Options for Earthquake Event Reduction in Stochastic Loss Modelling

Zur Erlangung des akademischen Grades einer

DOKTOR-INGENIEURIN

Von der KIT-Fakultät für
Bauingenieur-, Geo- und Umweltwissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

Shadi Shirazian

M.Sc. Structural

Geb. in Ardebil, Iran

Tag der mündlichen Prüfung: 21.06.2022

Referent: Prof. Dr. Friedemann Wenzel

Korreferent: Prof. Dr.-Ing. Lothar Stempniewski

Karlsruhe (2022)

Abstract

Infrastructures like transportation, power, and water networks facilitate life in modern societies, their construction is time-consuming and expensive, their role is vital in the aftermath of an earthquake; therefore, they should remain functional. However, they are vulnerable to earthquakes and any damage to them has a cascading effect that can significantly increase mortality and economic loss. But, although earthquakes are inevitable, losses are not. The seismic risk assessment provides invaluable information for the decisionmaker to apply risk mitigation strategies and allocate budgets to resilience projects and alleviate the adverse effects of earthquakes.

Stochastic seismic loss modeling is a probabilistic method to assess the risk in terms of loss exceedance curve for a range of return periods. In this method, each earthquake scenario is evaluated individually to calculate its consequent loss. It is especially suitable for risk assessment of spatially distributed structures like lifelines and portfolios. However, the requirement of including many earthquakes, and modeling the loss due to each of them makes this method computationally inefficient, and the complexity of loss modeling of infrastructures exacerbates this downside.

The scope of this thesis is to develop methodologies to reduce the required number of events and decrease the computational demand in stochastic risk assessment. So far, most of the studies that have been conducted to reduce the number of events have used a hazard-consistent approach. This means that the reduced event set should generate similar hazard values when compared to the results of the full event set with the assumption that similar hazards result in similar consequences. In the methodologies of this research, the effect of risk will be included in reducing the number of events. Also, the sensitivity of risk to hazard will be studied to see how different the resulted hazard estimation of the reduced dataset can become, while still developing a satisfactory risk assessment. Moreover, the applicability of the resulted reduced dataset in further risk assessments like retrofit prioritizations is explored.

The reduction methodology of this research applies clustering algorithms to cluster similar events and from each cluster, one representative event is selected for the reduced dataset. The definition of similarity is different in each one of the approaches of this research. The approaches are 1) Risk-Consistent: only risk influences clustering of events. 2) Weighted Risk-Hazard: a weighted combination of risk and hazard influences the clustering, where the effect of risk is more dominant. 3) Unweighted Risk-Hazard: an unweighted combination of risk and hazard influences the clustering. 4) Hazard-Consistent: only hazard influences the clustering. The comparison of results of approaches 1 to 4 shows that increasing the effect of risk in the reduction process improves the risk estimations with the reduced dataset however the hazard estimations might not be satisfactory. To get good estimations for both risk and hazard, greater sizes of the reduced dataset is required.

The case of study is a transportation network in Erzincan, a city in eastern Turkey located in a highly seismic area in the vicinity of active faults such as the North Anatolian fault. Erzincan has experienced earthquakes with magnitudes around 7 several times in the last 1000 years.

Keywords: Infrastructures, stochastic risk assessment, earthquake, data-reduction, clustering

Zusammenfassung

Infrastrukturen wie Verkehrs-, Strom- und Wassernetze erleichtern das Leben in modernen Gesellschaften, ihr Bau ist zeit- und kostenaufwändig, ihre Funktion ist nach einem Erdbeben von entscheidender Bedeutung; deshalb sollten sie funktionsfähig bleiben. Sie sind jedoch erdbebengefährdet, und jeder Schaden, der ihnen zugefügt wird, hat eine kaskadenartige Wirkung, die die Sterblichkeit und die wirtschaftlichen Verluste erheblich erhöhen kann. Doch auch wenn Erdbeben unvermeidlich sind, sind es die Verluste nicht. Die seismische Risikobewertung liefert den Entscheidungsträgern wertvolle Informationen für die Anwendung von Risikominderungsstrategien, die Zuweisung von Mitteln für Resilienz Projekte, und die Abschwächung der negativen Auswirkungen von Erdbeben.

Die stochastische Modellierung seismischer Verluste ist eine probabilistische Methode zur Bewertung des Risikos in Form einer Verlustüberschreitungskurve für eine Reihe von Wiederkehrperioden. Bei dieser Methode wird jedes Erdbebenszenario einzeln bewertet, um den daraus resultierenden Schaden zu berechnen. Sie eignet sich besonders für die Risikobewertung von räumlich verteilten Strukturen wie Lebensadern und Portfolios. Die Notwendigkeit, viele Erdbeben einzubeziehen und den auf jeden von ihnen entfallenden Verlust zu modellieren, macht diese Methode jedoch rechnerisch ineffizient, und die Komplexität der Schadenmodellierung von Infrastrukturen verschärft diesen Nachteil noch.

Ziel dieser Arbeit ist es, Methoden zu entwickeln, die die Anzahl der erforderlichen Ereignisse und den Rechenaufwand bei der stochastischen Risikobewertung verringern. Bisher haben die meisten Studien zur Reduzierung der Anzahl der Ereignisse einen gefährdungskonsistenten Ansatz verwendet. Das bedeutet, dass der reduzierte Ereignissatz ähnliche Gefahrenwerte erzeugen sollte, wenn er mit den Ergebnissen des vollständigen Ereignissatzes verglichen wird, wobei davon ausgegangen wird, dass ähnliche Gefahren zu ähnlichen Konsequenzen führen. Bei den Methoden dieser Forschung wird die Auswirkung des Risikos bei der Reduzierung der Anzahl der Ereignisse berücksichtigt. Außerdem wird die Sensitivität des Risikos gegenüber der Gefahr untersucht, um festzustellen, wie unterschiedlich die aus dem reduzierten Datensatz resultierende Gefahrenabschätzung ausfallen kann, während immer noch eine zufriedenstellende Risikobewertung entwickelt wird. Darüber hinaus wird die Anwendbarkeit des resultierenden reduzierten Datensatzes in weiteren Risikobewertungen wie der Priorisierung von Nachrüstungen untersucht.

Die Reduktionsmethodik dieser Forschung wendet Clustering-Algorithmen an, um ähnliche Ereignisse zu clustern, und aus jedem Cluster wird ein repräsentatives Ereignis für den reduzierten Datensatz ausgewählt. Die Definition von Ähnlichkeit ist in jedem der Ansätze dieser Forschung unterschiedlich. Die Ansätze sind 1) Risikokonsistent: nur das Risiko beeinflusst die Clusterung von Ereignissen. 2) Gewichteter Risiko-Gefährdung: Eine gewichtete Kombination aus Risiko und Gefährdung beeinflusst die Clusterbildung, wobei der Effekt des Risikos dominanter ist. 3) Ungewichteter Risiko-Gefährdung: Eine ungewichtete Kombination aus Risiko und Gefährdung beeinflusst die Clusterbildung. 4) Gefahr-Konsistent: Nur die Gefahr beeinflusst die Clusterbildung. Der Vergleich der Ergebnisse der Ansätze 1 bis 4 zeigt, dass eine Erhöhung des Risikoeffekts im Reduktionsprozess zwar die Risikoschätzungen mit dem reduzierten Datensatz verbessert, aber die Gefahrenschätzungen möglicherweise nicht zufriedenstellend sind. Um gute Schätzungen sowohl für das Risiko als auch für die Gefahr zu erhalten, ist eine größere Größe des reduzierten Datensatzes erforderlich.

Bei dem untersuchten Fall handelt es sich um ein Verkehrsnetz in Erzincan, einer Stadt im Osten der Türkei, die in einem stark erdbebengefährdeten Gebiet in der Nähe aktiver Verwerfungen wie der Nordanatolischen Verwerfung liegt. Erzincan hat in den letzten 1000 Jahren mehrmals Erdbeben der Stärke 7 erlebt.

Schlüsselwörter: Infrastrukturen, stochastische Risikobewertung, Erdbeben, Datenreduzierung, Clustering

Acknowledgements

I am profoundly grateful to my advisor Prof. Friedemann Wenzel for his guidance, help, constructive and motivating comments on this research, and of course, all the meetings, especially during the previous two years with Covid restrictions. Also, special thanks to my Korreferent Prof. Lothar Stempniewski for evaluating my dissertation and his invaluable comments on this research.

I am very thankful to all members of my doctoral committee: Prof. Stefan Hinz (the chair of the committee), Prof. Bozidar Stojadinovic, Prof. Thomas Kohl, Prof. Frank Schilling, Prof. Mustafa Erdik, and Prof. Shervin Haghsheno for their time, and their brilliant suggestions and comments on my dissertation.

Furthermore, I want to thank Dr. James Edward Daniell for his help during my Ph.D., especially in collecting data and resources. I also thank the Graduate School for Climate and Environment (GRACE) at KIT for providing the opportunities of attending helpful courses, workshops, and travels.

I want to express my sincere and lifelong gratitude to my dear parents, Dr. Abolfazl Shirazian and Farrin Karamati, for their unconditional love, kindness, and support. Also, to my lovely siblings Farnaz and Farzad, for all their help and encouragement. I dedicate this dissertation to my dear family, who is always the wind beneath my wings in my adventurous life, and I am truly blessed to have them.

Table of Contents

List of Figures

1 Introduction

1.1 Introduction

Lifelines like transportation networks, power, water network, etc. are considered assets in societies because they have an essential role in facilitating modern life, their construction is very expensive and time-consuming, and their role is crucial in a disaster aftermath. For example, transportation networks are vulnerable to earthquakes, but they should remain functional to provide access to the affected areas for rescue operations, possible evacuation planning, etc. and any damage to them can have a cascading effect and significantly increase mortality and economic loss. Therefore, it is required to evaluate the seismic vulnerability of transportation networks and estimate the possible loss due to their performance in an earthquake. Accordingly, the decisionmaker can apply risk mitigation strategies like retrofitting, increasing the redundancy of the networks, and allocating budgets to other resilience projects to increase the adaptability of the society for less frequent events.

Stochastic seismic loss modeling is one of the methods in probabilistic seismic risk assessment and its process for the case of the transportation networks is elucidated in a flowchart in Figure 1-1. This assessment is a sequential process that starts with collecting data about the seismicity of the area and the exposure (type and material of bridges, coordinates of the main component of the network, location of important origin-destination of the area). The former is used for generating a synthetic catalog of all possible events and modeling the ground motion for each event and the latter is used for selecting the type of fragility curves, and loss metric. The values of intensity measures at the location of the main components of the network are derived from each developed ground motion map and then the selected fragility curves are applied to estimate the damage to structures due to each event. By having an estimation of the damaged components of the network, network analysis is conducted for pre, and post-event conditions and loss metrics are calculated. This process is repeated for each event in the catalog and consequently, the seismic risk is estimated.

One of the advantages of using stochastic seismic loss modeling is the possibility of including the effect of spatial correlation of ground motions in the calculations. The spatial correlation that is observed in recordings of ground motions from previous earthquakes refers to the similarity in the variability of ground motion between pairs of sites that are affected by an earthquake and this statistical dependency dies down by increasing the distance between the sites (Jayaram & Baker, 2009) (Sokolov & Wenzel, 2013). Previous studies showed that (Crowley & Bommer, 2006) not including the spatial correlation in seismic risk assessment of spatially extended structures might lead to overestimation or underestimation in the results. Moreover, due to the connectivity of the components of the infrastructures, damage to a component can affect the performance of the whole network. Therefore, in loss modeling of infrastructures, it is required to evaluate the effect of an earthquake on all components and estimate their damage to calculate the loss to the functionality of the network due to that earthquake. These advantages, make stochastic seismic loss modeling a suitable method for risk assessment of infrastructures. However, the requirement including many events, and loss modeling of each of them makes this method computationally inefficient. a downside that gets exacerbated by the complexity of loss modeling of infrastructures.

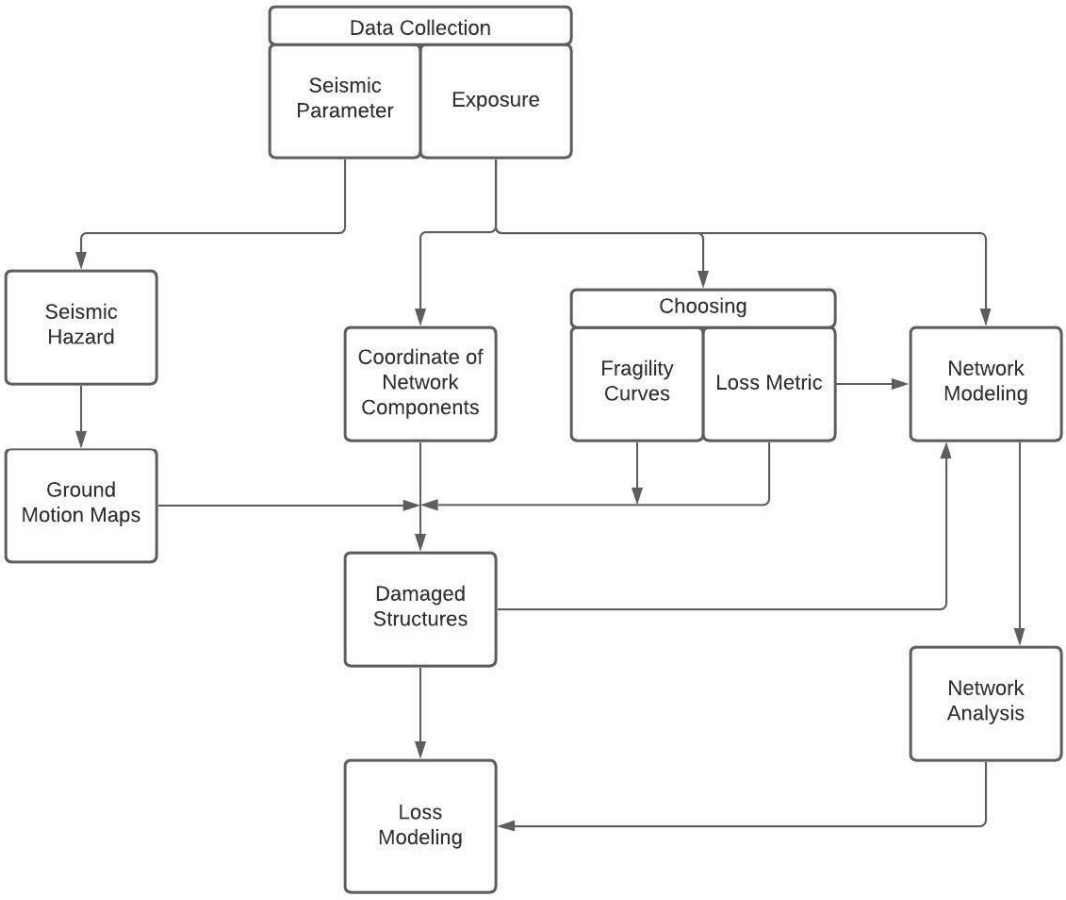


Figure 1-1 Flowchart of stochastic earthquake loss modeling of road network

1.2 Related Research

As mentioned before, the main challenge in stochastic loss modeling is the large number of events that should be considered in the calculation process. Since about two decades ago researchers implemented methods to alleviate this challenge that makes the assessment of spatially distributed systems like lifelines and portfolios computationally expensive. Most of these applied methods were Hazard-Consistent and with the assumption of events that produce similar hazard will have similar consequences. Some of these studies are briefly reviewed in following.

One of the pioneer studies in Hazard-Consistent event reduction methodologies was conducted by Chang et al. 2000 (Chang, Shinozuka, & Moore, 2000) in which 47 earthquakes were selected and their annual occurrence rates were iteratively adjusted to match the derived hazard curve with PSHA for one site (Los Angeles city hall) also produce ground motion levels at 16 track census same with a 475 return period hazard map of the region.

In a study of Kiremidjian et al. 2007 (Kiremidjian, Stergiou , & Lee, 2007) the importance sampling method has been postulated that is a very effective method in improving the simulation of the earthquakes by

applying weighted magnitude bins that result in generating fewer number of less important earthquakes and more of higher magnitudes that mostly cause damage to lifelines.

Lee et al.2008 (Lee & Graf, 2008) applied a Hazard-Consistent methodology in reducing the number of required earthquakes in risk assessment of their case study of a power network and a water network. From a full set of earthquakes in the region, for the case of the power network, they applied 3 steps in selecting their candidate earthquakes. first by defining a threshold for PGA, they eliminated all events that caused fewer PGA levels at the selected grid points in the region which resulted in selecting 79 earthquakes. Second, they eliminated the events that caused PGA levels above the first the threshold in less than two sites. third, by cross correlation of ground motions, they grouped events with similar magnitude and resulted ground motions, eventually 50 candidate earthquakes were selected to match generated hazard curve with the true hazard generated with PSHA. For their power network case of study. Similar method implemented for water network and 59 earthquakes were selected and their annual accordance rates were adjusted in a Hazard-Consistent manner.

Jayaram et al. 2010 (Jayaram & Baker, 2010) applied importance sampling for magnitude also residuals and generated 12500 events and by using K-means clustering in a Hazard-Consistent way (using intensity measure values of interested site in calculating distance metric of clustering) reduced the number of events to 150 events.

Vaziri et al. 2012 (Vaziri, Davidson, Apivatanagul, & Nozick, 2012) applied a Hazard-Consistent optimization method in selecting the reduced set of earthquakes that minimized the error in some selected return periods of hazard curves for all interested sites also errors in magnitude recurrence in sources. By applying this optimization method, they selected 20 earthquakes as their reduced dataset in their case of study.

Han et al (2012) (Han & Davidson. , 2012) proposed some Hazard-Consistent methodologies including sample variability minimization and a contribution factor. The contribution factor which is integrated over all selected sites and all selected returned periods to quantify the contribution of events in the generated hazard to rank the events accordingly and select only those with highest contribution in eventual reduced dataset. They studied a combination of their methods with some of Hazard-Consistent methodologies developed in previous studies such as importance sampling, optimization methods and compared the results. with their applied method for their case of study they could reduce the number of earthquakes from 3000 to 143.

Miller et al (2015) (Miller & Baker, 2015) proposed an optimization methodology to minimize the combination of hazard and risk errors. For the risk error in their optimization methods, they used a computationally inexpensive loss metric such as the ratio of damaged bridges in the network. The resulted reduced dataset later used to generate the risk curve of other system-level loss metrics that relatively demand more computation and their results showed a satisfactory match with their true risk curve.

1.3 Contribution

As mentioned before, despite the advantages of stochastic loss modeling in involving all possible uncertainties and spatial correlation, the downside of this method, especially in case of infrastructures, is being computationally intensive. Therefore, the scope of this research is to reduce the required number of events in stochastic risk assessment. In most of the previous studies, the focus was on developing a

hazard-consistent methodology for reducing the number of events in probabilistic seismic risk assessment by assuming that similar hazards will cause similar consequences. Therefore, the attempt in this thesis is to study the sensitivity of risk to hazard; in other words, how different the resulted hazard estimation of the reduced dataset can become but still develop a satisfactory risk assessment.

In this contribution, by applying clustering algorithms, 4 different approaches will be taken: **1) Risk-Consistent:** the focus is on developing a risk curve with the reduced event sets as similar as possible to the true risk curve (the one which is derived with the whole set of events). **2) Weighted Risk-Hazard**: using a weighted combination of risk and hazard in a way that the risk has more effect in selecting the events than a hazard. **3) Unweighted Risk-Hazard:** using an unweighted combination of risk and hazard in selecting the events means neither risk nor hazard has dominance in the reduction process. **4) Hazard-Consistent:** this method is similar to a part of the reduction methodology applied in Jayaram's study [Jayaram et al.2010] in using K-means clustering, the reduction is implemented in a way that with the reduced dataset a hazard curve similar to (as much as possible) the true one can be generated. All these approaches are implemented on the damaging event set, which means that the full event set will be reduced by a quick initial checking of their effect on the network. In this initial checking step, those events that do not cause damage to at least one bridge are eliminated.

Here, 3 different grouping method is applied to spot similar events. First is the simple spatial griding method which is implemented in the Risk-Consistent approach and selects events according to their location and loss values. The second is the K-means clustering algorithm which was applied by all 4 approaches mentioned above. The third is the density-based clustering which is robust in finding outliers and does not force all the events to be clustered this clustering method is used in the Risk-Consistent approach.

At the end of this research, to explore the potential of the reduced dataset in expediating the further risk assessment of the network (for example for retrofit prioritizations), a hypothetical retrofit scenario for the network is considered. the seismic risk assessment is performed for the retrofitted network, first with the generated reduced dataset in the Risk-Consistent approach (it should be mentioned that the reduced dataset is the one that is a result of the reduction based on the risk results of the pre-retrofit condition of the network). The next time, the risk assessment of the retrofitted network is implemented with the whole dataset of the events. Then the compatibility of the results of the two risk assessments is compared.

The case of study is a transportation network in Erzincan, a city in eastern Turkey which is a highly seismic area in the vicinity of active faults including the North Anatolian fault that has experienced several extreme events with magnitudes around 7 in the last 1000 years.

1.4 Organization of Thesis

Chapter 2:

In this chapter a quick review of the fundamental of risk assessment is implemented and some of the related tools, definitions and metrics are explained. Different methods of seismic hazard assessment and the loss modeling associated to them is briefly discussed with mentioning the pros and cons of each method.

Chapter 3:

In this chapter the modeling and methodology and the contribution of this research is explained and includes Information about the case of study and applied loss metrics and fragility curve and the software for hazard calculation and loss modeling as well. As an example of the results of hazard calculation and modeling the ground motion of some events are presented and the effect of including spatial correlation uncertainty in the modeling is also compared. Some of clustering methods are reviewed, and the applied methodologies of this research which is based of clustering data are illustrated.

Chapter 4:

This chapter includes the results of applying 4 different approaches of this research by using 3 different methods to cluster the events. The results of each methodology are compared for different loss metrics that are applied here. To check the sensitivity of the risk to hazard, the results of risk and hazard for each reduced dataset are compared. At the end, some of the reduced datasets are applied to explore their applicability in further risk assessments.

2 Seismic Risk Assessment

Earthquakes are sudden releases of energy in the form of waves toward the earth`s surface. The release of energy might be because of tectonic movements or volcanic activities. In the case of continuous tectonic movements; for example, when two tectonic plates are converging, cause compression on the surface of their sliding fault that is stored in the form of elastic strain energy. By increasing the compression, the asperities on the fault surface might be broken and release the stored energy in form of heat and wave, and that wave propagates toward the surface and shakes the ground (Kramer, 1996). Magnitude is a scale to measure the released energy as waves and the intensity measures (IMs) are used to quantify the level of ground shaking. The effect of the earthquake on the earth`s surface is not limited to ground shaking, and other effects are possible like surface ruptures, ground failure (like liquefaction, landslide), Tsunami, etc. Transportation networks are mostly vulnerable to ground shakings, surface ruptures, and ground failures. Due to their spatially extended structure, they might be constructed in the vicinity of faults. Sometimes bridges cross over gorges with their columns located on landslide-prone slopes or they might cross over rivers and located on saturated sandy soils that can be susceptible to liquefaction (Kiremidjian, et al., 2007). Damage to the transportation network due to earthquakes has a cascading effect and can cause significant loss to society. But, although earthquakes are inevitable, their consequent loss is not. The seismic risk assessment provides invaluable information for decision-makers to apply risk mitigation strategies and alleviate the adverse effects of earthquakes.

