4 has the optimal number of clusters of 2, while scenario 6 has the optimal number of clusters of 5. It can be observed that the McClain Indices for the optimal number of clusters for all 8 scenarios are between 0.53 to 0.56, indicating that the mean within-cluster distance is much lower than the mean between-cluster distance. The cluster sizes of the 31 clusters range from 21 (4%) to 347 (70%), with most of them range between 88 (18%) to 189 (38%). For the optimal number of clusters between 2 to 5, the cluster sizes of 18% to 38% for most of the clusters is appropriate, while small cluster size such as 4% allows the clustering process to capture the minority's response.

Table 4. Cluster prototypes of Scenarios 1-4.

Scenario	Estimated lambda	Number of Clusters	McClain Index	Silhouette Index	Cluster Sizes	Cluster Prototypes								
						Gender	Age	Highest education level	Occupation/ employment status	Household size	Residence type	Number of times of experiencing flood	Number of times of experiencing flood evacuation	Response
Scenario 1	1.66	4	0.56	0.10	42	Male	36.86	Degree/Adv. Diploma	Businessperson	5.21	Terraced House (2 storey or above)	2.50	1.31	Not Very Likely
					183	Female	20.26	Pre-U/Matric./STPM/ A-Level/Foundation	Student	5.21	Terraced House (2 storey or above)	0.36	0.05	Somewhat Likely
					156	Male	22.97	Degree/Adv. Diploma	Student	3.81	Apartment/Condo.	0.29	0.06	Very Likely
					117	Female	39.08	Degree/Adv. Diploma	Admin/Executive	4.33	Terraced House (2 storey or above)	0.61	0.06	Somewhat Likely
Scenario 2	1.66	4	0.54	0.12	35	Female	28.14	Degree/Adv. Diploma	Student	5.66	Terraced House (2 storey or above)	3.03	1.37	Not Very Likely
					178	Female	20.11	Pre-U/Matric./STPM/ A-Level/Foundation	Student	5.18	Terraced House (2 storey or above)	0.24	0.02	Somewhat Likely
					163	Male	23.48	Degree/Adv. Diploma	Student	3.83	Apartment/Condo.	0.32	0.07	Somewhat Likely
					122	Female	41.15	Degree/Adv. Diploma	Admin/Executive	4.34	Terraced House (2 storey or above)	0.71	0.14	Somewhat Likely
Scenario 3	1.65	4	0.55	0.10	45	Male	35.98	SPM/O-Level/UEC	Businessperson	5.13	Terraced House (2 storey or above)	2.56	1.27	Very Unlikely
					180	Female	20.40	Pre-U/Matric./STPM/ A-Level/Foundation	Student	5.19	Terraced House (2 storey or above)	0.29	0.06	Somewhat Likely
					162	Male	22.89	Degree/Adv. Diploma	Student	3.94	Apartment/Condo.	0.34	0.06	Not Very Likely
					111	Female	39.76	Degree/Adv. Diploma	Admin/Executive	4.23	Terraced House (2 storey or above)	0.58	0.05	Not Very Likely
Scenario 4	1.64	2	0.56	0.34	151	Male	39.60	Degree/Adv. Diploma	Admin/Executive	4.12	Apartment/Condo.	1.05	0.34	Very Unlikely
					347	Female	21.42	Degree/Adv. Diploma	Student	4.76	Terraced House (2 storey or above)	0.37	0.08	Very Unlikely

Table 5. Cluster prototypes of Scenarios 5-8.

Scenario	Estimated lambda	Number of Clusters	McClain Index	Silhouette Index	Cluster Sizes	Cluster Prototypes								
						Gender	Age	Highest education level	Occupation/ employment status	Household size	Residence type	Number of times of experiencing flood	Number of times of experiencing flood evacuation	Response
Scenario 5	1.70	4	0.53	0.14	44	Female	28.43	Degree/Adv. Diploma	Student	4.75	Terraced House (2 storey or above)	2.52	1.32	Somewhat Likely
					189	Female	20.45	Pre-U/Matric./STPM/ A-Level/Foundation	Student	5.21	Terraced House (2 storey or above)	0.26	0.02	Very Likely
					177	Male	24.50	Degree/Adv. Diploma	Student	3.69	Apartment/Condo.	0.28	0.04	Very Likely
					88	Female	45.00	Degree/Adv. Diploma	Businessperson	4.86	Terraced House (2 storey or above)	0.86	0.14	Very Likely
Scenario 6	1.64	5	0.55	0.10	21	Female	30.48	Degree/Adv. Diploma	Student	6.00	Terraced House (2 storey or above)	4.14	1.67	Somewhat Likely
					149	Female	20.87	Pre-U/Matric./STPM/ A-Level/Foundation	Student	5.41	Terraced House (2 storey or above)	0.34	0.07	Not Very Likely
					127	Male	24.60	Degree/Adv. Diploma	Student	4.17	Apartment/Condo.	0.36	0.06	Somewhat Likely
					105	Female	42.88	Degree/Adv. Diploma	Admin/Executive	4.41	Terraced House (2 storey or above)	0.77	0.17	Somewhat Likely
					96	Female	21.21	SPM/O-Level/UEC	Student	3.64	Apartment/Condo.	0.24	0.09	Not Very Likely
Scenario 7	1.64	4	0.55	0.11	42	Male	37.10	SPM/O-Level/UEC	Businessperson	5.24	Terraced House (2 storey or above)	2.38	1.33	Somewhat Likely
					185	Female	20.32	Pre-U/Matric./STPM/ A-Level/Foundation	Student	5.20	Terraced House (2 storey or above)	0.38	0.06	Not Very Likely
					165	Male	23.06	Degree/Adv. Diploma	Student	3.87	Apartment/Condo.	0.33	0.06	Somewhat Likely
					106	Female	40.47	Degree/Adv. Diploma	Admin/Executive	4.29	Terraced House (2 storey or above)	0.58	0.03	Very Unlikely
Scenario 8	1.66	4	0.56	0.11	48	Male	32.29	SPM/O-Level/UEC	Student	5.21	Terraced House (2 storey or above)	2.56	1.27	Not Very Likely
					180	Female	20.13	Pre-U/Matric./STPM/ A-Level/Foundation	Student	5.21	Terraced House (2 storey or above)	0.27	0.04	Very Unlikely
					152	Male	23.20	Degree/Adv. Diploma	Student	3.80	Apartment/Condo.	0.26	0.04	Very Unlikely
					118	Female	39.93	Degree/Adv. Diploma	Admin/Executive	4.32	Terraced House (2 storey or above)	0.65	0.06	Very Unlikely

Among all 31 clusters from the 8 scenarios, 4 clusters (13%) show "very likely", 12 clusters (39%) show "somewhat likely", 8 clusters (26%) show "not very likely" and 7 clusters (23%) show "very unlikely" for the responses to evacuate. "Somewhat likely" appears as the mode of the evacuation response in the clusters, and the probable response to evacuate ("very likely" and "somewhat likely") is about 52% under uncertainty. The percentage of responses, calculated based on the cluster sizes and responses, were represented in the bar charts as shown in Fig. 3.

As shown in Fig. 3, scenarios 1, 2 and 5 show a majority response of "very likely" and "somewhat likely", with scenario 5 showing a majority response of "very likely". These 3 scenarios are of high compliance level (i.e., the victims tend to obey evacuation commands), indicating that victims with high compliance level has a higher likelihood to evacuate. The likelihood of evacuation in scenario 5 is significantly higher than scenarios 1 and 2 as scenario 5 is of high water level. Scenario 6, although of high water level and high compliance level, has a balanced response between "somewhat likely" and "not very likely". The contrast between scenario 5 and scenario 6 indicates that in the case of late arrival or no show, some of the victims may have evacuated by their own during the case of high water level.

On the other hand, scenarios 3, 4 and 8 show a majority response of "not very likely" and "very unlikely", which indicates that victims with low compliance level has a lower likelihood to evacuate. When low compliance level and late arrival/no-show happen together, the likelihood of evacuation will be even lower, which can be observed from scenarios 4 and 8 showing a majority response of "very unlikely". The victims in scenarios 3 and 4 have low likelihood to evacuate as they may choose to not evacuate in the case of low water level, while in scenario 8 where the water level is high, the victims may have evacuated by their own instead of waiting for the evacuation planner.