Seismic risk assessment is a multidisciplinary task with the collaboration of seismologists, structural engineers, economists, etc. and it includes estimating intensity measures (IM) associated with each potential earthquake in the region and evaluating the probability of damages to vulnerable exposures using fragility curves (that show probability of exceeding damage versus IMs) and calculation of direct and indirect loss because of damaged structures. Quantifying the complex phenomenon of earthquakes with a few parameters and linking the knowledge between different disciplines requires assumptions and simplifications that increase the uncertainties in the process. The difference between methods of implementing seismic risk assessment is mostly related to how uncertainties are involved in the analysis. In this chapter, first, some of the related parameters and tools in seismic risk assessment are defined, and then methods of seismic hazard analysis and their associated method of loss modeling are briefly explained.

2.1 Definition of related measures and tools

2.1.1 Engineering Intensity Measures (IMs)

To quantify the strength of ground motion and include its effect on structural analysis, engineering intensity measures (IM) such as Peak Ground Acceleration (PGA), Peak Ground Velocity (PGV), Peak Ground Displacement (PGD), and spectral acceleration in certain periods (Sa_T) are applied. These measures are used in the analysis and design of structures for the effect of earthquakes. As mentioned above, ground motion prediction models (GMPMs) are applied to predict the IMs at the desired site or sites.

2.1.2 Ground Motion Prediction Models (GMPMs)

Ground motion prediction models provide a regression model from previous earthquake data to relate certain seismic parameters to the engineering intensity measures such as Peak Ground Acceleration (PGA), Spectral Acceleration at certain periods (Sa),and Peak Ground Velocity (PGV). Regression models inherently include residuals which is also the case in GMPMs. This means that the observed ground motion values can differ from the expected values by the model. These residuals result from uncertainties that are aleatory and epistemic. Aleatory uncertainties are inherent in probabilistic models due to the randomness and epistemic uncertainties are because of insufficient data and lack of knowledge about; for instance, the effective parameters, how to include them in the model, and what is the best form of the empirical model, (for example, for an identical case of earthquake scenario, different ground motions with different GMPEs can be modeled and many of them can be likely) (Douglas, 2010).

To address the aleatory uncertainty, in ground motion prediction models usually two types of uncertainties are considered that are called inter-event and intra-event uncertainties. The former is source related and deals with the uncertainties that are related to differences in the ground shakings between events and it is a constant value in a particular event for the affected sites, and the latter one is related to the uncertainties about the propagation of the energy of the earthquakes to the sites and differs from site to site at a particular event. The spatial correlation is observed in past earthquakes such that sites in a close distance from each other can show same variation in their ground motion values. Spatial correlation can be caused by local geologic conditions including soil conditions and near surface geologic conditions that may change on the scale of kilometers. To involve the spatial correlations in ground shaking calculations, the correlated intra-event residuals are considered in ground motion predictions and the correlation function for these residuals is applied from the available spatial correlation models such as (Jayaram & Baker, 2009). Furthermore, to address the epistemic uncertainties in ground motion predictions, some methods are applicable like using logic trees which are weighted combinations of a few/multiple GMPES, or alternative combinations of a few GMPEs ((Atkinson & Adams, 2013)).

To predict ground motions, the logarithmic mean values are predicted from the applied GMPMs and are added to the residual term which is a product of the standard deviation of the model to the random variables from a standard normal distribution as shown in equation 1:

$$\log y_{ij} = \log \overline{y_{\text{ij}}} + z_{T,ij}\sigma_T \text{ , where } \; z_{T,ij}\sigma_T = z_{E,i}\tau + z_{A,ij}\sigma \qquad (1)$$

In the equation above, y_{ij} is the predicted ground motion value for event i^{th} at site j^{th}, $\overline{y_{\text{ij}}}$ mean value predicted with the empirical equations of the applied GMPMs as a function of parameters like magnitude, distance metric, average shear wave velocity in the upper 30-meter depth from the surface, etc. $z_{T,ij}\sigma_T$ total residual term, $z_{T,ij}$ normalized total residual, σ_T total standard deviation that taken from the applied GMPMs, $z_{E,i}\tau$ inter-event residual term, normalized interevent residual, τ inter-event standard deviation obtained from the applied GMPM, $z_{A,ij}\sigma$ intra-event residual term, $z_{A,ij}$ normalized intra-event residuals, σ intra-event standard deviations obtained from the applied GMPM ((Jayaram & Baker, 2010) (Weatherill, 2014)).

2.1.3 Magnitude recurrence relationship

The relationship between the magnitude of earthquake and their frequency on a fault was studied by (Gutenberg & Richter, 1944) and introduced with the equation (2)

$$\log_{10} \lambda_M = a - bM \quad (2)$$

Where λ_M is the rate of occurrence of earthquakes with magnitudes larger than M, a and b values are empirical parameters and differ for different seismic areas and might change over time with getting more observations of the seismic activity of the area. This equation shows that if for example the b-value be equal to 1, the rate of occurrence of earthquakes with magnitudes greater than a certain magnitude value M_i, will be 10 times more than the rate of occurrence of earthquakes with magnitudes greater than M_{i+1}. For example, for b value equal to 1, earthquakes with magnitude greater than 5 will occur 10 times more often than earthquakes with magnitudes greater than 6.

2.1.4 Seismic Hazard and seismic risk

Seismic hazard is about a potential earthquake in a region and all the natural phenomena that are related to it such as ground motion, landslide, and liquefaction that have the potential of causing loss to vulnerable exposures (Wang, 2011). Seismic risk is the statistical likelihood that the vulnerable exposures in a region experience that loss due to the seismic hazard (U.S. Geological Survey, 2022). Through this thesis, the annual exceedance rate of a certain level of intensity measure at a specific site is referred to as hazard and the annual rate of exceedance of a certain level of loss due to the potential earthquakes in the region is referred to as risk.

2.1.5 Hazard Curve and Hazard Map

Hazard curve is a result of probabilistic seismic hazard analysis and presents the annual exceedance rate (or return period) of a certain level of intensity measure at a specific site. A hazard map, for a certain return period (like 100, 475, 975, or 2475 years), is derived from hazard curves of several sites that are spatially extended in the area (like grids). IM values for the related return period are obtained from site specific hazard curves and interpolated over the region to develop the hazard map that represents IM values for a specific return period in a spatial manner. These maps are applied in different disciplines such as structural engineering, and insurance.

2.1.6 Secondary seismic hazards

Following the earthquake ground shakings, several secondary hazards are possible such as surface rupture ground failure (like liquefaction, and landslide), tsunami, seiche, flood, and fire.

2.1.7 Fragility Curve

Fragility analyses are performed to estimate the seismic performance of structures conditional to intensity measure the results of this analysis represent the probability of exceedance a certain level of damage for a given intensity measure value (IM). Fragility analyses are conducted in two types of empirical and analytical based on the resource of their data. Empirical Fragility Analysis is based on engineering judgments during inspections of real damage to the structures due to an earthquake. Although the results of these types of fragility analyses might be accurate, they are subjective to the experience of the expert that generates them and because of lack of data they are applicable to limited areas. Analytical Fragility Analysis is based on computational analysis such as nonlinear time history, nonlinear static, or elastic spectra analysis and the derived seismic response of the modeling of structures. Previous studies show satisfactory consistency between the results of the analytical approach and the empirical ones (Shinozuka, Feng, & Lee, 2000).

Fragility curve, as the result of fragility analysis, is one of the key tools in seismic loss modeling of structures as it shows the probability of exceedance of a damage level in a structure for a given ground motion intensity measure. Damage levels are predefined classification and are usually defined as Slight, Moderate, Extensive and Complete. In Table 2-1 the HAZUS definition for damage limit state for concrete bridges is exactly quoted from (HAZUS, 2020).

In analytical fragility, the performance of the structure is evaluated based on applied Engineering Demand Parameters (EDP) (like column ductility) and a probabilistic model of seismic demand conditional to the intensity measure is developed. The probability of reaching or exceeding of seismic demand D from a specific damage level (known as capacity of the structure C) conditional to an intensity measure (IM) is defined as equation (3)

$$\text{Fragility} = P(D \geq C|IM) \tag{3}$$

By assuming lognormal distribution for the seismic demand and structural capacity models, the fragility equation can be written as equation (4)

$$Fragility = P(D \geq C|IM) = \Phi\left(\frac{\ln S_d - \ln S_c}{\sqrt{\beta_d^2 + \beta_c^2}}\right) \tag{4}$$

Where Φ is the standard normal distribution function, S_d is the median of the seismic demand and is a function of IM and S_c is the median of the seismic capacity, β_d is the logarithmic standard deviation of seismic demand and β_c is the associated logarithmic standard deviation of seismic capacity of the structure (Nielson, 2005) (Shirazian, Ghayamghamian, & Nouri, 2011).

Table 2-1 Damage classification for bridges based on HAZUS (HAZUS, 2020)

Level of damage	Descriptions
Slight	"Is defined by minor cracking and spalling to the abutment, cracks in shear keys at abutments, minor spalling and cracks at hinges, minor, spalling at the column (damage requires no more than cosmetic repair) or minor cracking to the deck."
Moderate	"Is defined by any column experiencing moderate (shear cracks) cracking and spalling (column structurally still sound), moderate movement of the abutment (<2"), extensive cracking and spalling of shear keys, any connection having cracked shear keys or bent bolts, keeper bar failure without unseating, rocker bearing failure or moderate settlement of the approach."
Extensive	"Is defined by any column degrading without collapse – shear failure - (column structurally unsafe), significant residual movement at connections, or major settlement approach, vertical offset of the abutment, differential settlement at connections, shear key failure at abutments."
Complete	"Is defined by any column collapsing and connection losing all bearing support, which may lead to imminent deck collapse, tilting of substructure due to foundation failure."

2.1.8 Loss metrics

Loss metric is used to quantify the loss of an exposure and evaluate the possible consequence of an event; in other words, Loss metric specifies the loss from the user perspective. Can be calculated based of the comparison of two pre-event and post-event conditions of the exposure and usually are measured in the form of these two states difference or their ratio. Based on the type of exposure different loss metrics are applicable. For example, in case of transportation network the performance can be quantified in different levels like component level and system level (functionality level). Tin the component level, loss metrics can be related to the direct physical damage to the structures of the components of the network like

damage to bridges as a main component of the network. The system level loss metric is related to the damage to the functionality of the network and the features that are desired to remain functional aftermath like connectivity or travel time in transportation network.

The loss metrics of component-level can also be combined with some of the system-level metrics to derive a total loss to the network. For example, in case of transportation network, the direct damage to the bridges can be converted to financial loss by having the loss ratio of the bridge (cost of repair to the replacement) and be added to the economic loss associated with the total time delay (as one of the system- level loss metrics) due to damaged network (by having an estimation of the value of the time) to calculate the total economic loss to the network due to an event. In the following of this section, some of the applied system-level loss metrics in the loss modeling of transportation networks are explained.

2.1.8.1 Simple Connectivity Loss (SCL)

This metric is one of the system level metrics which shows the damage in functionality of a network and is applied in loss modeling of transportation and power and water network as well (Franchin & Cavalieri, 2013), (Kyriazis, Franchin, Khazai, & Wenzel, 2014), (Poljanšek, Bono, & Gutiérrez, 2012). Requires a set of predefined origin-destination (or source-sink) pairs, by comparing the post-event with the prevent condition measures the average loss connectivity in the network and is defined by the equation 5:

$$SCL = 1 - \frac{\sum\left(\frac{N_{j,s}}{N_{j,0}}\right)}{n} \quad (5)$$

Where $N_{j,0}$ and $N_{j,s}$ are the number of sources (origins) connected to the jth Sink (destination) in normal condition and seismic condition respectively. n is the number of sinks (destinations).

2.1.8.2 Weighted Connectivity Loss (WCL)

This metric, like SCL, is a system level metric but is a weighted version for SCL. This means that it penalizes the rerouted connections. Unlike the SCL that is just affected by lost connections, this metric also involves the effect of rerouting in the measurement of loss to the network and this can be a better evaluation of loss to the network than the simple connectivity loss. WCL is defined by equation 6:

$$WCL = 1 - \frac{\sum\frac{W_{j,s}N_{j,s}}{W_{j,0}N_{j,0}}}{n} \quad (6)$$

Where like SCL, $N_{j,0}$ and $N_{j,s}$ are the number of sources (origins) that are connected to the jth Sink (destination) in normal condition (pre-event) and seismic condition (post-event) respectively. n is the number of sinks (destinations). $W_{,0}$ and $W_{j,s}$ are the weight of the path to the jth sink over all the connected sources in normal and seismic condition respectively. For road networks, the W (weight) can be defined as the inverse of number of edges from the source (origin) to the sink (destination) for example, $W_{j,s} = \sum_{i=1}^{N_{j,0}} I_{ji,s} \frac{1}{N_{ji,s}}$ and in same way for the weight in normal condition. (Franchin & Cavalieri, 2013).

In the next chapter, the results of two SCL and WCL loss metrics are compared by using an example of a simple graph.

2.1.8.3 Total Time Delay (TD)

One of the system-level losses in transportation network that might be of concern of decision-makers is the increase in travel time in the network due to the damaged components. Delay also can be converted to the economic loss by having an estimation for the time value for the uses of the network, fuel price, delay cost in commercial shipments, etc. Driver`s delay is a commonly used metric in measuring loss in functionality of a transportation network that is proposed by (Shinozuka, Murachi, Dong, & Zhou, 2003) and is calculated by equation (7). This metrics is based on a redundant transportation network sample; therefore, it assumes even in the extensive damage levels, an O-D connecting path is still possible through other roads like secondary or residential. Therefore, if a connection loses a bridge, the connection is still available in the network but slower and with a lower capacity in connecting paths.

Total delay travel time (TT) =$\sum_a x_a t_a(x_a)$ (7)

Driver`s Delay (DD) =$\sum_a x_a^{`} t_a^{`}(x_a^{`}) - \sum_a x_a t_a(x_a)$ (8)

Travel time at flow x_a: $t_a = t_a^0\left[1 + \alpha\left(\frac{x_a}{c_a}\right)^{\beta}\right]$ (9)

Where in equations above (7,8,9): x_a is flow at a link (in passengers per day), t_a is the travel time at flow x_a and is calculated by equation 9, where t_a^0 denotes the travel time in zero flow on link a (equals to the length of the link over its speed limit), c_a is the practical capacity of link a. α and β are empirical parameters that are usually taken as 0.15 and 4 respectively. Based on equation 9, if traffic flow in a link be equal to its practical capacity, the time of travel in that link will be 1.15 times of travel time in zero flow (Shinozuka, Murachi, Dong, & Zhou, 2003). The primed parameters in equation 6 denote the post-event condition means by subtracting TT in normal condition from TT in post-event, the driver`s delay is calculated.

2.1.9 Loss exceedance curve

Loss Exceedance Curve (LEC) is a result of probabilistic seismic risk assessment and presents the annual exceedance rates versus a range of values of loss related to the applied loss metric for the exposure and the exposure can be a single site, portfolio, or an infrastructure. For some reasons like the limited average lifespan of structures or making the results more understandable for other disciplines, usually, LEC is represented as the probability of exceedance of at least once in a time window (like 50 years). LEC is applied in risk mitigation planning and insurance companies. For example, if according to the derived loss exceedance curve for a structure, the probability of exceeding $100k economic loss in 50 years is 10%, then the stakeholder/decision-maker can decide whether this risk is acceptable or not. If the risk is too much to take, he/she can apply some risk mitigation strategies (like retrofitting) or can buy insurance. On the other, hand, the insurance companies, can use this information to estimate the insurance premium for the insurance coverage plan that they what to offer (Grossi, Kunreuther, & Windeler, 2005). The methods of generating the loss exceedance curve will be explained later in this chapter.

2.2 Seismic Hazard Analysis (SHA)

Seismic hazard analysis is the process of applying GMPMS to predict or model the level of ground motions at a site of sites of interest to be applied in loss modeling. Based on how uncertainties get involved in the calculation, seismic hazard analysis methods are classified as deterministic (scenario-based) and probabilistic methods.

2.2.1 Deterministic SHA (DSHA)

In the deterministic approach, also called the scenario-based, the seismic hazard is analyzed for one earthquake with a specific magnitude, location, and all other related parameters. The ground motion map is predicted with applying Ground Motion Prediction Equations. The loss modeling is implemented based on the derived ground motion map. The value of the intensity measure per location of the exposure is derived from the map and by using the pertinent fragility curves the probability of exceedance of damage levels associated with the predicted IM values are estimated. The loss to the exposure is then modeled according to the estimated damage and by calculating the applied loss metric.

This method is simple and fast in the calculation, provides an evaluation of the consequences of an earthquake scenario, and the output can be easily understood by stakeholders. But the downsides of this method in comparison with the probabilistic methods are that its result is subjective and limited to the selected earthquake scenario. Any decision (for risk mitigation or insurance etc.) based on the results of one earthquake scenario might be inadequate in case of occurrences of another scenario (Vaziri, 2009). Moreover, this method does not include rates (return periods) and hazard curves, and loss exceedance curves cannot be generated by this method.

2.2.2 Probabilistic SHA (PSHA)

In seismic hazard analysis, linking knowledge of earth science, which inherently includes estimations and approximations, between different disciplines like geology, seismology, and structural engineering needs simplifications and assumptions that cause uncertainties in the process. The simplifications that are applied to interpret the effect of an earthquake phenomenon on the area by intensity measures which are estimated by empirical equations that are functions of a few parameters like magnitudes and distance and soil condition etc. Even using a single value of Magnitude to scale earthquakes is a massive simplification and inherently a source of uncertainty. Not including uncertainties effect can cause underestimation or overestimation in hazard and loss modeling. The probabilistic methods attempt to include the effect of all possible earthquakes in the calculation and cover most of the uncertainties related to the earthquake reoccurrence model, source location, ground motion predictions, etc. Probabilistic methods, unlike the deterministic method, present the hazard and loss by their annual rate of exceedance (or return period). Two common methods for PSHA are classical and Monte Carlo (event based) PSHA.

2.2.2.1 Classical PSHA

In this method probabilistic method that was introduced by (Cornell, 1968), by using the total probability theory, the effect of all possible earthquakes with different magnitudes, probability of occurrence, source location and ground motion is included in prediction of the annual rate of exceedance of a given level of IM at a site of interest which is calculated with equation (10):

$$\lambda(IM > x) = \sum_{i=1}^{N_{Sources}} \nu(M_i > m_{min}) \int_{m_{min}}^{m_{max}} \int_{0}^{r_{max}} P(IM > x|m,r)\, f_{M_i}(m)\, f_{R_i}(r)\, dr\, dm \qquad (10)$$

Where ν is the rate of occurrence of earthquakes greater than m_{min} and $N_{Sources}$ is the number of earthquake sources and λ is the rate of occurrence of intensity measures greater than x in using a GMPE for ground motion modeling. $f_{Mi}(m)$ is the probability density function (PDF) for magnitude, $f_{Ri}(r)$ is the PDF for source to site distance (Baker, 2013).

Poisson models are commonly used in probabilistic seismic hazard analysis to model the probabilistic occurrence of earthquakes with the assumption of a random sequence for earthquakes (which means

that events occur independent of when, where, and how big was its previous event.) (Anagnos & Kiremidjian, 1988). Therefore, conditional on the assumption of the Poisson model for the occurrence of earthquakes, the probability that a certain level of IM at a site of interest is exceeded at least once within a given time window of t (for example 50 years) is calculated with equation (11) (Baker, 2013)

$$P(IM > x|t) = 1 - e^{-\lambda} \quad (11)$$

Where t is the window of time (for example 50 years), λ the mean annual rate of exceedance of the given IM.

In the case of risk assessment of a single site, the results of hazard calculation with classical PSHA as the exceedance rate of a ground motion level are applied to estimate the resulted loss using the fragility functions. But in the case of spatially extended structures like lifelines that the damage to a single site (as a component of the network) affects the performance of the whole network, this method is not applicable. Because the IM values for the same annual rate of exceedance in different sites is not necessarily related to the same earthquakes; therefore, cannot be applied in modeling the loss in the performance of the network (Vaziri, 2009). Moreover, the spatial correlation of ground motion can affect the results of loss modeling of multiple sites and ignoring it in the calculation can overestimate or underestimate the risk as shown in the study of (Crowley & Bommer, 2006). In Classical PSHA, the spatial correlation of intra event residuals is not included in the calculation, Therefore, this method would not be a good choice in the case of risk assessment of portfolios and infrastructures.

2.2.2.2 Monte Carlo (Event based PSHA)

This Monte Calo method (also called event based PSHA) is implemented in some different ways like using a historical catalogue, ore selecting the most credible events, or using a synthetic catalogue. Here the focus is on the conventional Monte Carlo method that generates a synthetic catalogue for an associated duration of T. By applying Monte Carlo sampling, all possible earthquakes with different magnitudes and on predefined fault sources or area sources are generated and develop a set of stochastic events in a way that represents the seismicity of area. Based on the law of large numbers, the longer associated duration of catalog is required, in ranges of hundreds of thousands, to get more realistic results and this requirement makes the Monte Calo method computationally inefficient.

Unlike the Classical PSHA method, in the ground motion modeling for each event in this Monte Carlo approach, spatial correlation of ground motion also can be modeled. This feature is one of the reasons that makes this method useful in risk assessments of spatially extended structures like portfolios and infrastructures. The hazard curve for a specific site is calculated by first sorting all the derived IMs at that site from each event, then counting the number of exceeding values on the list for each level of IM with including the effect of the annual occurrence rate of each event which is once in the duration of the stochastic event sets (Musson 2000). The annual rate of exceedance of IM=x at a site is calculated by equation (12):

$$\lambda(IM > x) = \frac{\sum_{i=1}^{n} \mathrm{H}(IM_i - x)}{T} \quad (12)$$

Where T is the duration associated with the stochastic event set and 1/T represents v as the occurrence rate of each event in the stochastic event set, and H is the heavy side function which is 1 when the intensity measure of the i_{th} event exceeds the given intensity measured value of x and zero otherwise, n is the

number of events in the stochastic event set. Like the classical PSHA, here also by the assumption of the Poisson distribution for the occurrence of the events, the probability of a certain level of IM, at a site of interest, is exceeded at least once within a given time window (for example 50 years) can be calculated by equation (11) (Ebel & Kafka, 1999).

In stochastic loss modeling, which is performed based on this event based PSHA, events are evaluated individually. Using the ground motion map of the event and the exposure related fragility curves, damage to the structures is estimated. then according to the applied loss metric, the loss values are calculated. Direct physical loss can be calculated for the structure from the estimated damage level; for example, by using loss ratios that relate the level of structural damage to the repair or replacement cost and by summing up the direct loss for all structures in the area, the total direct loss due to each event is derived. In the case of dependent exposures like a network, not only loss to the single structure is required to be evaluated but also the effect of that damage to the performance of the network is a matter of concern. Therefore, loss metrics that are related to the performance of the network (like connectivity or flow) are calculated by network analysis for the damaged network. In this method, the loss exceedance curve is generated based on the calculated loss values for each event in the stochastic event set. Whether the exposure is a single site or the network, the calculated loss values for each event are compared and for one certain loss values, the number of the exceedance of that loss values among event set is measured and the related exceedance rate is calculated with equation (13).

$$\lambda(l > L) = \frac{\sum_{i=1}^{n} \mathrm{H}(l_i - L)}{T} \qquad (13)$$

Where T is the duration associated with the stochastic event set and 1/T represents v as the occurrence rate of each event in the stochastic event set, and H is the heavy side function which is 1 when the loss value of i_{th} event exceeds the given loss value of x and zero otherwise, n is the number of events in the stochastic event set.

Some of the advantages of the Monte Carlo method are: 1) powerful in including uncertainties in modeling, 2) adaptable to also simulate a non-poison distribution of earthquakes (Musson, 2000), 3) the loss modelling implemented per event, therefore, resulted loss due to each earthquake is tractable which makes the output more understandable to people from other discipline and this item is valuable for the multidisciplinary task of risk mitigation decision makings. But the disadvantage of this method is the requirement of having many events to get more realistic hazard and risk results and these needs considering longer associated durations (like hundreds of thousands or million years) to include more realizations of earthquakes in the synthetic catalog and their consequent loss. This requirement makes the process of risk assessment computationally inefficient, especially in the case of infrastructures that the complexity of their loss modeling exacerbates this issue.

3 Modeling and Methodology

3.1 Introduction

In this chapter the case of study is introduced, then the implemented modeling, the applied software for the hazard calculations, and the loss modeling is mentioned and finally, the proposed methodologies of this research will be explained. As it was elucidated in a flowchart in the first chapter, the loss modeling is a sequential process. Therefore, research is started by collecting data about the exposure and the seismicity of the area. After collecting related data, the seismic hazard is analyzed by using OpenQuake software and the ground motion maps for each stochastic event are generated. Then the transportation network case of study is modeled by a digraph in MATLAB programing language and for each event, the ground motion map is interpolated to the location of bridges to get the intensity measure values at the location of bridges which is considered the main vulnerable components of the network. Having the fragility curve selected, the probability of exceedance of the extensive damage level for each bridge due to an event can be estimated. To assign the damage to the bridge, a Monto Carlo based method is applied that has been proposed in a study by (Poljanšek, Bono, & Gutiérrez, 2012); a random value between 0 and 1 is generated and compared with the derived value for the probability of exceedance of damage for the bridge that is derived from the fragility curve, If the value of the random number is lower than the derived probability value, the damage is assigned to the bridge. By this monte Carlo damage assignment method, the associated uncertainties such as assigning fragility curve to the structure can be considered that resulted in estimation of the performance of structures and loss modeling. The damaged bridges are eliminated from the network model (here digraph) and the network is analyzed for cases before and after the event to calculate the required parameters for the applied loss metrics. Except for the part of network analysis for the pre-event (normal) condition, the rest of the mentioned process is repeated for every single event in the generated stochastic seismic event set.

The high number of required events in stochastic (event based) loss modeling makes this calculation computationally intensive and in the case of infrastructure, the network analysis exacerbates this problem. Previously, some Hazard-Consistent methods have been proposed by researchers(such as (Chang, Shinozuka, & Moore, 2000) (Kiremidjian, Stergiou , & Lee, 2007) (Lee & Graf, 2008) (Vaziri, 2009) (Jayaram & Baker, 2010) (Han & Davidson. , 2012)) that the events eliminated in a way that a reduced dataset produces similar hazard results assuming that the eliminated events will produce similar loss to those which have been used as their representatives. To reduce the number of required events in stochastic seismic risk modeling, in this research some methodologies are applied to include the effect of risk in the data reduction process. Also, the sensitivity of the risk to hazard is evaluated to see how different a hazard estimation with the reduced dataset can become but still develop a satisfactory risk result.

The main structure of the reduction method in this research is to group the more similar events (similarity can be regarding to risk, hazard, or both) and choose one representative event from each group (cluster). Then sum up the occurrence rate of all events spotted in a cluster and use it as a new occurrence rate to the representative of that cluster. Doing this for each cluster, consequently, a reduced dataset is made such that will produce a similar loss exceedance curve to the original whole dataset. 3 different methods of grouping similar events will be applied. First is a simple spatial griding method inspired by grid-based clustering algorithm and it will be conducted in a Risk-Consistent manner. The size of grinding affect and

loss related layers will affect the size of resulted reduced dataset. The Second method applies the K-means clustering algorithm in clustering the similar events and this method has been implemented for different approaches as Risk-Consistent, Weighted Risk-Hazard, Unweighted Risk-Hazard and finally a Hazard-Consistent approach and their results will be compared. In the method of using K-means clustering, also two criteria in deciding about the optimum number of clusters will be used which gives practically give a hint about the optimum number of events in a reduced dataset. The third method in this research, applies the density-based clustering algorithm to apply the ability of this clustering method in dealing with outliers in the reduction procedure. Here, the density-based clustering will implement just for the Risk-Consistent approach. The details of each mentioned step are explained in this chapter.

3.2 Case of study and exposure

The case of study is Erzincan, a city in eastern Turkey, located in a highly seismic area within the vicinity of three active faults such as North Anatolian Fault, Northeast Anatolian Fault, and Ovacik fault, the line sources around Erzincan are shown in Figure 3-1 Faults in Erzincan Area. Over 1000 years ago this area experienced several earthquakes around magnitude 7 (Ambraseys, 2009). Just in the last century two extreme events occurred in the area, one in 1939 with a magnitude of 7.8 and caused significant damage and more than 30000 fatalities. The other one was in 1992 with a magnitude of 6.8 caused about 3-5 million dollars economic loss and half a million fatalities (Askan, et al.). According to the global seismic hazard map generated by GEM foundation, for a return period of 475 years, the PGA in Erzincan is more than 0.5 g. (Pagani, et al., 2018)

All the seismic parameters of the fault have been obtained from the Earthquake Model of the Middle East (EMME) Project (Danciu, Şeşetyan, Demircioglu, Erdik, & Giardini, 2016), and the selected GMPE logic tree is the same as the one that has been used in EMME project for the active shallow crust tectonic type including 4 weighted branches of GMPEs (including GMPEs developed by (Akkar, Sandıkkaya, & Bommer, 2014), (Akkar & Çağnan, 2010), (Chiou & ,Youngs. , 2008), (Zhao, et al., 2006)). The exposure data, maps, road network, faults shapefiles are collected from OpenStreetMap (www.openstreetmap.org/copyright). In modeling, mostly the primary and important roads of the city are modeled, and the rest of the road types are not included for sake of simplicity. It should be mentioned that the real seismic risk assessment of Erzincan is out of the scope of this study, and it has been modeled just as a case of a seismic area in testing our proposed methodology.

3.3 Hazard Calculations

By having seismic source information like tectonic type (shallow-active, subduction zone, ...), coordinate and geometry (area source fault or point source), and seismic parameters of seismic sources in the area, hazard can be calculated. As mentioned before, only faults (more than 50) have been used here and the area sources have not been included in the hazard modeling of this research, but this will not affect the methodologies that is going to be studied in this research because using the area source would just increase the number of events. In this research, an event based PSHA is conducted because as mentioned in chapter 2, Event-based PSHA is a probabilistic hazard calculation method that allows aleatory and epistemic uncertainties to be involved in the hazard modeling. The aleatory uncertainties as inter-event and intra-event residuals are considered in ground motion prediction modeling. Unlike classical PSHA, as the other probabilistic hazard method, in event based PSHA it is possible to consider the spatial correlation

phenomenon of ground shaking in our modeling. Moreover, the risk assessment based on this hazard calculation method provides a good vision of loss caused by each event.

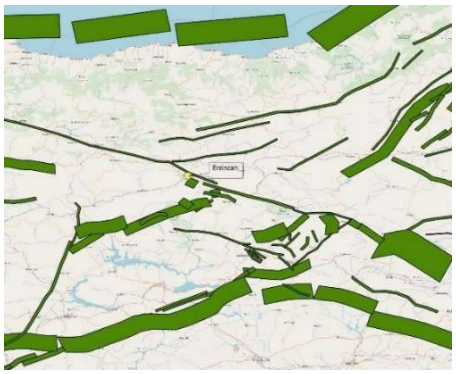


Figure 3-1 Faults in Erzincan Area. (Map data source: OpenStreetMap contributors, https://www.openstreetmap.org)

3.3.1 Software

The hazard calculations of this study have been performed by OpenQuake (OQ) A popular open-source earthquake hazard and risk modeling software written in Python programming language that has been developed by the Global Earthquake Model Foundation (GEM, 2022) and launched in 2012 (https://platform.openquake.org). During the past decade, OQ has been used in several international collaborative projects, in developing some official national seismic hazard maps, and in many scientific studies as well. This software can perform deterministic and probabilistic hazard calculations (classical and event base PSHA) and including all the uncertainties is possible in calculations. since the coding is open, it can be tested by other scientists and users, and it is updated frequently. These features make OQ a powerful and reliable software in hazard and risk assessments. Version 3.11.3 of this software is used in this research.

3.3.2 Modeling in OpenQuake: Event-based PSHA

To perform an event based probabilistic hazard calculation in OpenQuake, the required inputs are source model, source model logic tree, logic tree of GMPE, and a configuration file. Source model includes the geometry, tectonic type and other seismic parameters related to the seismic source. As mentioned before, the seismic data of this study has been derived from the EMME project input file that was a classical PSHA performed with OpenQuake. From the seismic data 52 faults in the interested region were selected and area sources are not included in this the hazard modeling as was explained before. The GMPE logic tree is required to define the GMPEs as an input to the software to be used in generating ground motions. The ground motion logic tree that is applied in hazard calculation is also the one that has been used in the Earthquake Model of Middle East (EMME) project for the active shallow crust tectonic region and involves 4 GMPEs (including GMPES generated by (Akkar, Sandıkkaya, & Bommer, 2014) , (Akkar & Çağnan, 2010), (Chiou & ,Youngs. , 2008), (Zhao, et al., 2006)). The configuration file involves information required for the hazard calculations such as the geometry of the region, local site condition, stochastic event set, investigation time, the spatial correlation model, and the hazard output. The spatial correlation model that is used in the hazard analysis of this research is a model developed by (Jayaram & Baker, 2009). For the site condition, a uniform soil type condition of rock (vs30=800m/s) for the area is considered.

Consequently, a stochastic event set with a duration of 200000 years is generated with OpenQuake including over 300000 earthquake events (ground motion maps).

By having the ground motion maps generated for each event in our stochastic event set, the loss modeling can be performed per each event. But first generated the hazard curve with the morel is checked. As mentioned in chapter 2, in event based PSHA, the hazard curve is calculated from ground motion values per site. By getting all values of PGA per each event at a site, in the stochastic event set with a duration of 200000 years, the annual exceedance curve can be calculated by equation 4 for a range of PGA:

$$\nu = \frac{\sum_{i=1}^{I} H(IM_i - IM)}{T} \quad (4)$$

Where ν is the rate of exceedance of a given IM, H is a Heaviside function (will be zero when IM_i is less than IM, otherwise equals to one), the I is the total number of events in the event set, T is the time the associated with the event set (200000 years). The generated hazard curve for a site in Erzincan is shown in Figure 3-2. The resulted curve shows the exceedance rate for PGA= 0.5 g is about 0.002, which means a return period of 425 years which is similar to the result of the official hazard map of the area for the return period of 475 that has been developed by classical PSHA (as mentioned in the section of the case of study in this chapter).

The curve of Figure 3-2 is not smooth for a longer return periods and longer duration of catalogs including a greater number of events will improve the hazard results. But considering a duration longer than 200000 (like 500000 or even 1000000 years) does not affect the methodology and scope of this research and would make the computation more intensive.

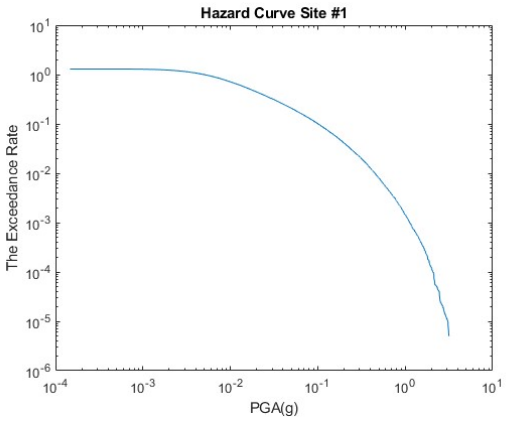


Figure 3-2 Generated hazard curve for a site in Erzincan with our Event-based hazard modeling in OQ

3.3.2.1 Ground motion prediction modeling in OQ

As mentioned in chapter 2, the uncertainties in predicting ground shakings are two types epistemic and aleatory. The epistemic is due to the lack of data and knowledge (for example unknown effective parameters in the model or what is the best form of the empirical model) and the aleatory is inherent in regression modeling due to randomness and simplifications and assumptions associated with the empirical modeling, (for example simplifications in scaling a complicated phenomenon like an earthquake with a simple parameter as Magnitude, relating the ground shaking to some few parameters like magnitude and distance, etc.) To consider epistemic uncertainties in hazard modeling, GMPE logic trees

are applied in hazard modeling. The aleatory uncertainties can be included in the ground motion calculation as inter-event and intra-event term. Inter-event is for the uncertainties between events means that similar events at the same area can produce different ground shaking but this uncertainty is constant for all sites in an event, (means that it is a source related uncertainty and all sites are experiencing a higher or lower variation in comparison with their shaking in other events) (Figure 3-4 c). The intra event term is about the variations in ground shaking within a particular event, which means that sites in the same soil condition and distance from the source can experience different shakings due to the same event. Normally, these intra-event variations are spatially correlated, and this correlation dies down by increasing the distance between sites, this distance is referred to as the correlation range. the reason for this correlation can be related to near surface local geologic conditions such as soil conditions that may change on the scale of kilometers. This correlation phenomenon can be included in the calculation by using spatially correlated intra-event term in the ground motion simulations. In following, it will be briefly discussed about how the aleatory uncertainties and the spatial correlation is included in hazard calculation with OpenQuake, for more detailed information see the OpenQuake ground motion toolkit and hazard book as well ((Weatherill), (Pagani, Monelli, Weatherill, & Garcia, 2014-08)):

Usually, in ground motion prediction, the value of intensity measure (IM) of interest due to event i at jth site is calculated with an equation like equation 1 which includes a combination of the logarithmic mean value of IM that is predicted from the applied Ground motion prediction model (GMPM) and the residual term which is a product of the standard deviation (defined in the GMPM) to the random variables from a standard normal distribution:

$$\log y_{ij} = \log \overline{y_{\text{ij}}} + z_{T,ij}\sigma_T \text{ , where } \quad z_{T,ij}\sigma_T = z_{E,i}\tau + z_{A,ij}\sigma \text{ , and } \sigma_T = \sqrt{\tau^2 + \sigma^2} \qquad (1)$$

In the equation above, y_{ij} is the predicted ground motion value for event ith at site jth, $\overline{y_{\text{ij}}}$ median intensity measure that is predicted with the empirical equations of the applied GMPEs as a function of parameters such as magnitude, distance metric, average shear wave velocity in the upper 30-meter depth from the surface known as Vs30. $z_{T,ij}\sigma_T$ is the total residual term, $z_{T,ij}$ is the normalized total residual, σ_T is the total standard deviation that is taken from the applied ground motion prediction model (GMPM). $z_{E,i}\tau$, is the inter-event residual term where $z_{E,i}$ is the normalized Inter-Event residual, τ is the inter-event standard deviation obtained from the applied GMPM. $z_{A,ij}\sigma$ is the intra-event residual term where $z_{A,ij}$ is the normalized intra-event residuals, and σ is the intra-event standard deviations obtained from the applied GMPE ((Weatherill, 2014) (Jayaram & Baker, 2009))

Same way in OpenQuake to generate ground motion maps for an area with N sites per each event in our generated stochastic data set, ground motion is simulated by the sum of three terms of logarithmic mean of ground motion, inter event residual term and intra event residual term. The inter event term is a random selection from a univariant normal distribution with zero mean and standard deviation of τ (this term is equal for all sites per event). The intra event term is sampled randomly from a multivariant normal distribution with a zero mean and a covariance matrix of Σ (as shown below), in case of not considering spatial correlation, this matrix would be a diagonal matrix (with zero off-diagonal elements) (Pagani, Monelli, Weatherill, & Garcia, 2014-08):

$$\Sigma = \begin{bmatrix} \delta_1^2 & \delta_1\delta_2\rho_{12} & \cdots & \delta_1\delta_N\rho_{1N} \\ \delta_2\delta_1\rho_{21} & \delta_2^2 & \cdots & \delta_2\delta_N\rho_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_N\delta_1\rho_{N1} & \delta_N\delta_2\rho_{N2} & \cdots & \delta_N^2 \end{bmatrix} \quad (2)$$

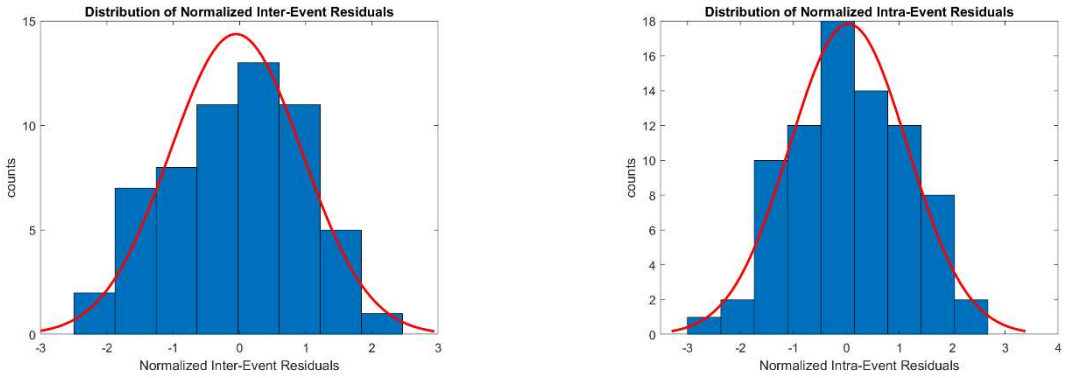
The correlation model that is used in the modeling of this study and the one that is available in OpenQuake as well is the model generated by (Jayaram & Baker, 2009) which is an empirical model based on data from previous earthquakes and is defined by equation 3:

$$\rho = e^{\frac{-3}{b}} \quad (3)$$

Where ρ is the correlation between two sites, h is the separation distance of the sites, and b is the range of correlation defined in the model as a function of the period of spectral acceleration Sa_T. In the case of PGA, the period is considered zero (Booth, 2007). Based on the spatial correlation model of Jayaram et al. 2009, for the case modeling in our study with the PGA as the applied intensity measure and considering the uniform vs30 of rock for the whole area, the expected distance range of correlation is about 40 km and after this distance, the correlation decreases.

To test the residuals in the generated ground motion maps by OpenQuake, the derived PGA values are decomposed by using equation 1. First To get the ground motion values, the truncation level in the configuration file of OpenQuake has been assigned equal to 3 so the normalized residuals for inter- event ($z_{E,i}$) and intra-event ($z_{A,ij}$) uncertainties, are selected randomly from a standard normal distribution within the 99.7% probability. The inter-event standard deviation and the normalized inter-event residuals per event is obtained directly from the OpenQuake outputs. The Figure 3-3 (a) shows the distribution of the applied normalized inter event residuals for these 58 events which can be fitted by a standard normal distribution. These 58 events are from the same source, with the same magnitude and the same applied GPME (Akkar & Çağnan, 2010). To calculate the interevent residuals the median values of PGAs are needed therefore the previous run was repeated but this time by assigning the truncation level equal to zero.

Figure 3-3 (b) shows the distribution of the normalized intra-event residuals between 79 sites of interest in our study for one of the mentioned 58 events. These values are derived by using equation 1 knowing the $y_{ij,}$ $\overline{y_{ij}}$, $z_{E,i}$ and having the standard deviation of inter and intra event residuals (τ and σ) from the applied ground motion prediction model which is Akkar et al. 2010 and are equal to 0.5163 and 0.6527 respectively. As shown in this figure, the distribution of applied normalized intra-events residuals in generated ground motion maps by OQ also can be fitted by a standard normal distribution.



a) Normalized Inter-event residuals b) Normalized Intra-Event residuals

Figure 3-3 distribution of normalized inter and intra event residuals from OpenQuake results

In the next section the effect of including spatial correlation model in hazard calculation by in OpenQuake in our generated ground motion maps is presented and compared with the case of not including spatial correlation.

3.3.2.2 Checking the spatial correlation in Intra-Event residuals in derived ground motion maps with OQ.

To check the effect of using spatial correlation model in the generated ground motion maps in the modeling of this research, semivariogram is used, which is the same geostatistical measure that has been applied in previous studies such as the study of (Jayaram & Baker, 2009) to model and express the correlation between intra-event residuals. The normalized intra-event residuals have been calculated by using the process that has been explained in the previous section, for two cases with and without including the spatial correlation model of Jayaram et al. 2009 in our hazard calculation by OpenQuake. To calculate it, the distance between each site pair is calculated. The maximum distance range is divided into bins with h width. Each site pair, based on their distance, are assigned to the related bin. The average semivariogram of normalized intra-event residuals between each site pair in each bin is calculated by equation 4 (Jayaram & Baker, 2009).

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{\alpha=1}^{N(h)} \left\{ z_{u_\alpha} - z_{u_\alpha + h} \right\}^2 \qquad (4)$$

Where $\hat{\gamma}(h)$, is the location-dependent semivariogram, N(h) is the number of site pairs in each bin of separation distance of h, Z_u and Z_{u+h} the normalized intra-event values for αth site pair in the pertinent bin separated by h. As is shown later in this section, plotting of the resulted $\hat{\gamma}(h)$ values versus the related distance is used to check the existence of a correlation between the normalized intra-evet residuals for two cases of modeling hazard with and without spatial correlation.

Figure 3-4 (a) shows the ground motion map of an event in our stochastic dataset in the case of using the spatial correlation model in hazard calculation. Figure 3-4 (b and c) is showing the median PGA values and the total inter-event residuals that are independent of the spatial correlation and will be the same for both cases of hazard calculation with and without spatial correlation as it can be observed in Figure 3-6 . Figure 3-4 d is showing the total intra-event residuals for the case of considering spatial correlation. Figure 3-5 is showing the resulted semivariogram using normalized intra event residuals of the map of Figure 3-4 calculated with equation 2. This plot shows the existence of correlation up to a distance range of 40 km, and the correlation dies down after this separation distance.

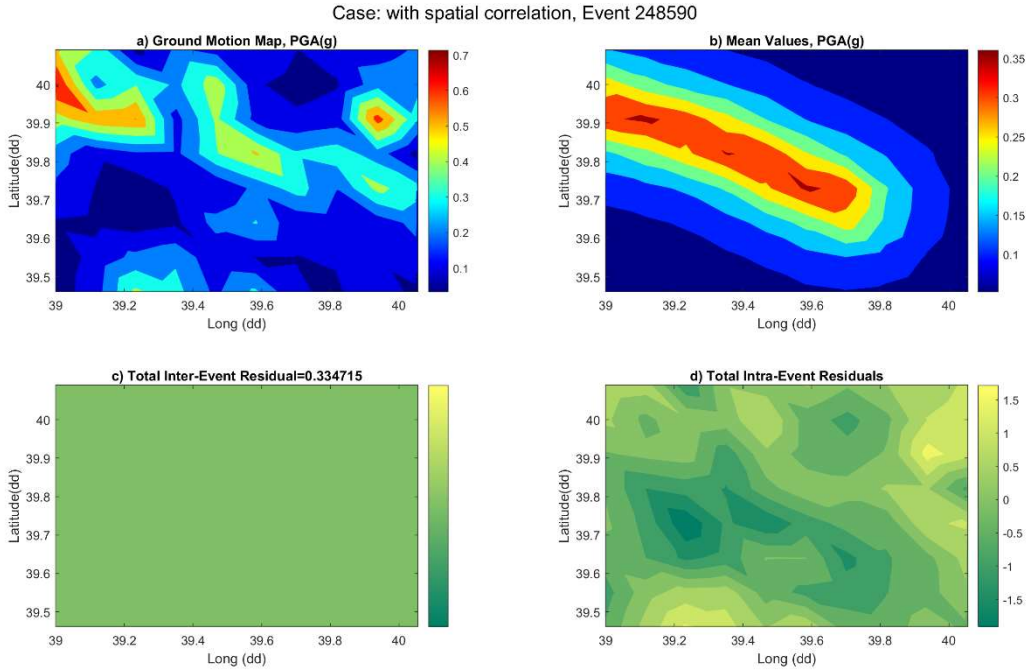


Figure 3-4 Ground motion map, the expected value, and uncertainties in case of including spatial correlation model, Event 248590. (1 decimal degree is about 111Km)

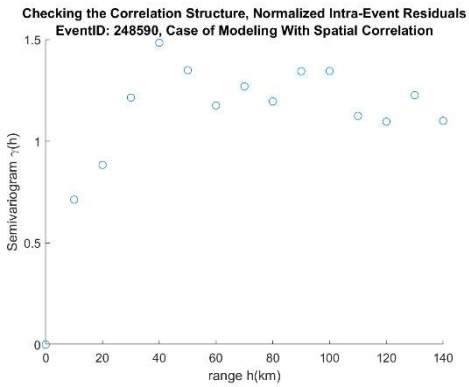


Figure 3-5 checking the correlation structure in results, case of including spatial correlation in calculation

Figure 3-6 (a) shows the ground motion map for the same event in the case of not including spatial correlation in the hazard modeling. As mentioned before and it is shown in Figure 3-6 b and C, the expected PGA values (median) and the inter-event residuals are like the related values in the case of including the spatial correlation as these two are independent of the spatial correlation. Figure 3-6 (d) shows the total intra-event residuals on the map in case of not using spatial correlation. Figure 3-7 presents the relation of the distance and the semivariogram and as is shown in this plot, there is no correlation between normalized intra-event residuals and the distance when spatial correlation is not included in hazard modeling and the values of semivariogram fluctuate in different distances without showing any pattern.