To determine the influence of each individual major influential element on the victims' likelihood to evacuate, the percentage of responses was calculated for all 4 responses in each category of influential factors as represented in Fig. 4 to Fig. 6. Fig. 4 implies that the effect of victims' risk perception on their likelihood of evacuation is lower. Regardless of low/medium water level or high water level, the likelihood to evacuate ("very likely" and "somewhat likely") and the likelihood to not evacuate ("not very likely" and "very unlikely") are both around 50%, indicating that risk perceptions must be considered together with other influential elements to better determine the likelihood of evacuation.

In Fig. 5, there is a strong contrast between high compliance level and low compliance level, in which victims who tend to obey evacuation commands have higher likelihood to evacuate (83.83%) while victims who are reluctant to obey evacuation commands have lower likelihood to evacuate (19.43%). This indicates that victims with higher compliance level are more likely to evacuate with the emergency planner, while victims with lower compliance level are more likely to not evacuate or to evacuate by their own.

Fig. 6 also shows a contrast between earlier arrival and late arrival of victims at pick-up location, whereby victims who arrive punctually or earlier at the pick-up location is more likely to evacuate (67.32%) than victims who arrive late or no-show (35.94%). However, the contrast is smaller than of compliance level, indicating that arrival pattern is less influential when compared to the compliance level of victims.

It can be concluded that among the three major influential factors, victims' compliance level to evacuation commands is the most influential factor on their likelihood to evacuate (83.83% vs 19.43%), followed by their arrival pattern at the pick-up location (67.32% vs 35.94%) and the risk perception (55.17% vs 48.09%). Therefore, when there is incomplete information on the three major influential elements, victims' compliance level can be prioritized when assessing their likelihood of evacuation.