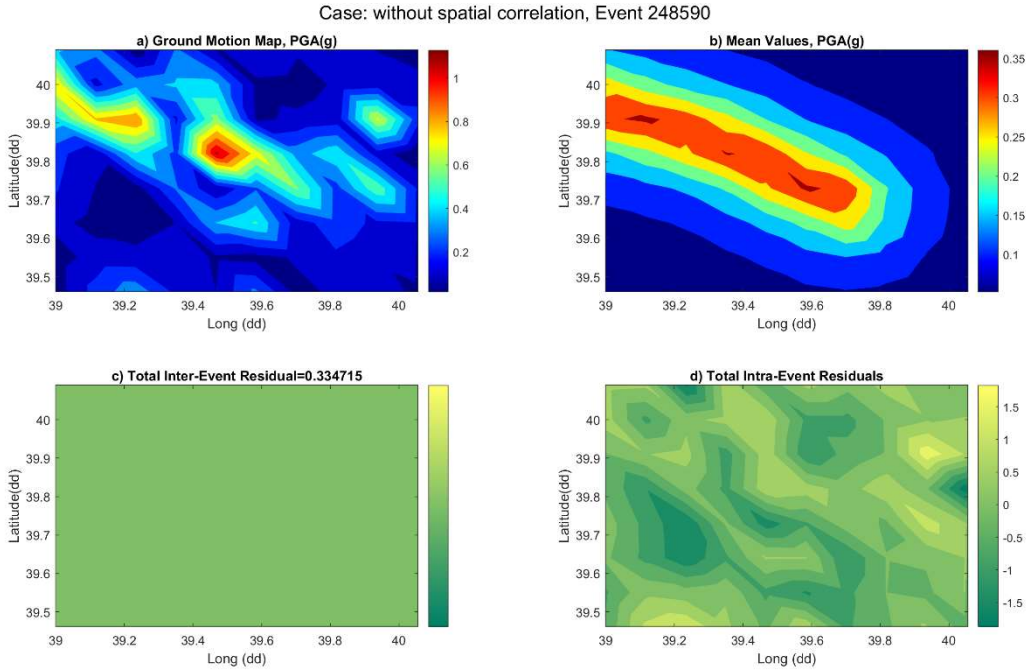


Figure 3-6 Ground motion map, the expected value, and uncertainties in case Not including spatial correlation model, Event 248590. (1 decimal degree is about 111Km)

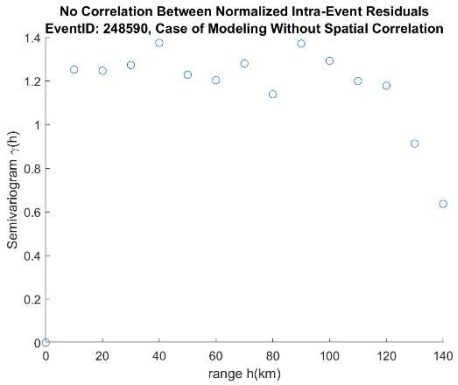


Figure 3-7 checking the correlation structure in results, case of NOT including spatial correlation in calculation

Moreover, it should be mentioned that not only the median value of ground motion but also the uncertainties have a significant effect on the resulted ground motion value for a site and the other interpretation for equation (1) can be the ground motion value at a site is exp(residual) times of the median value. For example, in Figure 3-6 by comparing the ground motion map with the median value and the maps of residuals, it is observed that the ground motion values can be very different than the median value and this is because of the uncertainties.

3.4 Loss Modeling

This part includes deciding about the type of fragility curves that will be applied in this research, the loss metrics that are suitable for our loss modeling, and the software and algorithms to calculate the required parameters to perform network analysis. Based on the data that are collected about the geometry of the

network and the location of the main components, which is just bridges in this study, and by having the ground motion maps from the results of our hazard calculations, the values of intensity measures in each bridge location are obtained to get the information of damaged or not damaged state of our bridge using fragility curves. Then network analysis is performed for two cases before and after the event to calculate our applied system level loss metrics. As mentioned before, the real risk assessment of the Erzincan is out of the scope of this study, and this is a simplified loss modeling in a seismic area to test our prescribed methodology.

3.4.1 Fragility curves

More information about the fragility curves and their application in loss modeling is reviewed in chapter 2. For this research, the fragility curves that are generated for bridges in Turkey from a study of (Avşar, yakut, & Caner, 2012) are applied. Because of the lack of detailed information about the structural type of the bridges in the case of study network, it is assumed that all have identical structures, therefore, an identical fragility curve is applicable for all our 87 bridges. The applied fragility curve is the curve generated for skewed greater than 30 degrees and multi-span single column reinforced concrete bridges as it was the most vulnerable bridge type in the study of Avsar et al. Here, only the curve associated with the extensive damage (collapse prevention) is used because it is the damage level of interest for this loss modeling. Figure 3-8 shows the applied curve in this research. The random damage assignment that has been explained in the introduction part of this chapter will be effective up to around PGA=1.7g and for greater PGAs (as the probability of exceedance is almost 1) the damage will be certainly assigned to the bridge that is in the affected area.

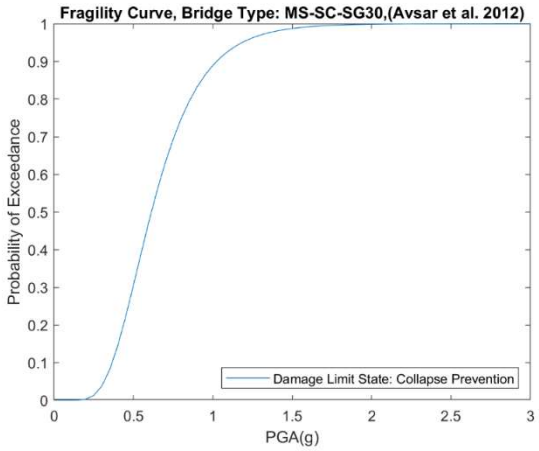
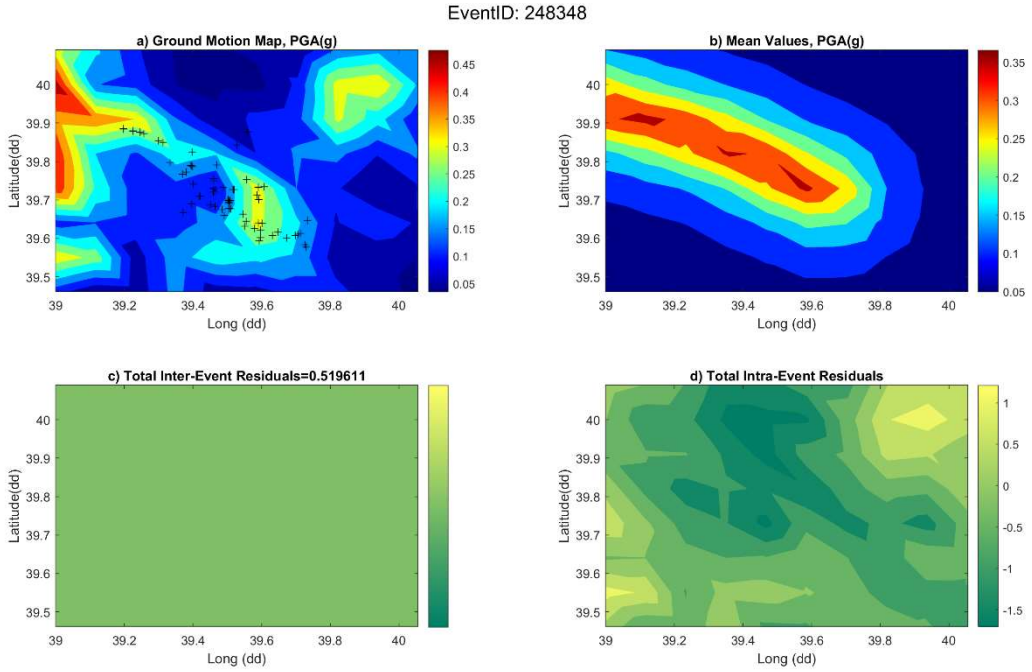


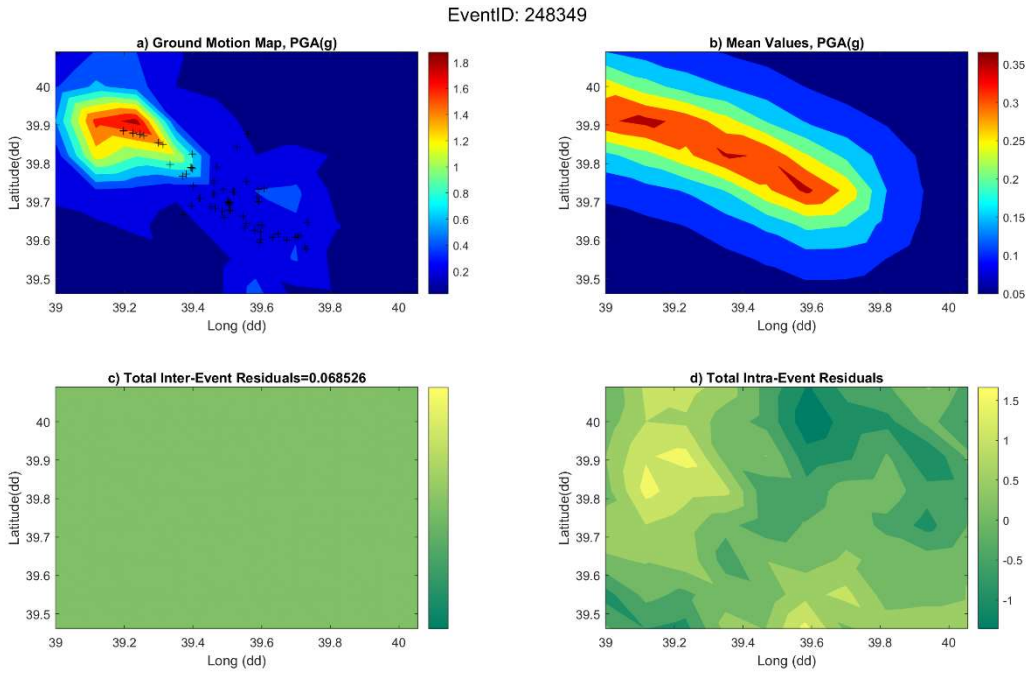
Figure 3-8 The applied fragility curve in this study, Bridge type: MS-SC-SG30 (AVSAR et al. 2012)

Figure 3-9 and Figure 3-10, are examples of two events (Event IDs= 248348 and 248349), related to the same source location, seismic characteristics, with the same magnitude 7.7, and the same applied GMPEs (which is (Akkar & Çağnan, 2010)) yet with different residuals that caused different ground motion maps causing different loss to the network. As shown in the b section of both figures, the expected values that are predicted by the GMPE are the same. The difference is in the intra-event residuals and the inter-event residuals that led to different ground motion maps. The location of the bridges of the network is shown by cross markers on ground motion maps in Figure 3-9(a) and Figure 3-10 (b), which shows the affected area of the network by hazard.



Number of Damaged bridges= 1

Figure 3-9 Ground motion map of Event 248348 and the location of the bridges of the network



Number of Damaged bridges=23

Figure 3-10 Ground motion map of Event 248349 and the location of the bridges of the network

Different residuals caused completely different damage to the network. This shows the sensitivity of the method of stochastic loss modeling to the number of events in the dataset. More events with more

realizations are required to result in a better estimation of the seismic risk. The requirement which is not usually possible because it makes the process computationally intensive.

3.4.2 Applied Loss metrics

For this study, some loss metrics that are related to road networks are applied. One of the applied loss metrics is at the component level to evaluate the direct damage due to an event and 4 loss metrics are at the system level to measure the effect of events on the connectivity of the network. As mentioned before, the real risk assessment of the Erzincan is out of the scope of this research, and this is a simplified loss modeling in a seismic area to test the proposed methodology of this research. However, the more sophisticated loss modeling of transportation networks normally includes consideration of different types of roads with specific flow demands or different flow demands according to the time of the day.

3.4.2.1 Total Length of Damaged Bridges (LDB)

This metric is one of the component level metrics and measures the direct damage due to an earthquake. Since the bridge is considered the most vulnerable component of the transportation network here, the only direct loss that is measured is related to the damaged bridges. Because the only damage level that is the concern of this loss modeling, the evaluation is binary and is just two cases of damage or not damages. Therefore, the applied component-based metric here will only reflect the collapse of the bridge. If the average construction price of the bridges per length is available, the total length of damaged bridges can be converted to a direct economic loss.

3.4.2.2 Total Time Delay (TD)

One of the system-level losses in a transportation network that might be the concern of decision-makers is the increase in travel time in the network due to the damaged components. Delay also can be converted to economic loss by having an estimation for the time value for the uses of the network, fuel price, delay cost in commercial shipments, etc. Driver`s delay is a commonly used metrics in measuring loss in functionality of a transportation network that is proposed by (Shinozuka, Murachi, Dong, & Zhou, 2003) and is calculated by equation (5). This metric is based on a redundant transportation network sample; therefore, it assumes even in the extensive damage levels, an O-D connecting path is still possible through other roads like secondary or residential. Therefore, if a connection loses a bridge, the connection is still available in the network but slower and with a lower capacity in connecting paths.

Total delay travel time (TT) =$\sum_a x_a t_a(x_a)$ (5)

Driver`s Delay (DD) =$\sum_a x_a^{`} t_a^{`}(x_a^{`}) - \sum_a x_a t_a(x_a)$ (6)

Travel time at flow x_a: $t_a = t_a^0\left[1 + \alpha\left(\frac{x_a}{c_a}\right)^{\beta}\right]$ (7)

Where in equations above (5,6,7): x_a is flow at link (in passengers per day), t_a is the travel time at flow x_a and is calculated by equation 7, where t_a^0 denotes the travel time in zero flow on link a (equals to the length of the link over its speed limit), c_a is the practical capacity of link a. α and β are empirical parameters that are usually taken as 0.15 and 4 respectively. Based on equation 7, if traffic flow in a link be equal to its practical capacity, the time of travel in that link will be 1.15 times of travel time in zero flow (Shinozuka, Murachi, Dong, & Zhou, 2003). The primed parameters in equation 6 denote the post-event condition means by subtracting TT in normal condition from TT in post-event, the driver`s delay is calculated.

Here the Driver`s delay is also applied as one of the system-level loss metrics. A fix traffic flow is considered for normal and seismic conditions as 1000 (passenger car per day) and a practical capacity for the roads in the pre-event condition as 2000 (passenger car per day) for all links. The speed limit in the pre-event condition is assumed to be 50 km/h. Following the estimation of change in link capacity and free flow speed limit due to damage levels of the link in a study implemented by (Shiraki, et al., 2007), it is assumed that in damage level of interest of this loss modeling, which is the extensive damage, the capacity will decrease by 50% and the zero flow speed limit as well. To calculate the preferred path by the network commuter, instead of using user equilibrium, it is assumed that the user just prefers the shortest path, and the Dijkstra algorithm is applied to find the shortest path between the O-D pair. Here, for simplicity, it is assumed that flow will remain fix in links, and in the case of a damaged bridge, only the practical capacity and speed limit are affected. It should be mentioned that in this study whenever the term of time delay or TD is used, it means the driver's delay or (DD), just because it is the only metric, among the metrics of this study, that is related to the time.

3.4.2.3 Simple Connectivity Loss (SCL)

This metric is one of the system-level metrics which shows the damage to functionality of a network and is applied in loss modeling of transportation and power and water network as well (Franchin & Cavalieri, 2013), (Kyriazis, Franchin, Khazai, & Wenzel, 2014), (Poljanšek, Bono, & Gutiérrez, 2012). Requires a set of predefined origin-destination (or source-sink) pairs, by comparing the post-event with the prevent condition measures the average loss connectivity in the network and is defined by the equation 8:

$$SCL = 1 - \frac{\sum\left(\frac{N_{j,s}}{N_{j,0}}\right)}{n} \quad (8)$$

Where $N_{j,0}$ and $N_{j,s}$ are the number of sources (origins) connected to the jth Sink (destination) in normal condition and seismic condition respectively. n is the number of sinks (destinations).

This metric is also applied here as one of the system level metrics because in infrastructure not only direct loss due to physical damage to the network is important but also the effect of the physical damage to the functionality of the network is one of the main concerns. Figure 3-1 shows our origin and destination points on the map which are 5 origin and 3 destination points. Considering the number of bridges in their path, the connection between these points is vulnerable to the earthquake. Here just a few O-D pairs are selected to increase the speed of the calculations.

3.4.2.4 Weighted Connectivity Loss (WCL)

This metric, like SCL, is a system-level metric but is a weighted version of SCL. This means that it penalizes the rerouted connections. Unlike the SCL that is just affected by lost connections, this metric also involves the effect of rerouting in the measurement of loss to the network and this can be a better evaluation of loss to the network than the simple connectivity loss. WCL is defined by equation 9:

$$WCL = 1 - \frac{\sum \frac{W_{j,s}N_{j,s}}{W_{j,0}N_{j,0}}}{n} \quad (9)$$

Where like SCL, $N_{j,0}$ and $N_{j,s}$ are the number of sources (origins) that are connected to the jth Sink (destination) in normal condition (pre-event) and seismic condition (post-event) respectively. n is the number of sinks (destinations). $W_{,0}$ and $W_{j,s}$ are the weight of the path to the jth sink over all the connected

sources in normal and seismic condition respectively. For a road network, the W (weight) can be defined as the inverse of the number of edges from the source (origin) to the sink (destination) for example, $W_{j,s} = \sum_{i=1}^{N_{j,0}} I_{ji,s} \frac{1}{N_{ji,s}}$ and in the same way for the weight in normal condition. (Franchin & Cavalieri, 2013)

3.4.2.5 Distance-based Weighted Connectivity Loss (DWCL)

The weighted connectivity loss penalizes the rerouted cases means that although a connection is not lost at least an effect of the increase in the path due to the damage to one of the critical components in one of the connections between an O-D pair is included in the measurement which can make this metric more reliable than the simple connectivity loss.

In the last section, the applied weight was (1/ (the increase of the number of edges in the path)), But in this research, the algorithm for finding the path is the shortest path, therefore, distance is the criteria in finding the path. This means that, a rerouted path will be longer than the pre-event path but does not necessarily include a greater number of edges in the rerouted state. The shortest path between a given origin-destination pair might include more edges in comparison with the other possible paths and this might lead to negative WCL values.

In this loss modeling, also a distance-based weight is used in calculating the connectivity loss and comparing the result with WCL in the previous section. To distinguish these two weighted loss metrics that are used here, this modified version of WCL is called Distance based Weighted Connectivity loss (DWCL). Which is practically the previous one with a different weight. The applied weight is the inverse of the distance in each path and the equation is modified as shown in equation 10. Figure 3-11 shows a normalized histogram of the differences between WCL and DWCL values from the results of this study. Although differences are negligible, the DWCL does not involve any negative value and apparently, it represents the loss of connectivity better by including the effect of the increase in distance as a parameter in this measure.

$$DWCL = 1 - \frac{\sum \frac{(DW_{j,s} N_{j,s})}{(DW_{j,0} N_{j,0})}}{n} \quad (10)$$

Where $N_{i,0}$ and $N_{i,s}$ are the number of sources (origins) connected to the jth Sink (destination) in normal condition and seismic condition respectively. n is the number of sinks (destinations). DW is the distance-based weight and is defined as $DW_{i,0}=1/D_{ji,0}$, in which $D_{ji,0}$ is the distance from ith source to the jth sink in normal condition. DW_{j0} is equal to the sum of all $DW_{i,0}$ over the connected sources to jth sink in normal condition. Also, $DW_{i,s}=1/D_{ji,s}$ is the distance from ith source to the jth sink in seismic condition. DW_{js} is equal to the sum of all $DW_{i,s}$ over the connected sources to jth sink in seismic condition.

As an example of the result of using DWCL and its difference from the other connectivity loss metrics, a comparison of the calculated loss value with SCL and WCL and DWCL is shown in Figure 3-12 for a case of a simple graph. This figure presents the non-seismic and seismic conditions of the graph. The calculated loss value for the same case of damage to a graph is 0, 0.25, and 0.4 for SCL, WCL, and DWCL respectively. This comparison shows that using the distance of the path penalizes rerouting in the resulted loss values more than WCL when edge numbers are applied as weights. Both weighted metrics have higher loss values than the SCL that does not represent the rerouting as a loss or damage to the network. For more examples of the calculation of SCL and WCL loss metrics see (Franchin & Cavalieri, 2013).

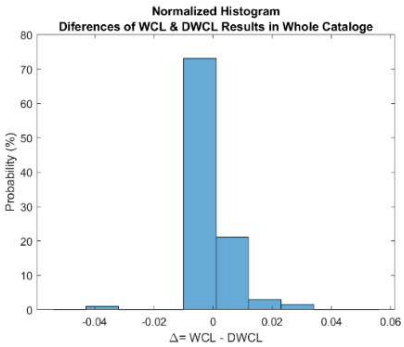


Figure 3-11 Normalized histogram of differences of WCL and DWCL values among cases in this study.

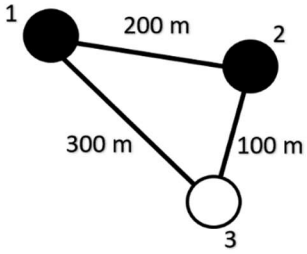
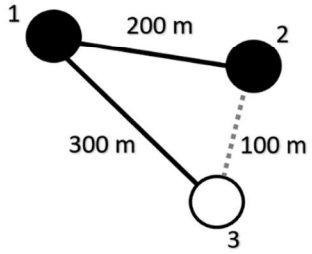
Normal (Non-Seismic) Condition		
	$N_{1,0}$=1 $N_{2,0}$=1 *Weighted based on number of edges (link) in the path:* $W_{1,0}$=1/1=1 $W_{2,0}$=1/1=1 *Weighted based on distance of the path:* $DW_{1,0}$=1/300 $DW_{2,0}$=1/100	**Legend** Source (Origin) Sink (Destination) Undamaged link (edge) Damaged Link (edge)
Seismic Condition		
	$N_{1,s}$=1 $N_{2,s}$=1 *Weighted based on number of edges (link) in the path:* $W_{1,s}$=1/1=1 $W_{2,s}$=1/2=0.5 *Weighted based on distance of the path:* $DW_{1,s}$=1/300 $DW_{2,s}$=1/(300+200) = 1/500	$\mathrm{SCL} = 1\text{-}\frac{1}{2}\left(\frac{1}{1}+\frac{1}{1}\right) = 0$ $\mathrm{WCL} = 1\text{-}\frac{1}{2}\left(\frac{1}{1}+\frac{0.5}{1}\right) = 0.25$ $\mathrm{DWCL} = 1\text{-}\frac{1}{2}\left(\frac{\frac{1}{300}}{\frac{1}{300}}+\frac{\frac{1}{500}}{\frac{1}{100}}\right) = 0.4$

Figure 3-12 Comparison of the calculation of the connectivity loss with SCL, WCL and DWCL

3.4.3 Loss Modeling

The network is modeled as a digraph (directed graph) in MATLAB and the components of a network are represented by nodes (vertices) and edges (links). Now by having the ground motion maps per each event

ground motion values can be interpolated to the location of bridges that the location of the center of bridges is modeled as a node in our digraph. By having the value of intensity measure of the interest (which is Peak Ground Acceleration (PGA) in our study), the probability of exceedance of a damage level is derived from the selected fragility curve. To assign the damage level to the bridge a Monte Carlo method is used that has been applied in the study of (Poljanšek, Bono, & Gutiérrez, 2012). Which is about generating a random variable between 0 and 1 and comparing it with the value of the probability of exceeding the damage level derived from the fragility curve. If the value of the random number is less than the probability, the bridge considers as damaged. As was mentioned in section 3.4.1 and shown in Figure 3-8, This random damage assignment method will be effective up to around PGA=1.7g, and for greater PGAs the damage will be certainly assigned to the bridge that is in the affected area as the probability of exceedance is almost 1.

The damage level of concern in this study is the extensive damage therefore assigning this damage to a bridge means that bridge is lost. Therefore, the next step is the elimination of damaged bridges from our network. By having the total length of damaged bridges, the LDB loss metric can be calculated in this step.

The downside of modeling bridges as nodes in the digraph is that by eliminating a node from a graph or digraph, it is possible to get new updated names for all vertices and this might cause errors in further network analysis because the names of origin-destination nodes will be affected. Therefore, in this step, for each event, after eliminating the damaged bridges from the network, the program that is written in MATLAB for loss modeling of this research finds the updated name of our desired o-d pairs by their coordinate and applied their new name in further commands to avoid any possible mistake. Hence, this step is not required if someone models the bridge as a link of the graph or digraph.

The next step is calculating the other loss metrics that are at system-level and network analysis is needed to perform. To simplify network analysis, only 6 origin points and 3 destination points have been considered which are shown on the map in Figure 3-13. But as mentioned before, these points are more vulnerable to the damage of bridges in the network due to the existence of bridges in their connecting path. to calculate the connection between each origin-destination pair, the shortest path is calculated by using the Dijkstra algorithm, which was proposed by (Dijkstra, 1959). In the next section, more explanations are provided about this shortest path algorithm. By implementing network analysis and comparing the pre-event condition with the result of the post-event condition, the 4 system-level loss metrics can be calculated which are time delay, simple connectivity loss (SCL), Weighted connectivity loss (WCL), and our defined metric Distance based weighted connectivity loss (DWCL).

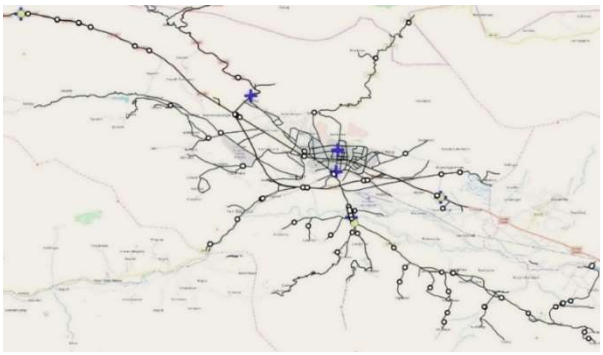


Figure 3-13 The assumed important nodes of the network that were vulnerable to damage of the bridges. the blue marks are the origins, and the green ones are the destination nodes the white circles are locations of bridges. (Map data source: OpenStreetMap contributors, https://www.openstreetmap.org)

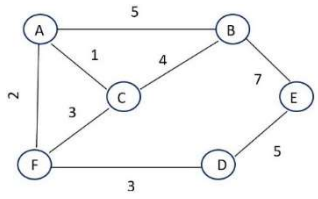
3.4.3.1 Dijkstra Algorithm

This algorithm was proposed by Dijkstra, a Dutch computer scientist in 1959, and it finds the shortest path from one starting node in a graph to other nodes and is one of the popular algorithms that is applied in network analysis. The process includes making an information table and two tracking lists that are for the visited and unvisited vertices and updating them in each step. Here visiting a node means calculating the shortest distance from the unvisited neighbors of that node from the starting node. The steps for this algorithm are as follows and an example of finding the shortest path in a graph is shown in Table 3-1, steps are:

1- Create two lists of visited and unvisited vertices
2- Create the information table including 3 columns of names of the nodes, the shortest distance to the starting vertex, and the previous vertex
3- In the second column of the information table, allocate initial values for the starting point equal to zero, and for the rest of the node infinity
4- Visit an unvisited node that has the shortest value in column 2 of the information table
5- Update the values on the information table and the tracking lists
6- Repeat step 3 to 5 until there is not any unvisited vertex left on the list

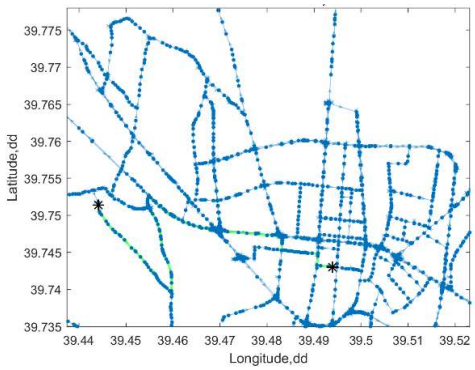
Table 3-1 an example of the shortest path calculation with Dijkstra Algorithm

Vertex	Shortest distance from the starting vertex=A	Previous vertex
A	0	
B	5	A
C	1	A
D	5	F
E	10	D
F	2	A

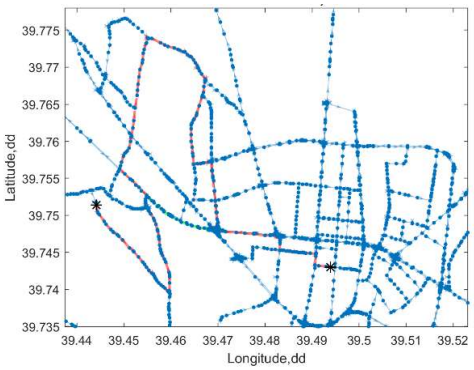


3.4.3.2 Examples of rerouting and loss connection in the model due to a damaged bridge

In section 7.4.2.1 more detail and the steps of this algorithm are explained. In Figure 3-14 and Figure 3-15 the modeling of the network in two cases of a rerouted connection and a lost connection is shown.

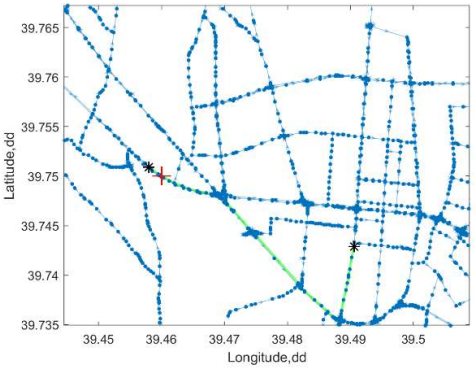


Green path: path before rerouting

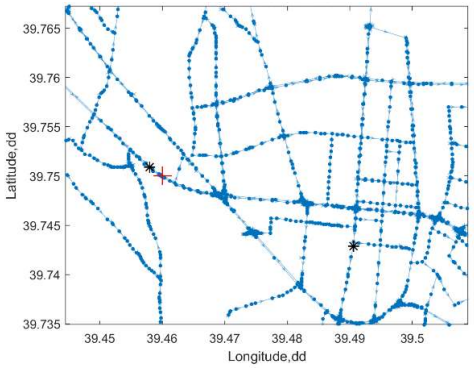


Red Path: Path After Rerouting

Figure 3-14 An example of modeled rerouted path



Green path: path before losing the connection



Red Mark: the damaged bridge that caused losing the connection between the black marks

Figure 3-15 An example of a modeled lost connection

3.5 Review on Some Clustering Algorithms

Clustering or unsupervised classifying, which is also widely used in Data Mining and Machine learning, is a technique of getting and evaluating abstract data based on their features and dividing them into smaller subsets in a way that similarity between members of a subset is maximum. In clustering, the distance metric that uses in comparing dissimilarities chiefly controls the formation of clusters; therefore, based on the type of features an appropriate distance metrics should be applied. In the following some of the common distance metrics in clustering are explained:

3.5.1.1 Distance Metrics

In clustering the distance metric chiefly controls, the formation of clusters; therefore, based on the type of features an appropriate distance metrics should be applied. Some popular distance metrics are as follows:

3.5.1.1.1 Euclidean Distance

Euclidean distance is the most popular distance metric and samples for A and B with I number of features are calculated by equation (11):

$$D_{Euclidean} = \sqrt{\sum_{i=1}^{I}(A_i - B_i)^2} \quad (11)$$

3.5.1.1.2 Manhattan Distance

Manhattan distance is the sum of the distance of features, and for samples, A and B with I number of features is calculated by equation (12)

$$D_{Manhatta} = \sum_{i=1}^{I}|A_i - B_i| \quad (12)$$

3.5.1.1.3 Minkowski Distance

It is the generalized form of two previews distance metrics and for samples, A and B with I number of features is calculated by equation (13)

$$D_{Minkowski} = \left(\sum_{i=1}^{I}|(A_i - B_i)|^r\right)^{\frac{1}{r}}, r \geq 1 \quad (13)$$

3.5.1.2 K-means Clustering

One of the most popular partitioning clustering algorithms is K-means clustering which clusters data into the predefined K number of groups or clusters. The main challenge in the K-means clustering method is that the number of clusters should be determined in advance, and the number of clusters affects the accuracy of clustering. The steps of this algorithm:

1. Determine the number of clusters (K)
2. Pick a center point for each of k clusters, there are two options for this step:
   2.1. Randomly select k observation out of the dataset
   2.2. Create k random points in each of k clusters
3. Determine a distance metric and calculate the distance of sample in dataset from the K centroid points and create a distance matrix which consists of k columns and rows in number of the dataset
4. Assign a centroid point to each sample by comparing the columns in each row and find the closest centroid point for that row (that sample from the dataset)
5. Calculate the mean in each cluster and update the centroid points of each cluster with the calculated mean
6. Repeat steps 2 to 5 until no change happens to the initial centroid points
7. Repeat the whole process several times with different initial centroid points to make sure that you get the best clustering outcome

To find the suitable number of clusters, a measure for accuracy can be determined, like the Sum of Squared Errors (SSE), and can be used as the distance between each point in a cluster and the centroid point, as shown in equation (14), and the mean of this SSE for all clusters is considered as an accuracy measure for the whole clustering. By increasing the K value, the error decreases. If the calculated errors be plotted versus the number of Ks, usually in a specific number of K the rate of decrease in errors is changed and reduced significantly. This point is called the elbow point and its related K number can be used, to some extent, as a satisfactory clustering number (Aghabozorgi, 2020).

$$SSE = \sum_{i=1}^{n_j}\left(x_i - C_j\right)^2 \quad (14)$$

Where the x denotes the ith member of jth cluster and C_j is the center point of j_{th} cluster. Member of the K-means Clustering algorithm falls short in dealing with the outliers as they force all the data to be grouped in a cluster which in some cases might be regardless of the insufficient similarity of the members of a cluster. Moreover, the shape of clusters is always spherical in K-Means clustering. In contrast, Density-based clustering, which is explained in the next section, is a robust method of dealing with outliers and the shape of clusters is arbitrary.

3.5.1.3 Density-based clustering Algorithm (DBSCAN method)

One of the popular density-based clustering algorithms is Density-based Spatial Clustering of Applications with Noise (DBSCAN) proposed by (Ester, 1996). This algorithm is also very effective when it is used for spatial data. One main assumption for this method is that the member of a cluster should be at a certain distance from other members of the cluster. therefore, to shape the clusters, R and M values should be defined initially as the maximum distance between members of a cluster and the minimum number of members in a cluster, respectively. One of the advantages of this method is that clusters could have any shape. Also, in contrast to the K-means that all the data including outliers are clustered, in the DBSCAN method, the outliers would be detected (Gan, Chaoqun, & Jianhong, 2020).

In this algorithm, the dataset is evaluated and classified as a core, border, or outlier. A core is a point that has M members around it in a maximum R distance. A border point is a point that is in the required vicinity of a core (R) but does not have M members in its neighborhood therefore is not a core. An outlier is neither a core nor a border point. The steps of this algorithm are as follows (Aghabozorgi, 2020):

1- defining R and M value
2- classify the dataset as a core, a border, or an outlier
3- Clusters are shaped by the collection of all core points that are in R distance to each other including their related border points.

As the R and M should be defined in advance, determining the R value can be challenging especially in the case of samples with multiple features.

3.6 Proposed Reduction Methodologies

As it is mentioned at the beginning of this chapter, the methodology of this research in the reduction is applying partitioning clustering algorithms to cluster all events in less dissimilar groups and represent each group/cluster by one event from each cluster. The main target here is to produce a risk curve by this reduced event set as similar as the risk curve that is developed with the whole package for each loss metric. Therefore, the occurrence rate of the representative event of each group is modified with the sum of the occurrence rate of all the members in the cluster. One of the effective items in clustering is selecting the features involved in calculating the distance metric. In the simple Grid-Based clustering only the loss values are used to differentiate the clusters. But in the K-means clustering algorithm 4 different approaches are applied based on the selected features to calculate the Euclidean distance as the distance metric of the clustering algorithm and those approaches are Risk-Consistent, Weighted Risk-Hazard, Unweighted Risk-Hazard, and Hazard-Consistent. Each clustering procedure and the approaches are explained in the following:

3.6.1 A simple spatial griding method (a Risk-Consistent approach)

This method is inspired by grid-based clustering but does not follow any known methods in the grid-based clustering algorithm. In *the first step*, to imitate the spatial distribution of the whole dataset scenarios, the area is divided into small spatial grids as our initial criteria for grouping the events. In *the second step*, the range of a calculated loss metric for all scenarios is divided into small bins. Scenarios (events) that are spotted in each grid due to the coordinate of their sources, are grouped in their pertinent loss bin as the secondary clustering criteria. As a result, the scenarios (events) that are grouped in a cluster are in the same region, and the same loss bin. This procedure is schematically shown in Figure 3-16. In *the third step*, in each group, one scenario (event) with the maximum magnitude is selected as the representative of the other scenarios (events) in that group and the occurrence rate of the events in a group is summed up and is assigned to the representative scenario of that group/cluster. The reason behind this modified rate for the cluster representative scenario is to make sure that the resulting loss exceedance curve with the reduced dataset will be similar to the one that is generated with the whole dataset of events.

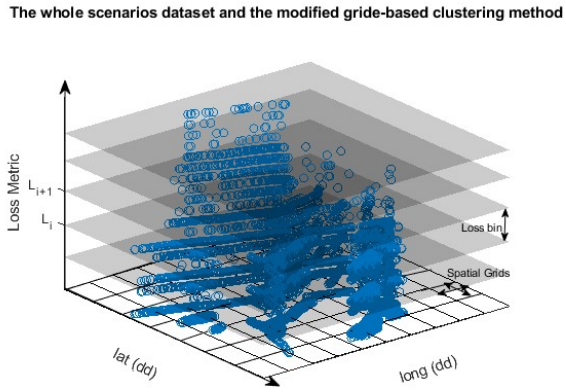


Figure 3-16 Modified Grid based Clustering Method.

3.6.2 Using K-Means Clustering algorithm

For calculating dissimilarities, the distance metric is applied in k-means clustering and one of the common distance metrics is Euclidean distance. In this study, also Euclidean distance is applied in k-means clustering. The features that are applied as the effective items in clustering of the events are selected based on the approaches that are wanted to be pursued. In following the procedure is explained more.

3.6.2.1 Risk-Consistent

In this approach, like what have been implemented in the simple grid-based clustering, our main parameter to compare the events with is the value of a specific loss metric for each event. Therefore, the loss metric value is the only feature to be involved in the calculation of the distance metric. Finally, after clustering all the events, one event that is spatially close to the center of each cluster is selected as the representative of the cluster and the occurrence rate of the representative is replaced by the sum of the occurrence rates of all the events in that cluster.

In addition to the elbow point as one of the criteria to select the optimum K number, in this Risk-Consistent approach, a risk-based method has been proposed to estimate the optimum clustering numbers related

to the resulted risk. In this method, the clustering for several k numbers is performed, and for each run, the average sum of squared errors (the error in this optimization method means the difference between the resulted risk curve by the reduced data set with the comparison with the resulted risk curve by the whole data set) for 3 specific return period of 475, 975, and 2475 years is calculated and plotted versus the pertinent K number of that clustering. Errors changes by increasing the number of clusters and the point in which the slope of decreases in alleviated can be selected as the estimation for the optimum number of clusters, in an assumption that those 3-return periods are the most important return periods for the analyst. Equation 15 shows the applied distance metric in this approach:

$$Distance\ Metric = \sqrt{(L_A - L_B)^2} \tag{15}$$

Where L_A and L_B Denotes the loss value in event A and B respectively.

3.6.2.2 *Weighted Risk-Hazard*

In this approach, multi-feature clustering is applied. This means that in addition to considering the loss metric as one of our features of clustering samples, the effect of hazard is also involved in the clustering by considering the value of engineering intensity measure (here is PGA) in 5 different locations of bridges in our network. In calculating the Euclidean distance, a 50% weight is considered for the loss metric feature and 10% for each of the intensity measures (IMs), to have the dominancy of loss in clustering. This approach is called here as Weighted Risk-Hazard clustering. These 5 different locations of the bridges are also the same location that will be compared the effect of the reduction on their hazard curve with their derived hazard curve by the whole dataset. In this clustering, after standard scaling the data in Python programming language, the features are multiplied with the determined weight matrix and then performed the clustering. it is expected to see better hazard curve results in the location of bridges that are involved in this clustering, but these improvements should not be significant and at the same time, a less consistent risk result in comparison with the Risk-Consistent approach. Therefore, to get better consistency in results, a greater number of clusters might be needed which leads a larger reduced dataset. Equation 16 shows the applied distance metric in this approach:

$$Distance\ Metric = \sqrt{0.5\ *(L_A - L_B)^2 + 0.1 * \sum_{j=1}^{5}\left(PGA_{Aj} - PGA_{Bj}\right)^2} \tag{16}$$

Where L_A and L_B Denotes the loss value in event A and B respectively. The PGA_{Aj} and PGA_{Bj} is the values of PGAs in site j affected by event A and B respectively.

3.6.2.3 *Unweighted Risk-Hazard*

In this approach, like Weighted Risk-Hazard one, the same features are applied. But unlike the previous approach, after standard scaling of the samples, no weighting is applied in calculating the Euclidean distance. Thus, there is no dominancy this time and all the features are treated the same in clustering. the goal is to slightly increase the effect of the hazard and check the difference with the time that the 50% weight is considered for the loss values in the clustering. the prediction here is that better hazard curves but less consistent loss exceedance curves for the similar clustering number with the Risk-Consistent approach will be generated. Probability to get better consistency a greater number of clusters is needed

and consequently a greater number of events in a reduced dataset. Equation 17 shows the applied distance metric in this approach:

$$Distance\ Metric = \sqrt{(L_A - L_B)^2 + \sum_{j=1}^{5}\left(PGA_{Aj} - PGA_{Bj}\right)^2} \quad (17)$$

Where again L_A and L_B Denotes the loss value in event A and B respectively. The PGA_{Aj} and PGA_{Bj} are the values of PGAs in site j affected by event A and B respectively.

3.6.2.4 *Hazard-Consistent*

In this approach, only 5 PGAs in 5 sites, which are locations of bridges, are considered as features in calculating the Euclidean distance but like the other 3 approaches, here also the representative of each cluster gets the sum of the occurrence rates of all members in a cluster. The first part of providing this reduced dataset is all the events that do not cause damage to at least one bridge in the network are eliminated. This means that here it is not expected that the results be purely Hazard-Consistent because the clustering samples are just the mentioned damaging events. Of course, if the whole complete list of events (including damaging and not damaging) were used by a reasonable number of clusters the real hazard curve could be generated, but the goal here is first to slightly include the loss effect of the event in the resalted subset of catalogue but keep the hazard effect dominant. Therefore, just in comparison with the other 3 approaches, here this approach is called as Hazard-Consistent. Finally, the risk curve and the hazard curve for reduced datasets are developed and compared with our previous risk-involved approaches with K-means clustering. Equation 18 shows the applied distance metric in this approach:

$$Distance\ Metric = \sqrt{\sum_{j=1}^{5}\left(PGA_{Aj} - PGA_{Bj}\right)^2} \quad (18)$$

Where PGA_{Aj} and PGA_{Bj} denote the values of PGAs in site j affected by event A and B respectively.

3.6.3 Density based clustering approach

As mentioned before, one of the advantages of the density-based clustering algorithm is its ability to spot outliers and not force all the events to be clustered. Therefore, the goal is to use this algorithm to test the possibility of increasing similarities in clusters and at the same time keep and include the outliers individually in the reduced dataset. Here, this clustering is just applied in the Risk-Consistent approach. For this algorithm two parameters of R and M (distance limit and a minimum number of neighbors) are required to be predefined, finding the optimal number of clustering might not be as straightforward as the elbow point in k-means clustering. Therefore, for this method, the process was iterated several times by changing these R and M values to get a reasonably small number of clusters and have a similar value of clusters to the results of K-means for each loss metric. In this way, the comparison of the result of this method with the other two previous methods, k-means and simple spatial griding method, in the Risk-Consistent approach will be easier.

3.7 Conclusion

In this chapter, the exposure model and the case of study as well as a sample of derived hazard curve and some of the ground motion including the effect of considering the spatial correlation model on generated ground motion maps were covered. Then the applied loss metrics were defined including a modified version of WCL that is called here as the distance-based weighted consistency loss (DWCL). To explain the

methodology of this research, some basic information about clustering algorithms and their applications were reviewed. finally, the reduction methodologies that are proposed and applied in this research were explained. The simple spatial griding method was introduced which is a Risk-Consistent approach. Then the main proposed reduction method which is based on the k-means clustering algorithm was explained that is implemented in 4 different approaches: Risk-Consistent, Weighted Risk-Hazard Consistent, Unweighted Risk-Hazard Consistent, and Hazard-Consistent. Also, the third methodology was explained that applies density-based clustering to check the effect of excluding outliers from the grouping event on the resulted risk curves with the reduced dataset. In the next chapter, the results of these approaches will be discussed and compared in detail.

4 Results and Discussions

As mentioned before, the scope of this research is to reduce the required number of earthquake scenarios to be applied in stochastic loss modeling. To do so, 4 different approaches were adapted as Risk-Consistent, Weighted Risk-Hazard, Unweighted Risk-Hazard, and Hazard-Consistent. The applied methodology in this study is based on some clustering algorithms and an assumption that one member of each cluster can represent the risk effect of all other similar scenarios that are members of the same cluster. The requirement for this reduction method is to sum up the occurrence rates of the eliminated similar scenarios and add them to the rate of the remaining representative of the cluster. What means similarity is defined according to the approach that is taken from the mentioned 4 different approaches.