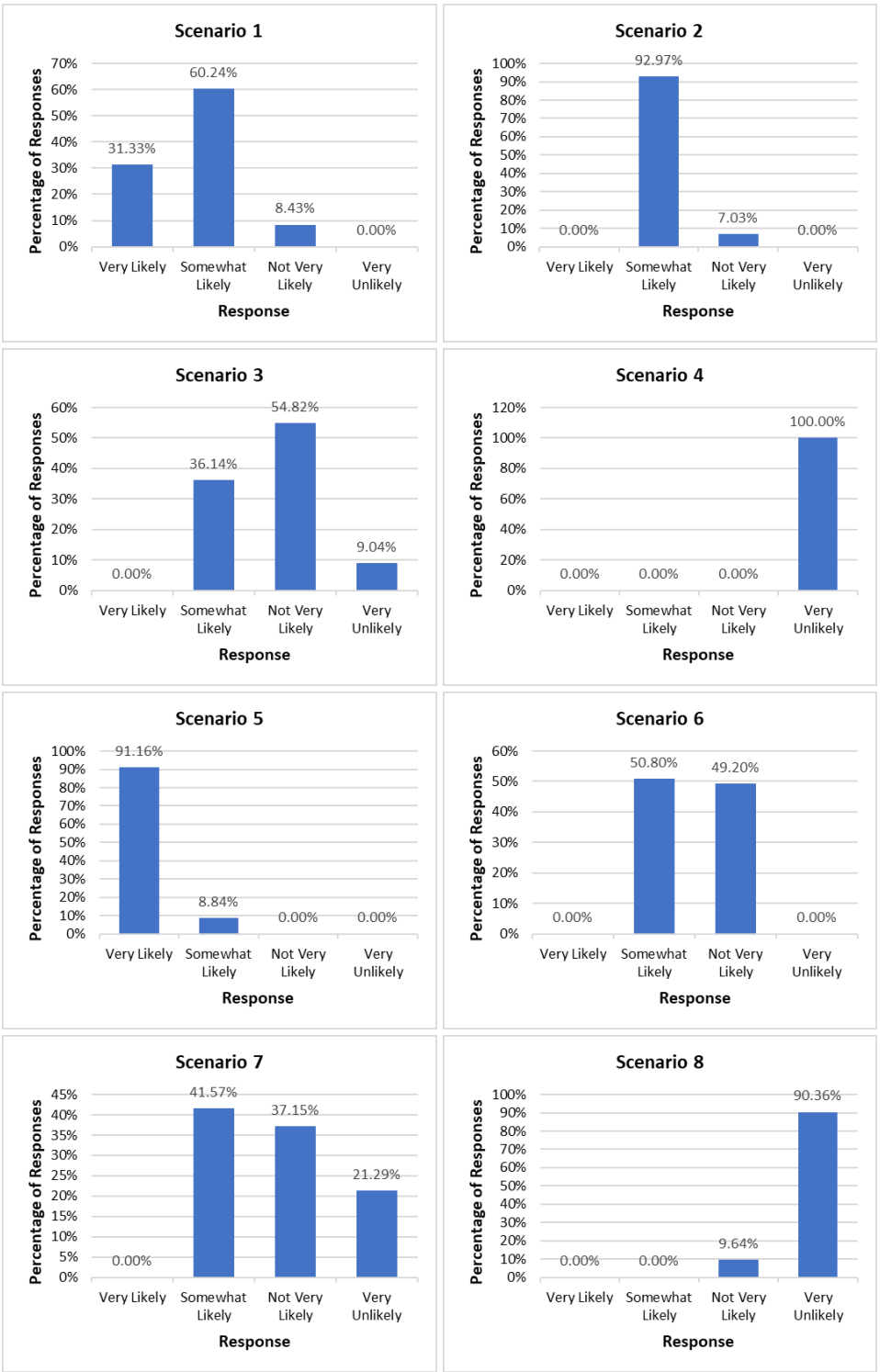

Fig. 3. The responses for all 8 scenarios.

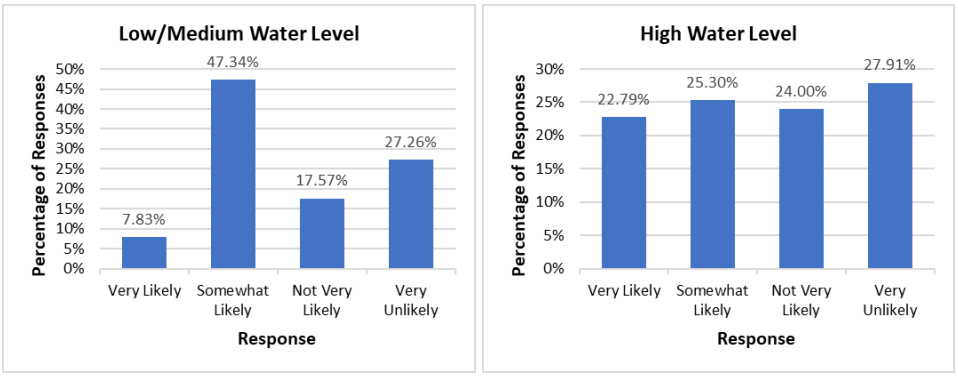

Fig. 4. The responses for risk perceptions.

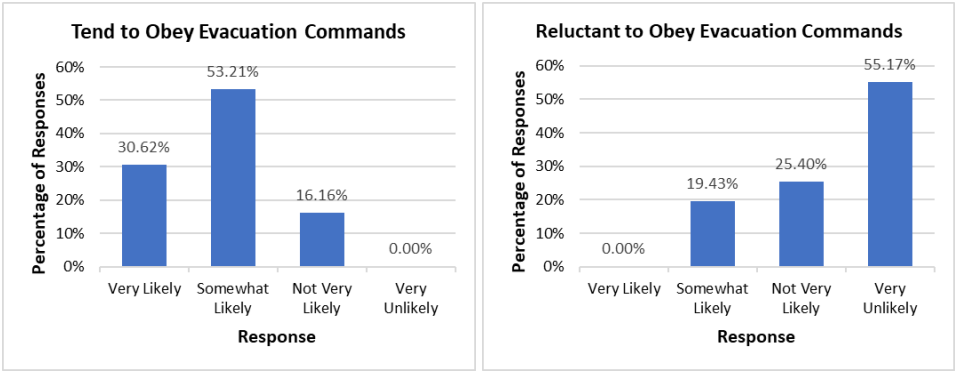

Fig. 5. The responses for compliance levels.