The order of presenting the results in this section is first the result of our simple spatial griding reduction algorithm are stated for two loss metrics as Simple Connectivity Loss (SCL) and time delay. The approach in this method is Risk-Consistent. In the second section of this chapter, the results of applying the k-means clustering algorithm to group similar scenarios are explained. The 4 different approaches were studied by this clustering algorithm for all of the applied loss metrics. The result of each approach is presented in the order of each loss metrics separately, which means that for example, in the case of Length of Damaged Bridges (LDB) as the applied loss metric, the results of all 4 approaches for this loss metric are presented first and their differences are compared. then the same is implemented for the other loss metrics separately. In the third section of this chapter, the results of the last method in grouping earthquake scenarios which is the density-based clustering algorithm (DBSCAN) are explained. Only the Risk-Consistent approach is studied by applying DBSCAN for two different cases of loss metrics, LDB and Distance-based Weighted Connectivity Loss (DWCL). Finally, a hypothetical network improvement scenario is considered, and the compatibility of the risk curve generated by the reduced dataset (that has been developed based on the loss results in the pre-retrofit condition) with the whole dataset of new loss results after the system improvements. This application is checked for the results of the Risk-Consistent approach for 2 metrics which are again LDB and DWCL. Table 4-1 shows a list of upcoming sections of this chapter that are related to the results of the applied reduction methods in this research. In the following, all the explained steps will be illustrated in detail.

Table 4-1 Upcoming Sections of this chapter about Results of applied Reduction Methods

Method	**Loss Metric**	**Approach**
5.1 Simple Spatial Griding	*5.1.1* TD	• Risk-Consistent
	5.1.2 SCL	
5.2 K-means clustering	*5.2.1* LDB	• Risk-Consistent • Weighted Risk-Hazard • Unweighted Risk-Hazard • Hazard-Consistent
	5.2.2 TD	
	5.2.3 SCL	
	5.2.4 WCL	
	5.2.5 DWCL	
5.3 Density based clustering	*5.3.1* LDB	• Risk-Consistent
	5.3.2 DWCL	

It should be mentioned that to perform loss modeling with the whole data set to produce a reference loss curve, at first the events that did not affect the most of the area, were removed and considered as not important. Second, in calculating the loss metric values, the first step was to calculate the direct loss which is faster than performing the network analysis for the rest of loss metrics. During the evaluation of the

possibility of the direct loss for each event, each event that did not damage at least one bridge of the network was eliminated (here, the bridge is considered the most vulnerable component of the network). By doing these eliminations about 16000 events remained out of about 360000 events and this smaller but damaging subset of the whole package was used to be reduced further with the reduction methodologies of this research.

4.1 Results of reduction by simple Spatial Griding Method

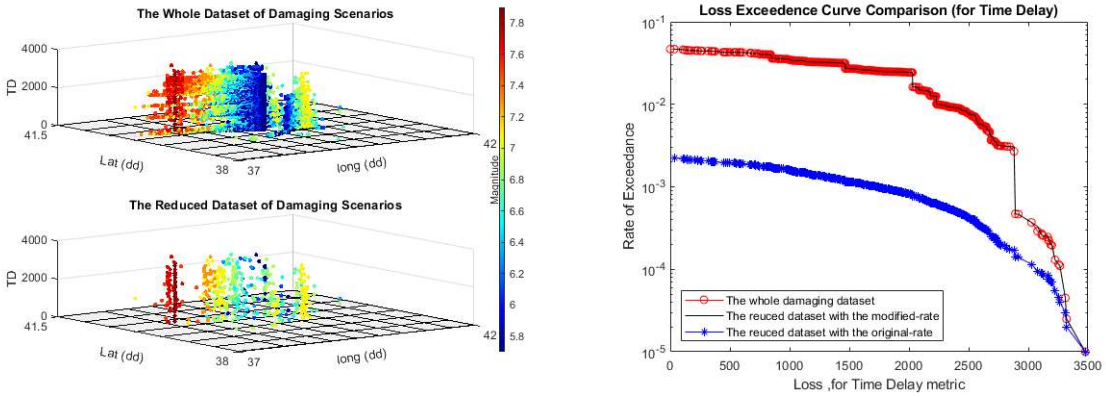
In reference to the chapter of modeling and methodology, which demonstrated the simple spatial griding method in an explicit way, this reduction methodology is based on a Risk-Consistent approach. the goal of this method is to get a risk curve from the reduced list of earthquake scenarios, as similar as possible to the generated risk curve from the whole list of stochastic earthquake scenarios for the region.

The algorithm includes spatially griding the area to a predefined size of grids and the range of values of the applied loss metric divided into equal bins (this can also be an unequal layering by applying logarithmic scale in diving loss values) which provides several layers over the area of a spatial grid which enclosed by two vertically adjacent loss layers shapes a cubic cluster. Then the algorithm identifies all events that are occupied the same cubic cluster. This means that clustering is based on the seismic source coordinates and its consequent loss as well.

The reason for including the effect of coordinates in the grouping is to keep a similar spatial pattern with the whole event set. After grouping the scenarios as explained, a representative from each group should be selected. This step can be performed by several criteria like the closest scenario to the center of the denser area in the cubic cluster, the one with the highest magnitude, the highest value of the consequent loss, etc. Here the magnitude was chosen as an effective parameter in selecting the representative scenario, as the goal was to keep important events in our subset. This method was implemented by applying two different loss metrics; time delay (TD) and Simple connectivity loss (SCL) which will be presented in the following sections respectively. At the end of this section, the results and pros and cons of this method will be discussed.

4.1.1 Results for total time delay (TD)

The time delay loss metric values were applied and divided into several equal subsets to produce loss layers over the grided area. The size and number of resulted cubic-shaped clusters affect the accuracy of the developed loss exceedance curve by the results in comparison with the results of the whole dataset. Moreover, if there are a few repetitive distinct values in loss results the maximum possible number of layers will be limited to the number of distinct values in loss results. Figure 4-1 shows a result of clustering with 70 spatial grids and 90 loss layers that produced a subset of 453 representative earthquake scenarios. Figure 4-1 (a) shows the spatial distribution pattern of the reduced data set and the whole one, it seems spatial griding helped to keep the desired pattern. Also, by choosing the highest magnitude in each cubic-shaped cluster a good number of severe events remained in the resulted subset. Figure 4-1 (b) is the comparison of the generated loss exceedance curve (risk curve) with the reduced data set in two conditions: the blue line is the curve without changing the occurrence rate of representative events and the black line shows the effect of performed modification in occurrence rate to develop a compatible risk curve with the true risk curve that is generated with the whole dataset.

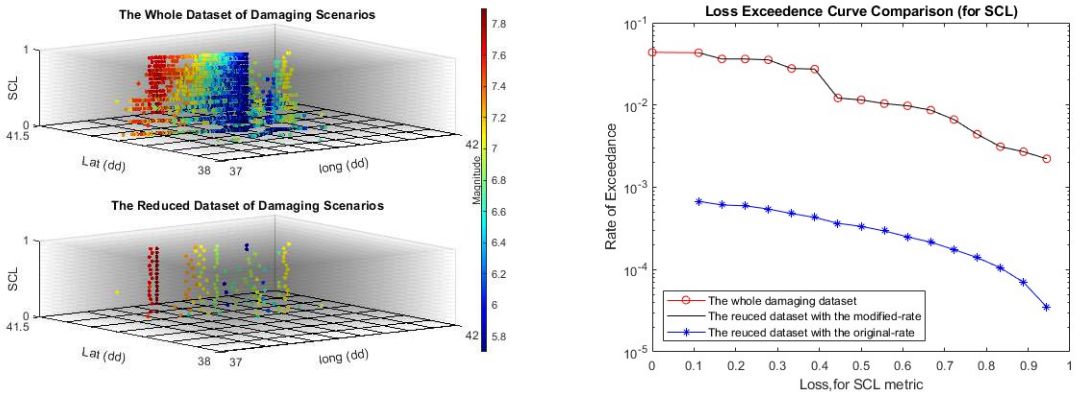


a b

Figure 4-1a) Comparison of Reduced Dataset of Damaging Scenarios with the whole one, Reduction method: Simple Spatial Griding Method, Loss Metric: TD, b) comparison of Loss Exceedance curve, (unit of y axis is 1/year).

4.1.2 Results for Simple Connectivity Loss (SCL)

The simple spatial griding method was performed with the values of SCL to develop loss layers. For 18 vertical loss layers and 70 spatial grids and 144 representative reduced datasets were produced. Figure 4-2 shows the results with this loss metric. Figure 4-2 (a) shows the spatial distribution of the reduced dataset and compares it with the reference dataset, similar pattern, and a good subset of events from the aspect of including events with high magnitude is produced. Figure 4-2 (b) shows the loss exceedance curve resulting from the reduced dataset and the whole one. It also shows the effect of our rate modification in predicting accurate results by the comparison of the blue line as a result of a reduced subset without rate modification and the black line is related to the same dataset but with the modification occurrence rate for the representative earthquake event in the reduced dataset. Due to a limited number of distinct values in our SCL results, which makes a set of layers unoccupied in case of increasing the number of layers, for fixed size and number of spatial griding fewer datasets were obtained, and better risk results than the case of TD in the previous section.



a b

Figure 4-2a) Comparison of Reduced Dataset of Damaging Scenarios with the whole one, Reduction method: Simple Spatial Griding Method, Loss Metric: SCL, b) Comparison of Loss Exceedance Curve, (unit of y axis is 1/year)

4.1.3 Discussions on the results of simple Spatial Griding Method

In this section, the simple spatial griding was covered with two examples of loss metrics. As it was shown, the number of cubic shape clusters, as well as the applied loss metric and their values, will affect the number of events in the resulted reduced dataset. To discuss the pros and cons of this method, it can be mentioned that the simplicity of the steps makes this reduction method fast and easy, this method evaluates all the scenarios based on their potential loss and includes their risk contributions in the reduced dataset, it can follow the spatial distribution of the whole dataset (to some extent), and at the end can provide a smaller Risk-Consistent dataset. But to address the cons; the obligation of providing predefined grid sizes and loss layers results in a biased selection. On the other hand, this methodology forces all the events to be identified as a member of a cluster and not leave anyone as an individual event. Therefore, the definition of similarities in each cluster might not be realistic in an aspect of seismicity of the events and finally, all the clusters are similar in shape.

4.2 Results of reduction by K-Means clustering Algorithm

In this section, the results of applying K-means clustering for our reduction method will be illustrated. The method for selecting the representative from each cluster is selecting the closest event to the center point of each cluster based on our distance metric values. As mentioned at the beginning of this chapter, here four different approaches will be checked for each loss metric and will be explained one by one. In each subsection of this section, the result of 4 different approaches for each loss metric will be discussed. For the k-means clustering, it will be attempted to find an estimation of an optimum number of clustering by applying two methods. One is the elbow point which is a common method in the K-means clustering for finding a reasonable number of clusters and it was explained in the chapter of modeling and methodology. The second, which is proposed in this research, includes the error of estimating the risk due to the number of clusters which practically is equal to the size or reduced dataset. To do so, the average error in the estimation of loss for selected 3 return periods of 475, 975, and 2475 years by the reduced datasets with the comparison with the true dataset for each clustering number is calculated and plotted versus the number of clusters. It is anticipated that, like the Elbow point method, the sudden change in decreasing slop of this plot will be the optimum point and the result of selecting the number of clusters based on this anticipation is tested in this chapter. Here the results of the latter are the one that is used more because it is related to the error in the estimation of loss values. Thus, when an optimum number of clustering is mentioned in this section, it means the results of the second loss (risk error-based) method.

In the following subsections, the resulted loss exceedance curve with the reduced dataset is compared with the loss exceedance curve of the whole dataset and the unmodified risk rate of the reduced dataset as well. This comparison is shown for the applied 5 loss metrics in this study. Also, the effect of the number of clusters on the accuracy of the results is evaluated by implementing these comparisons for 3 different clustering numbers of the minimum, optimum, and maximum numbers of clusters in this study.

Moreover, to check what happens to hazard curves generated with the reduced dataset, in each part also the hazard curve of the reduced dataset without and without including the modification rate in the reduced dataset is presented and compared with the hazard curve with the whole dataset. This comparison is carried out and compared for 3 different clustering numbers: minimum, optimum, and the maximum clustering number that was performed in this study.

At the end of this chapter, the reduced dataset by k-means clustering in the Risk-Consistent approach was used to assess the seismic risk in case of a hypothetical retrofit plan for the network. Therefore, 15 bridges out of 87 bridges in our modeled road network were randomly selected and their performance was improved in a way that they will never damage in the next run of the loss modeling. The results are shown and compared with the whole dataset for two loss metrics LDB and DWCL.

Presenting results of k-means clustering here is in the order loss metric as LDB, TD, SCL, WCL, and DWCL. For each loss metric 3 different approaches of the Risk-Consistent, Weighted Risk-Hazard, and Unweighted Risk-Hazard are studied, and the results of all in addition to the Hazard-Consistent approach are shown in similar clustering numbers that are used in the Risk-Consistent approach for each loss metric.

4.2.1 Results for Total Length of Damaged Bridges (LDB)

Here reduction approaches were tested for the case of considering the total length of the damaged bridge as the effective loss metric in clustering. As it was explained in the second chapter, the total length of damaged bridges is a direct/physical loss to the network and can be converted to the direct economic loss by having the average construction cost of the bridge per length. In the following, the order of presenting the results is the Risk-Consistent approach, Weighted Risk-Hazard, Unweighted Risk-Hazard, and Hazard-Consistent approach. This subsection will be wrapped up with a discussion part to compare the results by applying LBD in 4 different approaches.

4.2.1.1 Risk-Consistent approach with LDB Loss Metric

In this part, the Risk-Consistent approach was adopted for the case of LDB as the loss metric. The order of the following plots is first the elbow point method in finding the reasonable number of clusters and second is the mentioned risk-error-based method of estimating an optimum clustering number. The loss exceedance curve generated by reduced dataset with this clustering method will be compared with the reference risk curve which is the one that has been generated with the whole package. The effect of different number of clusters in the accuracy of risk results is compared by 3 loss exceedance plots with 3 different clustering numbers (min, optimum due to the risk error and the maximum number that we performed in this study). At the end the resulted hazard curve with the reduced data set is shown and compare it with the hazard result of the whole dataset. hazard plots are also shown for different clustering number to compare their accuracy.

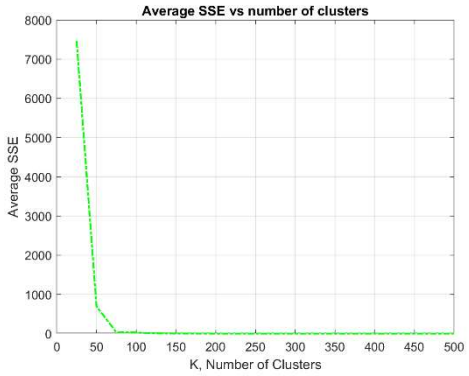


Figure 4-3Elbow point for selecting the K (for Case LDB and Risk-Consistent)

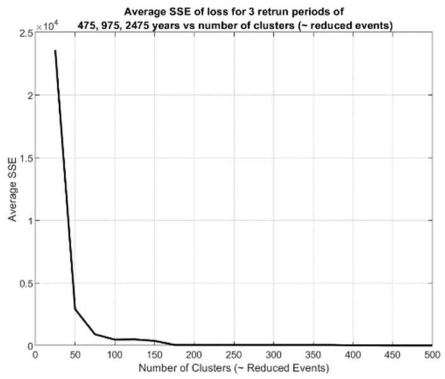


Figure 4-4 the Average SSE of loss vs Clustering No. for 475 ,975,2475 (for Case LDB and Risk-Consistent)

As it is shown in Figure 4-3 the estimated of acceptable clustering number is about 50 clustering but according to Figure 4-4 these values, from the risk error aspect, it is estimated as 150.

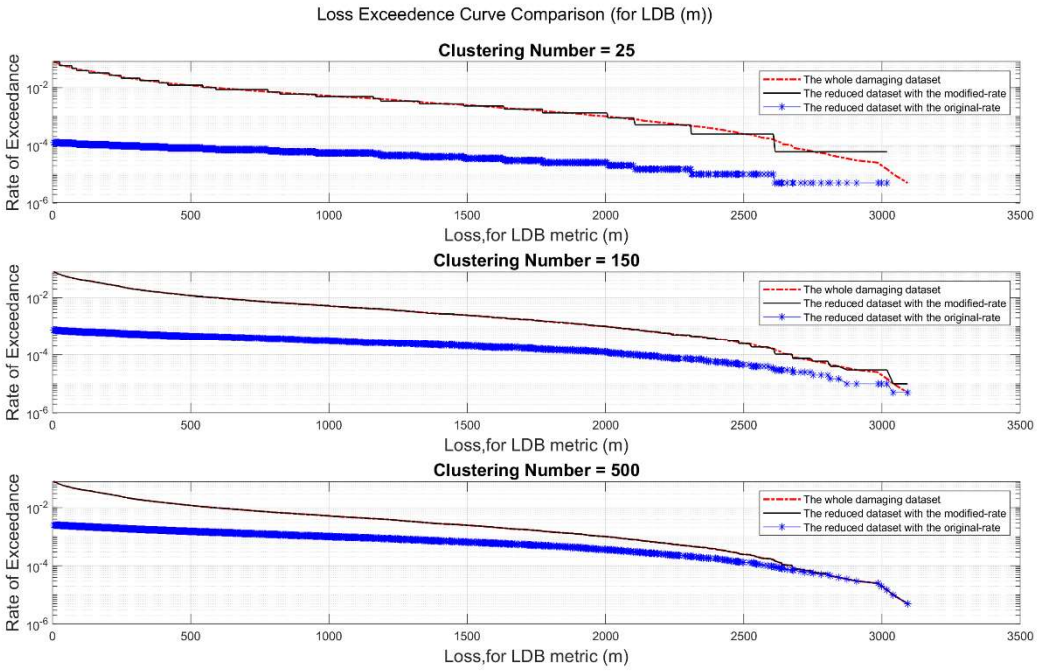


Figure 4-5 Comparison of LECs for reduced dataset with min, error based optimum, max number of clusters (for Case LDB and Risk-Consistent), (unit of y axis is 1/year).

Figure 4-5 shows the comparison of resulted loss exceedance curve for reduced data set (the black curve) with the whole data set (the red curve), the effectiveness of applying the modification for the rate of representative earthquake scenarios is in increasing the accuracy of the risk result with the reduced dataset can be observed (the blue curve). As a reminder, the modification here is the sum of the occurrence rate of all the members in the clusters that we use as the occurrence rate of the representative earthquake event. This figure also shows the effect of clustering number in the accuracy of the results and the optimum number of clustering predicted by the risk error curve shows a satisfactory accuracy.

Figure 4-6 to Figure 4-10 show the resulted hazard curves with the reduced dataset in 5 different bridge location in the area and compare it with their reference hazard curve with the whole data set of events.

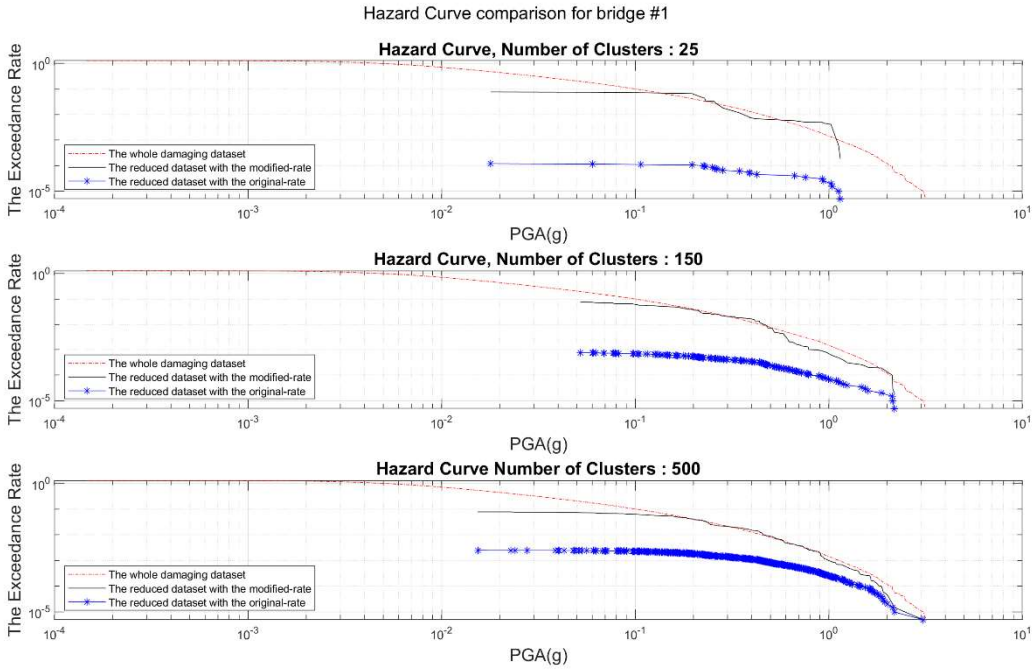


Figure 4-6 Hazard Curve Comparison for Bridge number 1 (for Case LDB and Risk-Consistent), (unit of y axis is 1/year).

as it can observe, despite the good compatibility of the resulted loss exceedance curve (LEC) with the reference LEC curve even in few numbers of clustering (which lead few numbers of events in the reduced dataset) and covering a large range of loss values, the hazard curves are not accurate and mostly underestimate the hazard risk for the selected locations. This is a result of applying Risk-Consistent approach. And by comparing the trend of increasing in the accuracy of the results by increasing the number of clusters, it can be concluded that, in applying Risk-Consistent approach in our data reduction, to get both Risk-Consistent and Hazard-Consistent results it is needed to increase the number of events in the reduced dataset.

Here to judge the sufficiency of the reduced dataset in estimating the hazard estimation a measure is applied which is from a loss perspective. In this measure, the compatibility of the generated hazard curves for the location of the selected bridges by the reduced dataset and the whole dataset, the associate PGA values for 3 different return periods of 475, 975, and 2475 years are read from both curves of the reduced dataset and the whole set. Then, for derived PGA values, the probability of the exceedance of a damage level of interest (in this study collapse) is calculated by the related fragility function for the bridge. The error is calculated by averaging the error in the probability of exceedance for the case of each return period, as shown in equation (1).

$$Error = \frac{1}{I}\sum_{i=1}^{i=I}\frac{\left(P\left(EDP \geq d | IM_{T_{i,Ref}}\right) - P\left(EDP \geq d | IM_{T_{i,Reduced\ Set}}\right)\right)}{P\left(EDP \geq d | IM_{T_{i,Ref}}\right)} \quad (1)$$

Where, $P\left(EDP \geq d|IM_{T_{i,Ref}}\right)$ is the probability of exceedance of desired damage level for the derived Intensity Measure (IM) from the reference hazard curve for the i_{th} return period and $P\left(EDP \geq d|IM_{T_{i,Reduced\ Set}}\right)$ is the probability of exceedance of the desired damage level for the derived IM value from the generated hazard curve by the reduced dataset. I is the total number of considered return periods to check the error which is 3 here (475, 975, and 2475 years).

As an example of checking the compatibility of the derived hazard curves with equation 1, this error is calculated for hazard curve in for the derived hazard curves in Figure 4-10 for the location of bridge number 5. The values of PGAs for 3 mentioned return periods are read from reference hazard curve which is from the whole dataset and the hazard curve derived from the reduced dataset. it is concluded that, for the reduced dataset by the Risk-Consistent approach for the LDB loss metric, for the clustering number of 25 (size of reduced dataset is 25), this error of predicting the probability of exceedance of the desired damage level by the pertinent fragility of the bridge is 68.7%, for the clustering number of 150 (size of reduced dataset is 150), this error is 45%, and for the clustering number of 500 (size of the reduced dataset is 500), this error decreases to 32% which is still a significant error. According to the applied fragility curve (see chapter of modeling and methodology), the range of probability of damage for the pertinent PGA values of theses 3 return periods that are obtained from the reference hazard curve (the red curve) is 59%-95%.