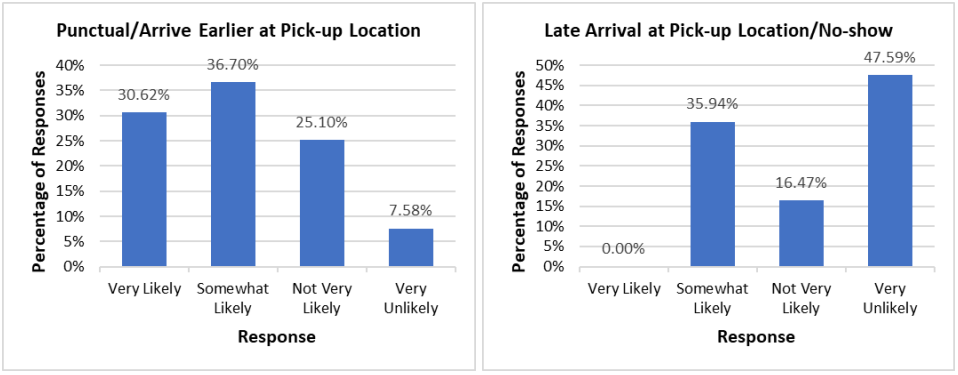

Fig. 6. The responses for arrival patterns.

5.2.2 Results validation

To ensure the reliability of the clustering result, Silhouette Index was adopted to validate the clustering result. Silhouette Index can be used to measure the quality of the clustering result by evaluating the diversity within a cluster [53-54]. Silhouette Index has the advantages of low computational complexity and simple interpretation rules [55], yet still ensures the high accuracy in performance validation [38], making it a performance index which is frequently used in various studies.

Silhouette Index ranges from -1 to +1, and a higher Silhouette Index indicates a better clustering result [38]. Scenario 4 shows the highest Silhouette Index which is 0.34, while for the other 7 scenarios, the Silhouette Index ranges from 0.10 to 0.14. It can be noticed that Scenario 4 with the Silhouette Index of 0.34 only has 2 clusters, while the other scenarios with the Silhouette Index of 0.10-0.14 have 4-5 clusters. This indicates that Silhouette Index prefers a smaller number of clusters as smaller number of clusters maximizes the mean dissimilarity, which increases the Silhouette Index. Nevertheless, a positive Silhouette Index indicates that the mean dissimilarity is larger than the mean distance between the object and other objects in the same cluster [50], implying that the clustering result is appropriate.

Concisely, the clustering result is useful in identifying the evacuation responses of victims in various scenarios, which can be further applied to quantify the demand level of victims during the evacuation planning. Scenario 5 has the highest likelihood of evacuation followed by Scenarios 1 and 2, while Scenarios 4 and 8 have the lowest likelihood of evacuation. Results of Scenarios 4 and 8 implies that with low compliance level and late arrival/no-show at pick-up location, the victims are very unlikely to evacuate with the evacuation planner regardless of the water level. Besides, it is also noticeable that victims' compliance level to evacuation commands and their arrival pattern at pick-up locations are influential to their likelihood of evacuation, which are in line to the facts revealed by Bayram [1] and Amideo et al. [19].

6 Conclusions

The frequent happening of floods necessitates the need of demand (behavior) analysis for better evacuation planning. Thus, this study is carried out to examine the victims on their likelihood of evacuation with the disaster planner in various hypothetical scenarios. The k-prototype clustering is adopted to group the victims according to similar characteristics to identify their probable response to evacuate. The optimal number of clusters ranges from 2 to 5 with McClain Index ranges from 0.53 to 0.56. Notably, scenarios 1, 2 and 5 have high likelihood of evacuation while scenarios 4 and 8 have low likelihood of evacuation, and the overall probable response to evacuate is 52% for all scenarios. The clustering results also imply that compliance level and arrival pattern are influential to the likelihood of evacuation. Result verification is performed via Silhouette Index, which ranges from 0.10 to 0.34, ensures that the result is reasonable and reliable.

Several limitations are identified from this study. Firstly, the analysis is only performed for the Central Region of Malaysia, therefore the results might not reflect the responses of other flood-prone regions. Besides, non-probabilistic sampling method (snowball sampling and convenient sampling) is adopted when conducting the survey, causing the response to be dominated by young students, and less responses of pregnant women and people requires special care could be captured. Future works can be done by focusing on other flood-prone regions for further analysis and adopt probabilistic sampling method for the survey to ensure that the responses reflects the targeted population well so that the demand analysis is applicable for the respective flood-prone regions for effective evacuation planning.

This research was supported by the Ministry of Higher Education (MoHE) Malaysia through Fundamental Research Grant Scheme project FRGS/1/2022/TK02/UTAR/02/2. The authors also want to thank student assistants in conducting the survey and all respondents of survey who had taken part voluntarily.

References

1. V. Bayram, Surveys in Oper. Res. and Mgmt. Sci. **21**, 63-84 (2016)
2. The Star, https://www.thestar.com.my/news/nation/2006/12/21/typhoon-utor-to-blame (2006)
3. Daily News, https://archives.dailynews.lk/2010/11/06/wld04.asp (2010)
4. CNN, https://edition.cnn.com/2023/09/16/world/global-rain-flooding-climate-crisis-intl-hnk/index.html (2023a)
5. The Guardian, https://www.theguardian.com/world/2023/oct/05/india-floods-death-toll-lhonak-lake-injuries-missing-sikkim (2023)
6. Channel News Asia, https://www.channelnewsasia.com/asia/malaysia-flood-johor-highest-rainfall-1991-evacuate-deaths-3326076 (2023)
7. CNN, https://edition.cnn.com/2023/03/05/asia/johor-malaysia-floods-intl-hnk/index.html (2023b)
8. The Star, https://www.thestar.com.my/lifestyle/living/2023/03/12/johor-floods-this-is-only-the-beginning-for-malaysia (2023)
9. M. Gama, B.F. Santos, M.P. Scaparra, EURO J. on Compt. Optim. **4**, 299-323 (2016)
10. E. Aroca-Jiménez, J. M. Bodoque, J. A. García, Sci. of the Total Environ. **746**, 140905 (2020)
11. M. J. Alam, M. A. Habib, E. Pothier, Int. J. of Disaster Risk Redn. **53**, 102016 (2021)
12. M. J. Alam, M. A. Habib, Trans. Res. Part D **97**, 102946 (2021)
13. W. K. Anuar, L. S. Lee, S. Pickl, H. V. Seow, Appl. Sci. **11**, 667 (2021)
14. H. Yang, E.F. Morgul, K. Ozbay, K. Xie, Trans. Res. Record **2599**, 63-69 (2016)
15. M.B.B. Lim, H.R. Lim Jr., M. Piantanakulchai, Asia Pac. Mgmt. Rev. **24**, 106-113 (2019)
16. K. E. Schlef, L. Kaboré, H. Karambiri, Y. C. E. Yang, C.M. Brown, Environ. Sci. and Policy **89**, 254-265 (2018)
17. F. Mahdavian, M. Wiens, S. Platt, F. Schultmann, Int. J. of Disaster Risk Redn. **49**, 101685 (2020)
18. W. Yi, L. Nozick, R. Davidson, B. Blanton, B. Colle, Trans. Res. Part B **95**, 285-304 (2017)
19. A. E. Amideo, M. P. Scaparra, K. Kotiadis, Eur. J. of Oper. Res. **279**, 279–295 (2019)
20. S. Liu, L. E. Quenemoen, J. Malilay, E. Noji, T. Sinks, J. Mendlein, Am. J. of Public Health **86**, 87-89 (1996)
21. Q. Yin, G. Ntim-Amo, D. Xu, V.K. Gamboc, R. Ran, J. Hu, H. Tang, Int. J. of Disaster Risk Redn. **78**, 103126 (2022)
22. P. K. Anyihodo, R. A. Davidson, T. Rambha, L.K. Nozick, Trans. Res. Part A **159**, 200-221 (2022)
23. S. Chantarat, S. Oum, K. Samphantharak, V. Sann, J. of Asian Economics **63**, 44-74 (2019)