These derived error values show by Risk-Consistent approach the accuracy of hazard curves are decreases and is not compatible with the reference hazard curve and to improve the hazard curve, the size of reduced dataset showed be increased. In the next part, the goal is to slightly increase the effect of hazard into the clustering to see its effect in the accuracy of both loss exceedance curve and hazard curve.

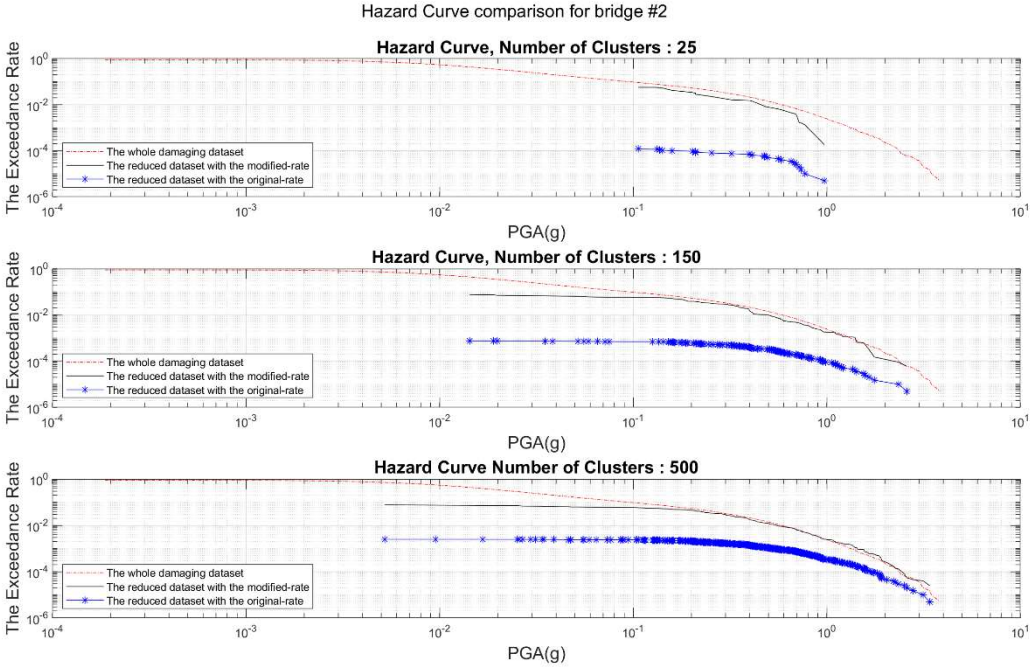


Figure 4-7Hazard Curve Comparison for Bridge number 2 (for Case LDB and Risk-Consistent), (unit of y axis is 1/year).

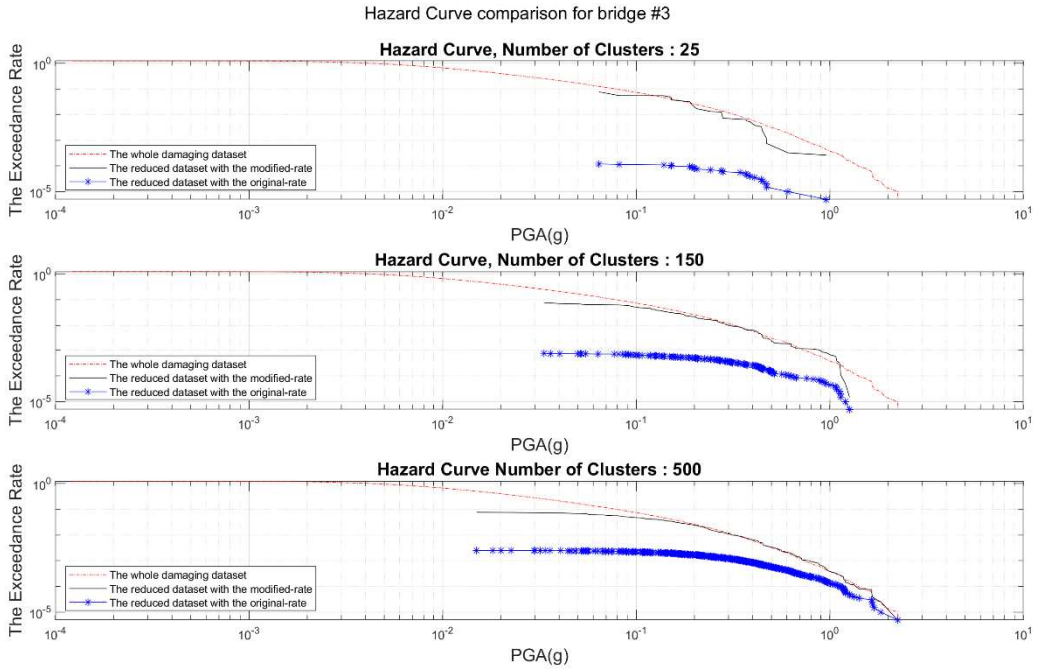


Figure 4-8 Hazard Curve Comparison for Bridge number 3 (for Case LDB and Risk-Consistent), (unit of y axis is 1/year).

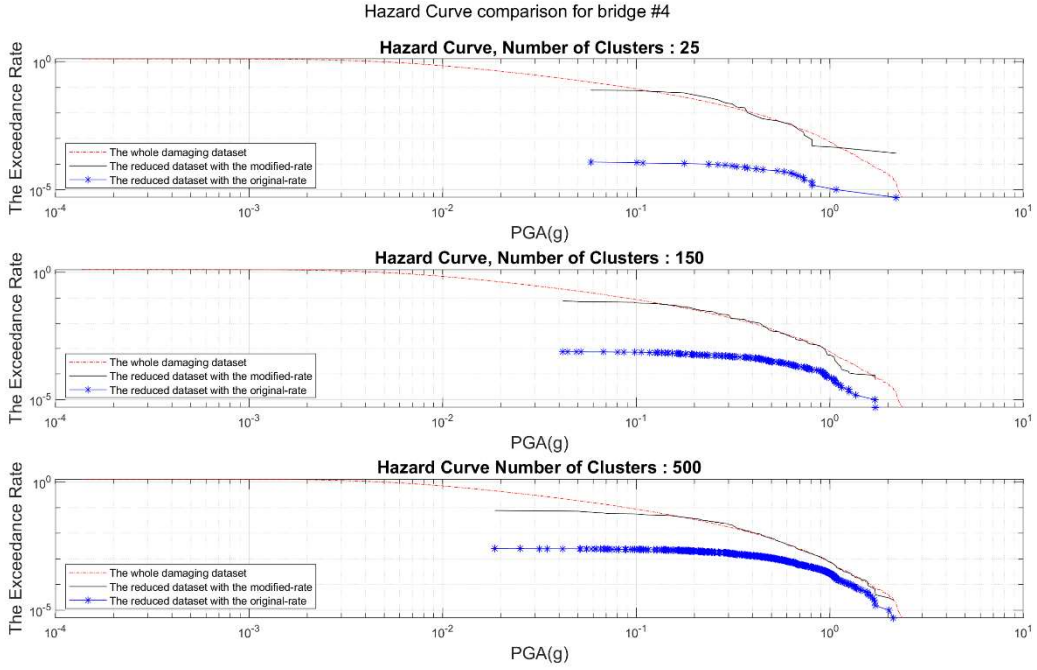


Figure 4-9 Hazard Curve Comparison for Bridge number 4 (for Case LDB and Risk-Consistent), (unit of y axis is 1/year).

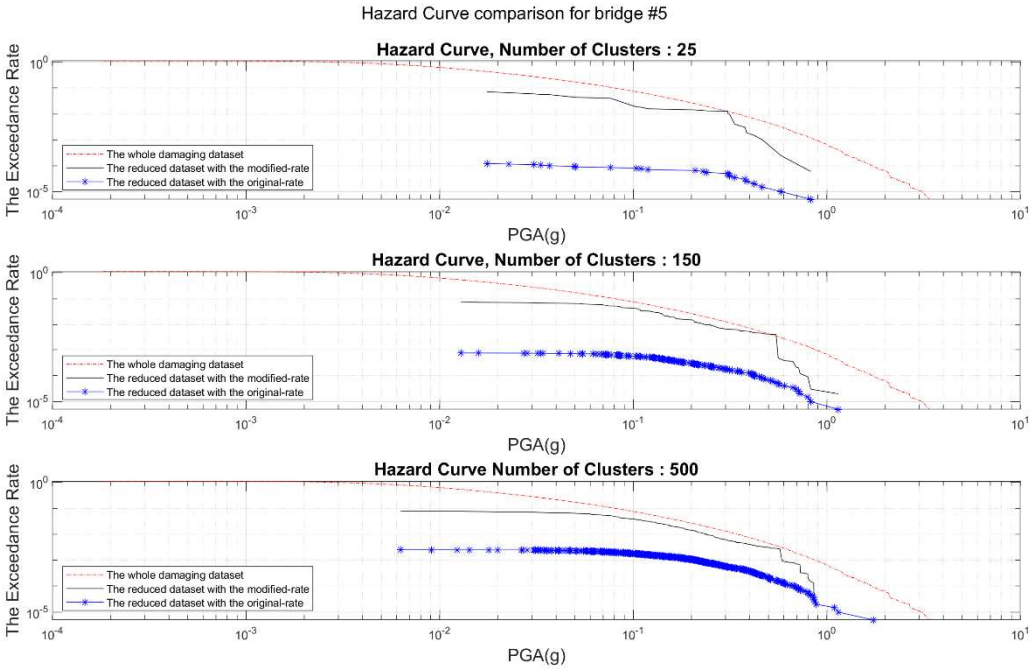


Figure 4-10 Hazard Curve Comparison for Bridge number 5 (for Case LDB and Risk-Consistent), (unit of y axis is 1/year).

4.2.1.2 Weighted Risk-Hazard approach with LDB Loss Metric

In this section for examining our approach for LDB which refers to the direct loss values, a combination of the loss values and intensity measure values (which is PGA in this study) is used in the clustering. Means that the dominancy of the risk is decreased by involving the hazard. To do so, 6 features for each clustering samples (our events) are used which one of them is the loss value of our loss metric of interest, here LDB, and the other 5 features are the values of PGAs in the location 5 bridges (those that their hazard curve will be studied here) all related to same event. Figure 4-11 shows the locations of these bridges on the network. This bridge locations are selected in a way that the hazard in each part of the network (north, east, west, south and center) be involved also these locations have different hazard curves. it should be mentioned that IM features for weighted clustering that are used here can be related to any important location of the network and can be even more than 5 locations if the dominancy of the loss remains 50%.

In calculation of distance metric in clustering, as mentioned in chapter of modeling and methodology, a wight of 0.5 is used for the loss value and 0.1 for each 5 PGA values. In following, the curves with the same clustering numbers that were used in the Risk-Consistent approach with LDB is used to provide a better comparison of the effect of this approach. Therefore, the results of elbow point or the risk error plot for this clustering method are not presented for this approach. Obviously, the application of these elbow point and risk error plots is possible for this approach in case of requirement of an estimation of optimum number of clustering for this apprach.

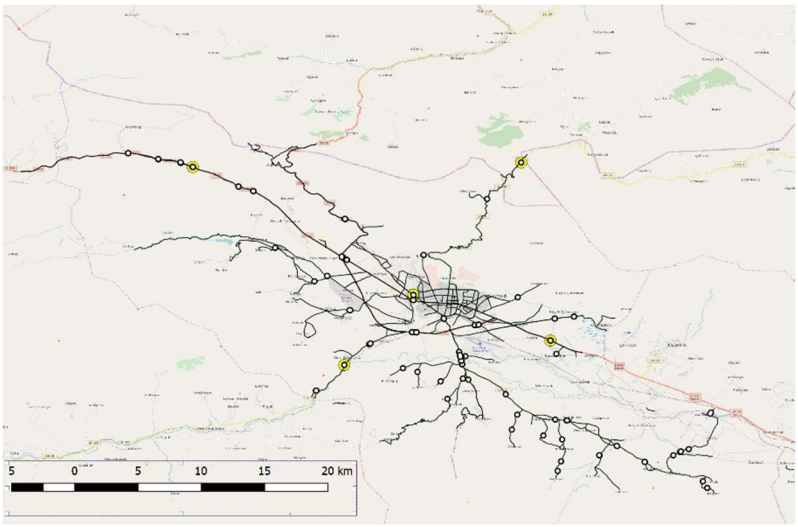


Figure 4-11 location of 5 selected bridges on the network (yellow markers). (Map data source: https://www.openstreetmap.org)

Figure 4-12 is about the results of the risk estimation with the reduced dataset of this approach. As it is observed in comparison with the Risk-Consistent approach, that even in a few numbers of clusters a wide range of loss values were covered by the curve, here more discrepancy is observed with the reference LEC for the value of LDB loss metric with the whole dataset in comparison with what was resulted with the Risk-Consistent approach. This is what was expected and the accuracy of few numbers of events in the reduced data set is sacrificed to get better values in hazard estimation. As it is shown in this figure that by increasing the number of clusters the risk error can be decreased. But still the curve with the 150 events in the reduced package, have a negligible error in estimating the loss values for the return periods of 475, 975 and 2475 years.

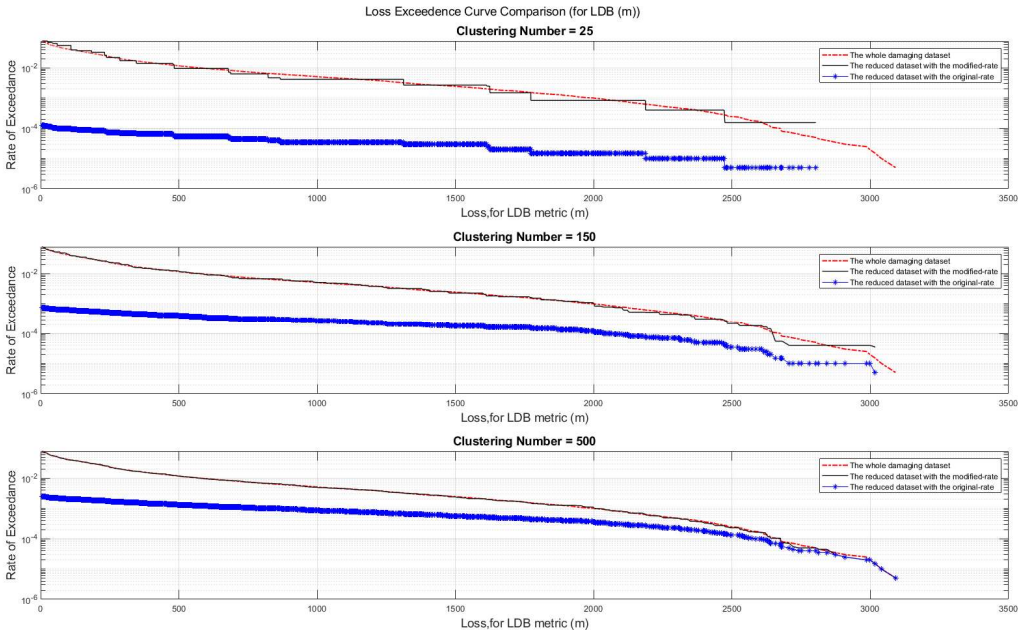


Figure 4-12 Comparison of LECs for reduced dataset with min, 150, max number of clusters (for Case LDB and Weighted Risk-Hazard), (unit of y axis is 1/year).

Figure 4-13 to Figure 4-17 show the hazard curves generated by the Weighted Risk-Hazard approach. Some improvements can be observed especially in the less frequent hazards and the error in frequent hazard are relatively significant. Again, the trend of decrees in error by increasing the clustering number is observed, that is because of our initial step in reduction by removing all the not damaging events.

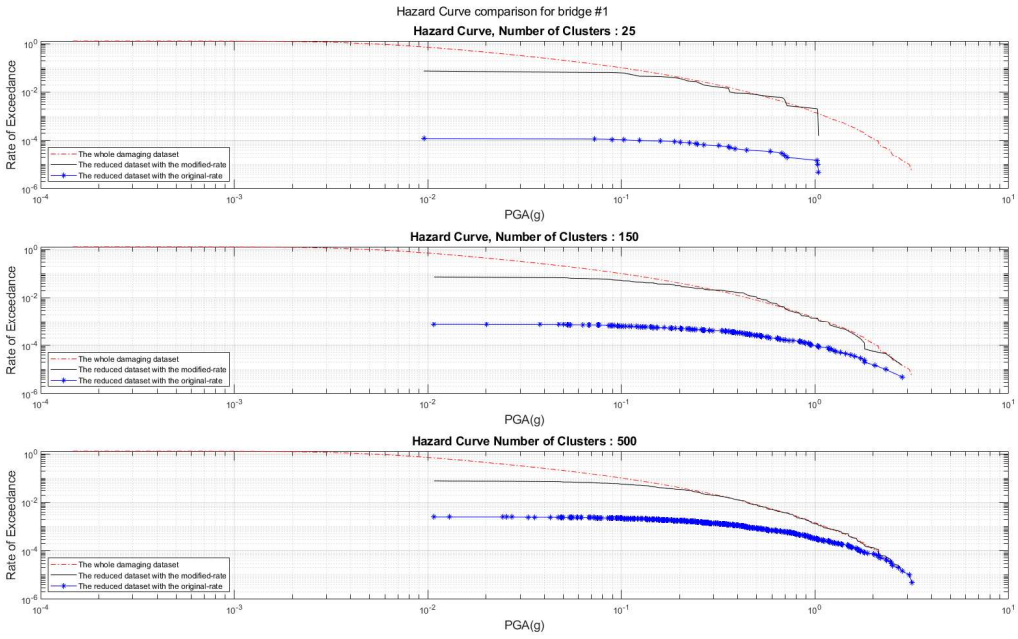


Figure 4-13 Hazard Curve Comparison for Bridge number 1 (for Case LDB and Weighted Risk-Hazard), (unit of y axis is 1/year).

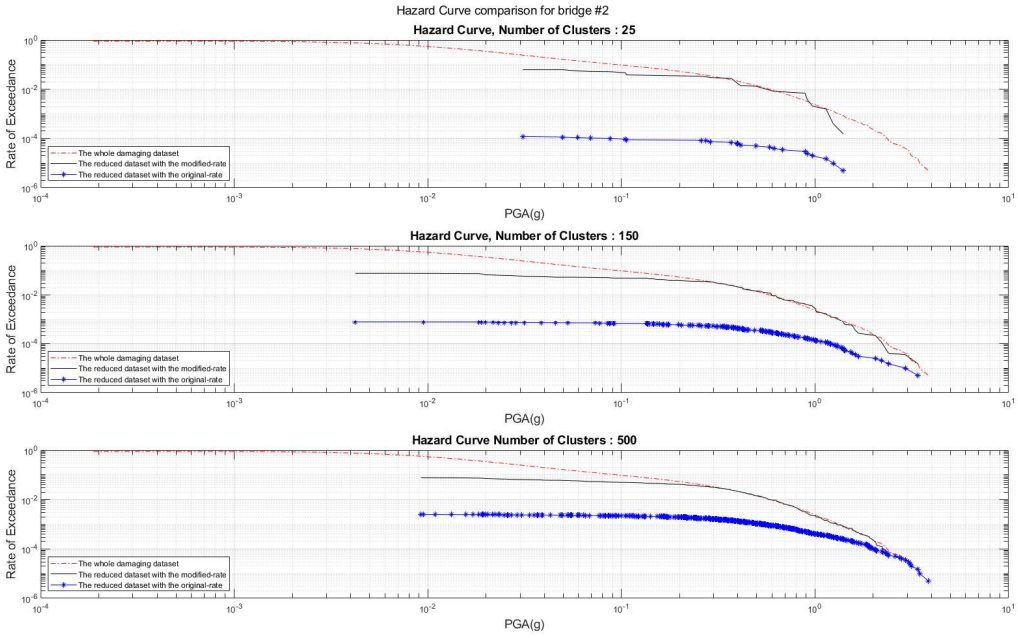


Figure 4-14 Hazard Curve Comparison for Bridge number 2 (for Case LDB and Weighted Risk-Hazard), (unit of y axis is 1/year).

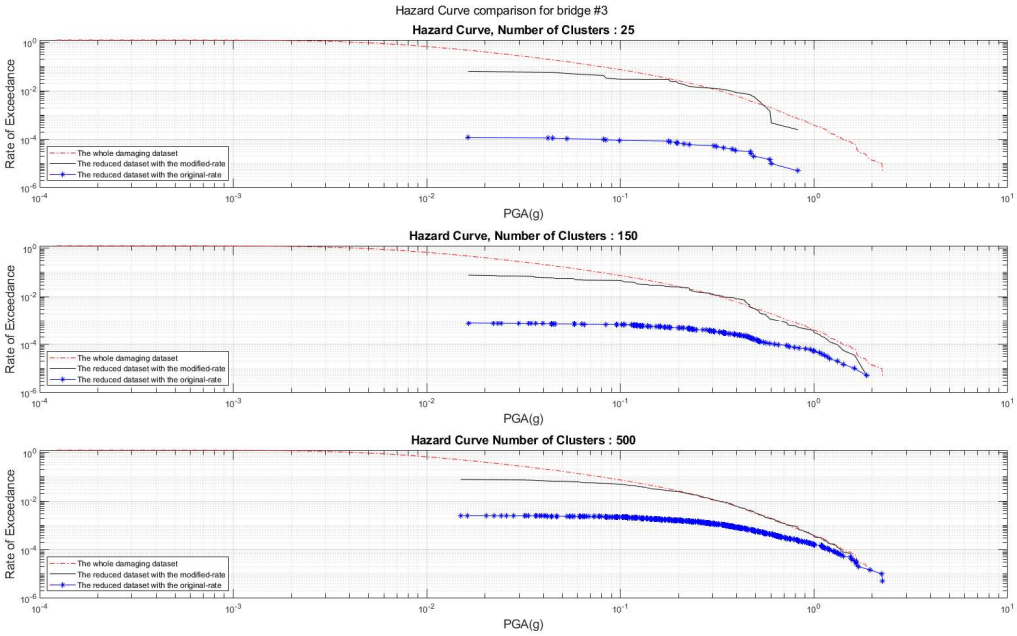


Figure 4-15 Hazard Curve Comparison for Bridge number 3 (for Case LDB and Weighted Risk-Hazard), (unit of y axis is 1/year).

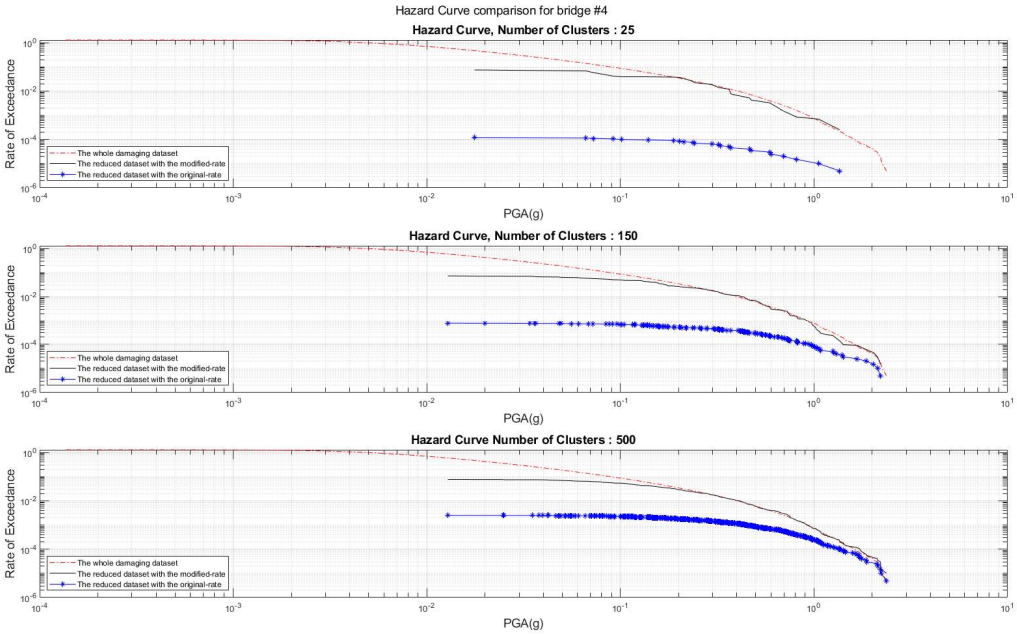


Figure 4-16 Hazard Curve Comparison for Bridge number 4 (for Case LDB and Weighted Risk-Hazard), (unit of y axis is 1/year).