24. M. B. B. Lim, H. R. Lim Jr., J. M. L. Anabo, Int. J. of Disaster Risk Redn. **56**, 102137 (2021)
25. G. Ntim-Amo, Q. Yin, E.K. Ankrah, Y. Liu, M. A. Twumasi, W. Agbenyo, D. Xu, S. Ansah, R. Mazhar, V.K. Gamboc, Int. J. of Disaster Risk Redn. **80**, 103223 (2022)
26. S. D. Wong, A. J. Pel, S. A. Shaheen, C. G. Chorus, Trans. Res. Part D **79**, 102227 (2020)
27. X. Zhao, R. Lovreglio, D. Nilsson, Autom. in Constr.**113**, 103140 (2020)
28. A. J. Pel, M. C. J. Bliemer, S. P. Hoogendoorn, Trans. **39**, 97-123 (2012)
29. M. Diakakis, Sust. **12**, 4409 (2020)
30. A. N. Qazi, Y. Nara, K. Okubo, H. Kubota, IATSS Res. **41**, 147–152 (2017)
31. K. Lin, H. Chen, C. Y. Xu, P. Yan, T. Lan, Z. Liu, C. Dong, J. of Hydro. **584**, 124696 (2020)
32. X. Xu, L. Zhang, M. Trovati, F. Palmieri, E. Asimakopoulou, O. Johnny, N. Bessis, Appl. Soft Comput. **111**, 107667 (2021)
33. R. Sreejith, K. R. Sinimole, Sust. Cities and Soc. **87**, 104257 (2022)
34. C. Li, X. Wu, X. Cheng, C. Fan, Z. Li, H. Fang, C. Shi, Environ. Res. **176**, 108568 (2019)
35. S. Sulastri, L. Usman, U.D. Syafitri, Indo. J. of Stat. and Its Appl. **5**, 228-242 (2021)
36. J. Soria, Y. Chen, A. Stathopoulos, Trans. Res. Record **2674**, 383–394 (2020)
37. Z. Jia, L. Song, Math. Problems in Eng. **2020**, 1-13 (2020)
38. R. Aschenbruck, G. Szepannek, Arch. of Data Sci., Series A **6**, 1-12 (2020)
39. Y. G. Choi, S. Ann, J. Kim, IEEE Access **11**, 75945-75954 (2023)
40. K. R. N. Kumar, H. Lahza, B. R. Sreenivasa, T. Shawly, A. A. Alsheikhy, H. Arunkumar, C. R. Nirmala, Comput. Syst. Sci. and Eng. **46**, 3239-3260 (2023)
41. Z. Šulc, M. Matějka, J. Procházka, H. Řezanková, Metodološki Zvezki **14**, 37-48 (2017)
42. J. Ji, W. Pang, Y. Zheng, Z. Wang, Z. Ma, L. Zhang, Appl. Math. & Info. Sci. **9**, 2933-2942 (2015)
43. R. Danielis, L. Rotaris, Trasporti Europei **13** (1999)
44. M. Bočkarjova, P. Rietveld, E. T. Verhoef, Safety, Rel. and Risk Anal.: Theory, Methods and Appl. **1**, 2781-2788 (2008)
45. A. F. Hulio, V. Varghese, M. Chikaraishi, Clim. Risk Mgmt. **42**, 100571 (2023)
46. I. Mohamad, D. Usman, Res. J. of Appl. Sci., Eng. and Tech. **6**, 3299-3303 (2013)
47. Z. Huang, *Clustering large data sets with mixed numeric and categorical values*, in Proceedings of the 1st Pacific-Asia Conference on Knowledge Discouvery And Data Mining, PAKDD, 23-24 February 1997, Singapore (1997)
48. G. Szepannek, The R J. **10**, 200-208 (2018)
49. J.O. McClain, V.R. Rao, J. of Marketing Res. **12**, 456-460 (1975)
50. P. J. Rousseeuw, J. of Compt. and Appl. Math. **20**, 53-65 (1987)
51. The Straits Times, https://www.straitstimes.com/asia/se-asia/floods-displace-over-6500-people-in-malaysia (2023a)
52. The Straits Times, https://www.straitstimes.com/asia/se-asia/heavy-rain-causes-floods-in-some-parts-of-malaysia (2023b)

53. Z. Ansari, M. F. Azeem, W. Ahmed, A. V. Babu, World of Comput. Sci. and Info. Tech. J. **1**, 217-226 (2011)
54. R. Novidianto, H. Wibowo, D. R. Chandranegara, Kinetik: Game Tech., Info. Syst., Comput. Network, Comput., Elec., and Ctrl. **6**, 109-116 (2021)
55. A. Dudek, Classif. and Data Anal.: Theory and Appl. **28**, 19-33 (2020)