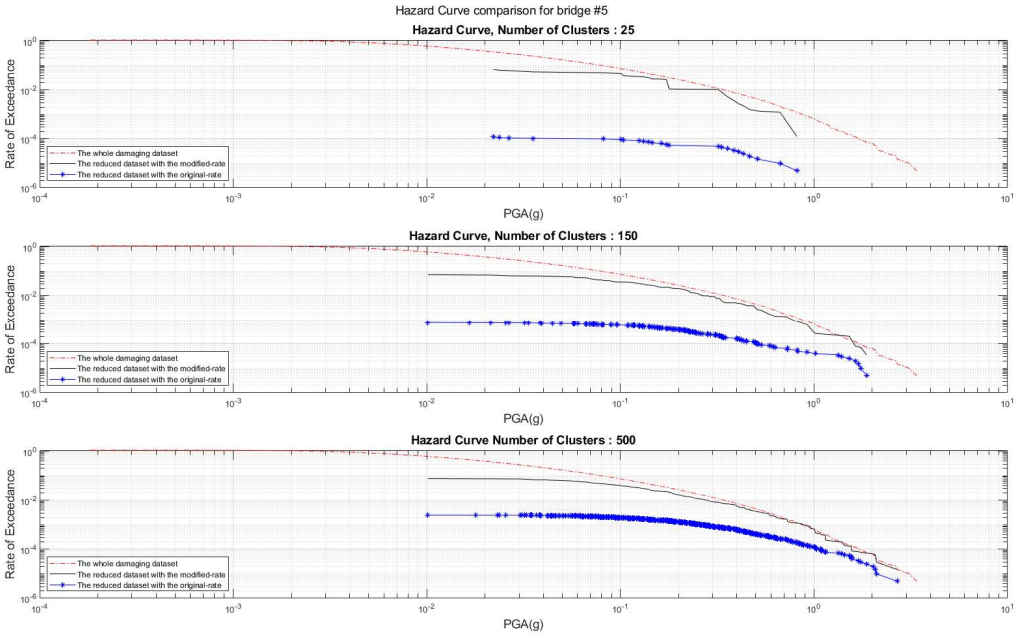


Figure 4-17 Hazard Curve Comparison for Bridge number 5 (for Case LDB and Weighted Risk-Hazard), (unit of y axis is 1/year).

In this section the dominancy of the risk in clustering was slightly decreased by including the hazard effect. In next part, all the 6 features in clustering are going to be treated equally and the results will be compared.

4.2.1.3 Unweighted Risk-Hazard approach with LDB Loss Metric

In this part unlike the previous approach, loss values are not dominant in calculating the dissimilarity between events to make the less dissimilar group of events. The same features as the previous approach are used but without any weights included. Here again, the same clustering numbers that were presented in two previous approaches were used for this approach to provide a better comparison of the effect of approaches.

In Figure 4-18 the comparison of loss exceedance curve with the reference curve shows fewer range of loss values is seen in lower clustering numbers. More errors in estimation of less frequent losses in clustering number of 150 in comparison of two previous approaches.

Figure 4-19to Figure 4-23 shows the hazard curve generated by the dataset that has been reduced by applying Unweighted Risk-Hazard k-means clustering for loss metric case of LDB. A slightly better hazard curves than the weighted one is seen but much more better hazard curve than the Risk-Consistent method. Still the reduced dataset falls short in predicting frequent hazard and the reason for it is that at the first step of the reduction method of this research, all the events that do not cause any damage to at least one bridge were eliminated. Therefore, even by increasing the clustering number, there are errors in that part of the hazard curves (lower PGAs) for all 5 bridge locations.

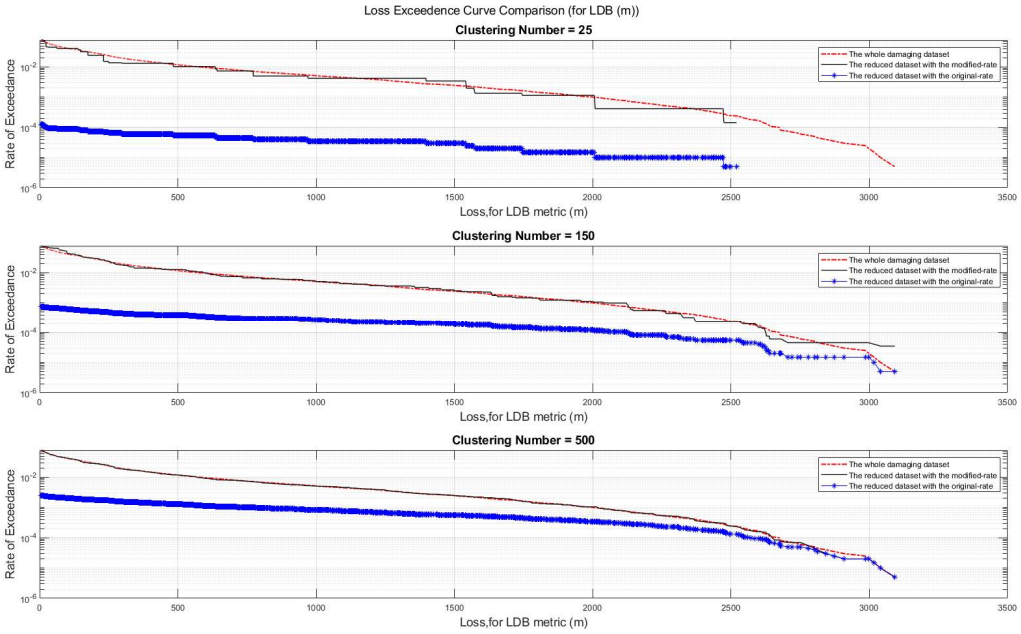


Figure 4-18 Comparison of LECs for reduced dataset with min, 150, max number of clusters (for Case LDB and Unweighted Risk-Hazard), (unit of y axis is 1/year).

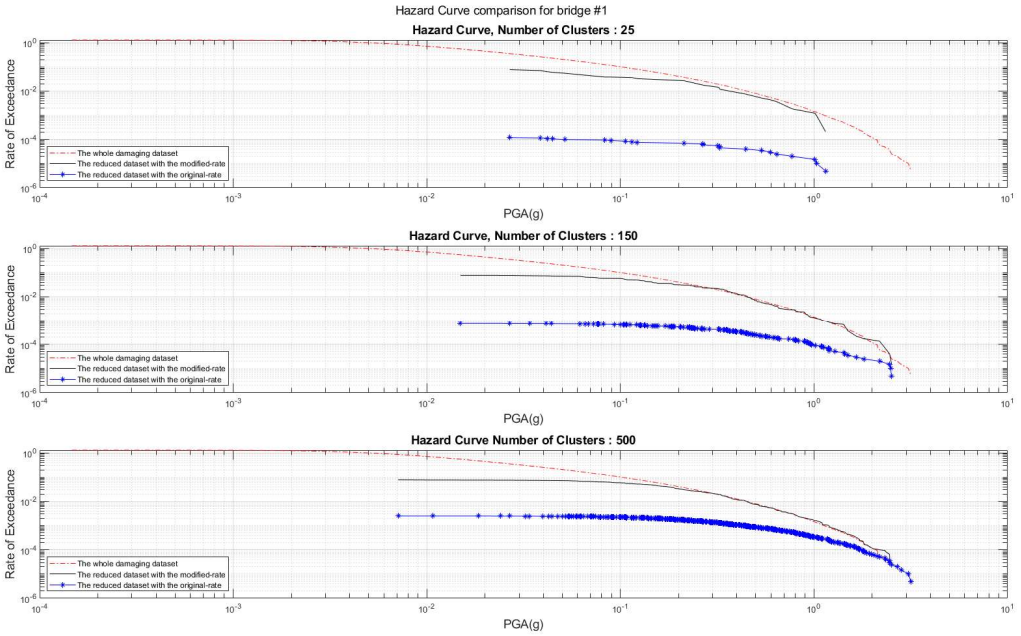


Figure 4-19 Hazard Curve Comparison for Bridge number 1 (for Case LDB and Unweighted Risk-Hazard), (unit of y axis is 1/year).

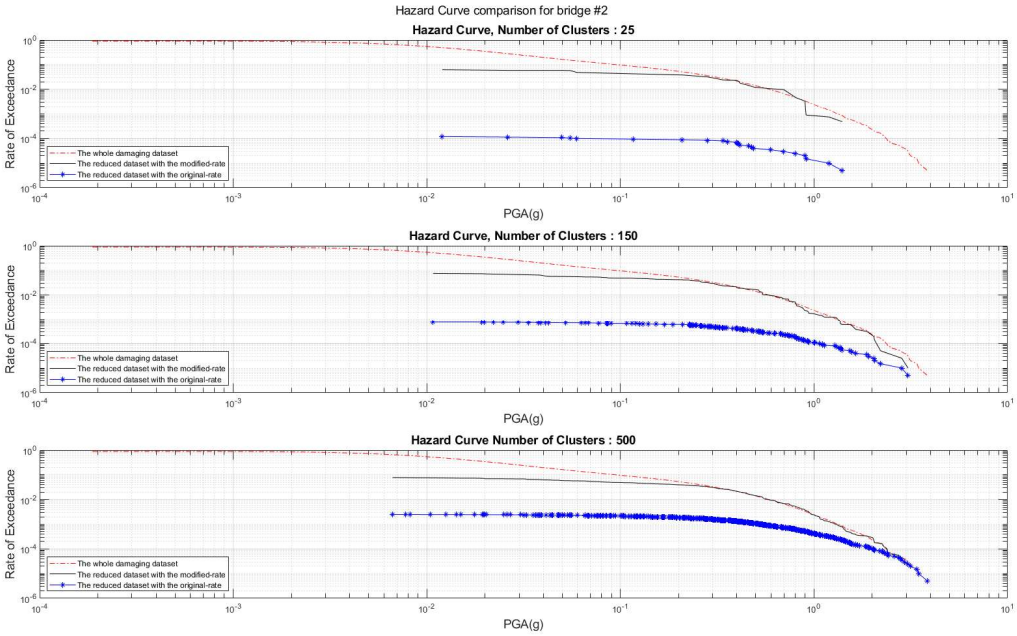


Figure 4-20 Hazard Curve Comparison for Bridge number 2 (for Case LDB and Unweighted Risk-Hazard), (unit of y axis is 1/year).

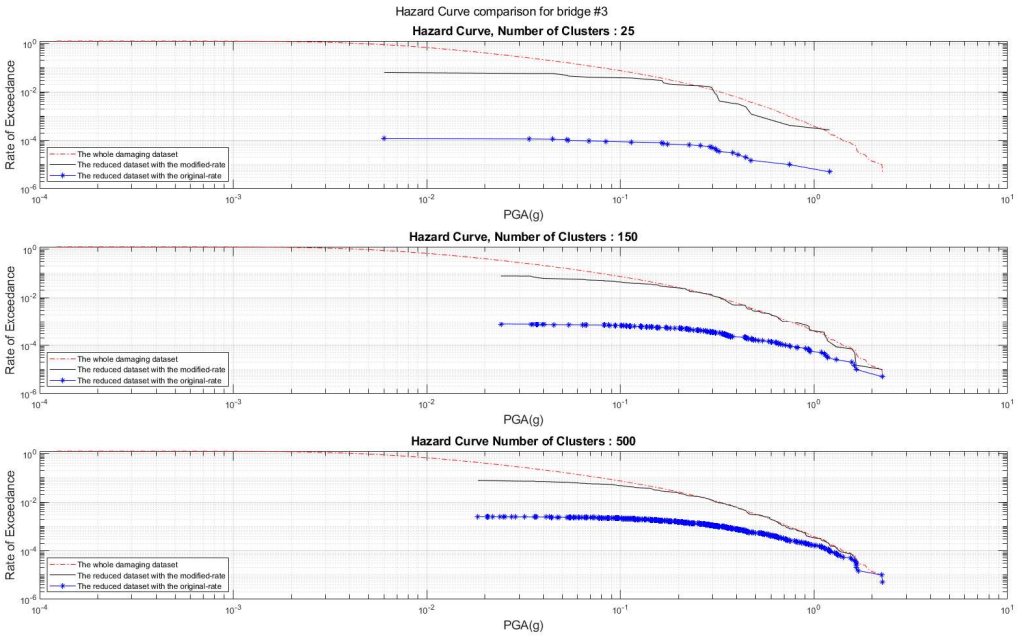


Figure 4-21 Hazard Curve Comparison for Bridge number 3 (for Case LDB and Unweighted Risk-Hazard), (unit of y axis is 1/year).

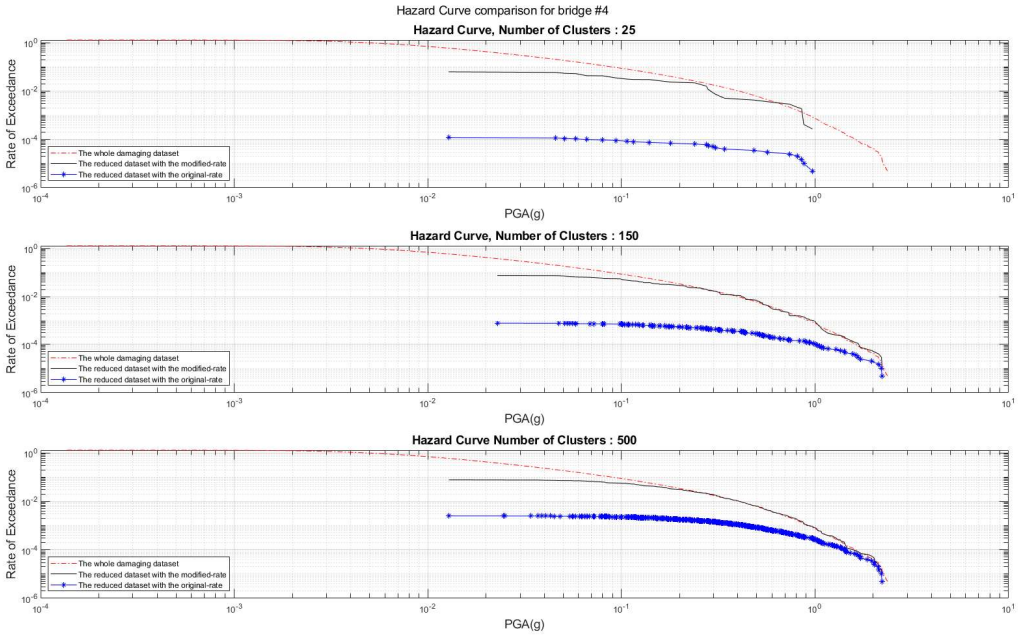


Figure 4-22 Hazard Curve Comparison for Bridge number 4 (for Case LDB and Unweighted Risk-Hazard), (unit of y axis is 1/year).

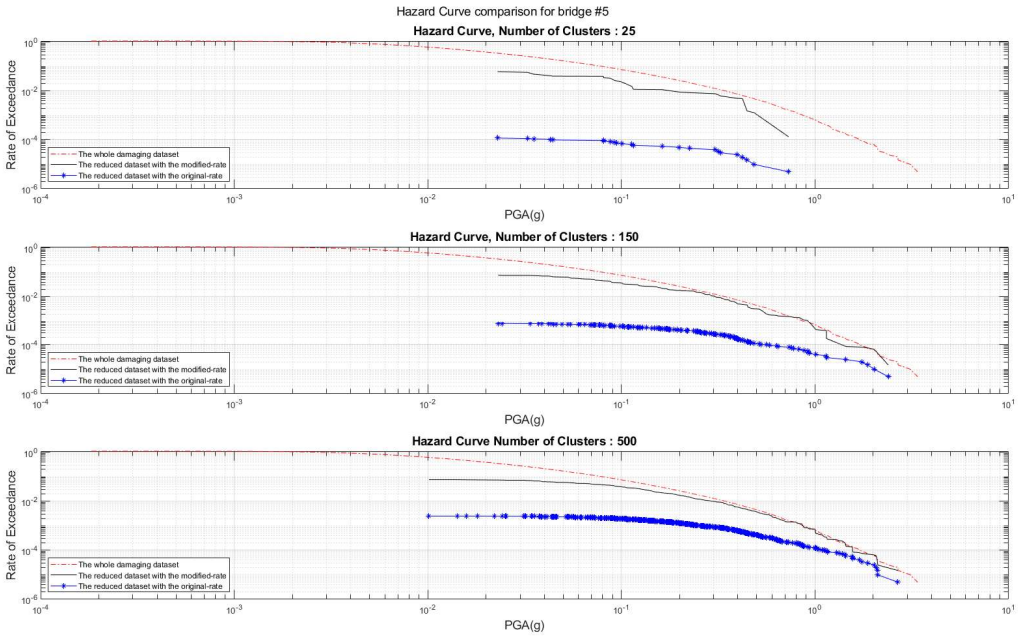


Figure 4-23 Hazard Curve Comparison for Bridge number 5 (for Case LDB and Unweighted Risk-Hazard), (unit of y axis is 1/year).

In this section the effect of treating risk and hazard`s effect on clustering equally was tested and as what was expected, the accuracy of loss exceedance carve were decreased with fewer number of events in the reduced data set and in return relative to Risk-Consistent method, here a better estimation for less

frequent hazards were resulted although there are still some errors especially with 150 number of events. In the next part, only the effect of hazard is considered in the clustering by k-means, and a better consistency with the reference hazard curve in fewer number of clusters is expected.

4.2.1.4 Hazard-Consistent approach with clustering number like 3 previous approaches with LDB

In this chapter, the 5 PGA values are considered as the effective features in the clustering. As it was mentioned before, due to elimination of all non-damaging earthquakes scenarios in the first step of this reduction method, therefore, the effect of loss is somehow included in this reduction as well.

Figure 4-24 shows the comparison of the resulted loss exceedance curves, and many loss values are missing in less clustering numbers and the estimated loss exceedance curve with the 150 events in the reduced dataset shows underestimations in estimating less frequent loss values for example the one with the return period of 2475. Still the reduced event set with 500 events shows relatively reasonable results for this approach which the loss values were not directly included.

Figure 4-25 to Figure 4-29 shows the results of hazard with the Hazard-Consistent approach and as it was expected there is some errors in predicting hazard of frequent and low values of PGAs which reflect the first step of data reduction.

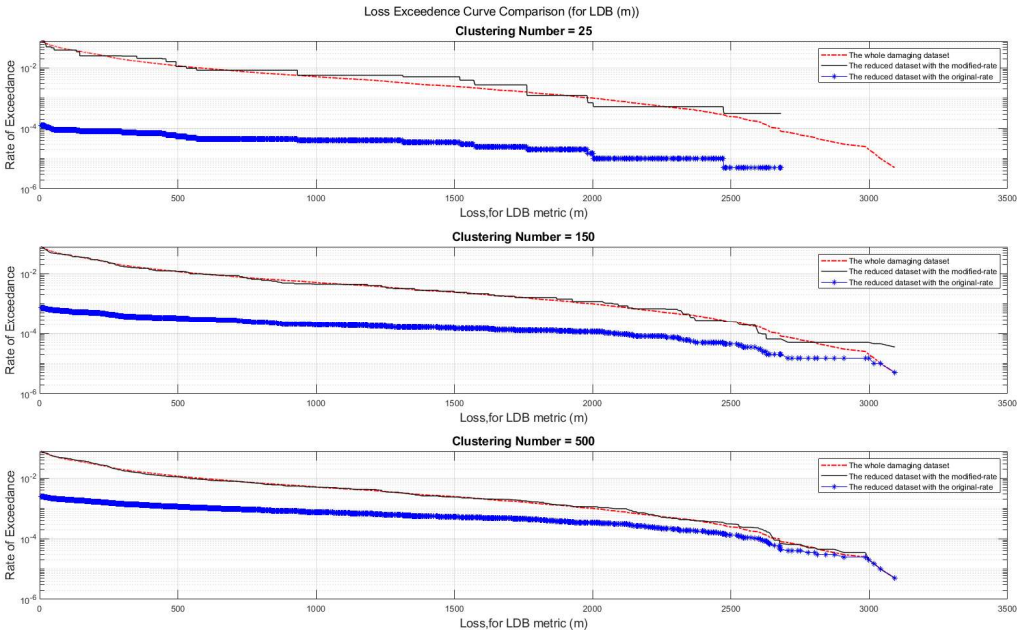


Figure 4-24 Comparison of LECs for reduced dataset with min, 150, max number of clusters (for Case LDB and Hazard-Consistent), (unit of y axis is 1/year).

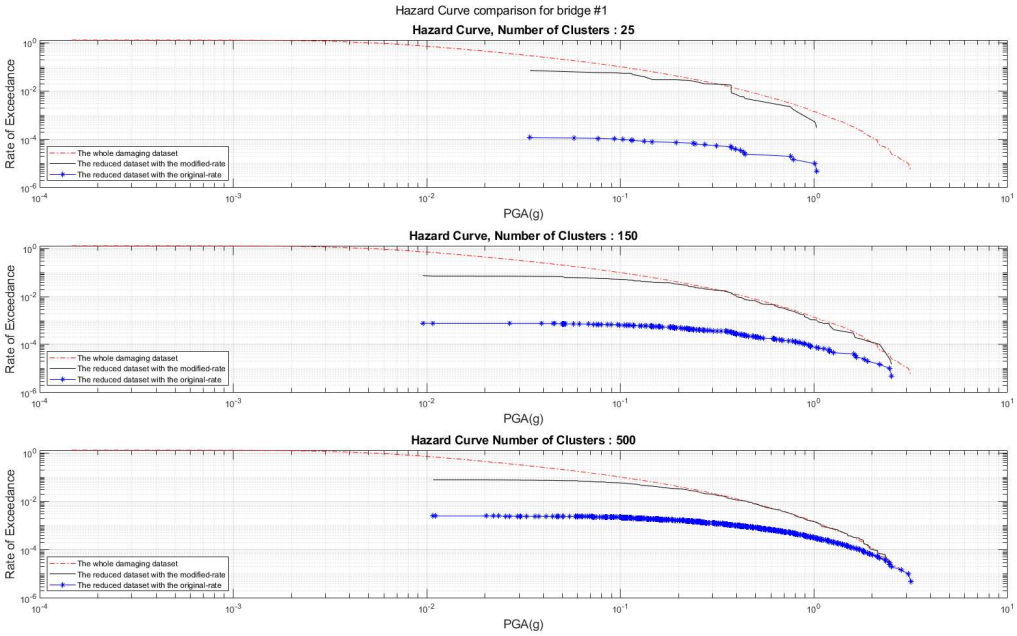


Figure 4-25 Hazard Curve Comparison for Bridge number 1 (for Case LDB and Hazard-Consistent), (unit of y axis is 1/year).

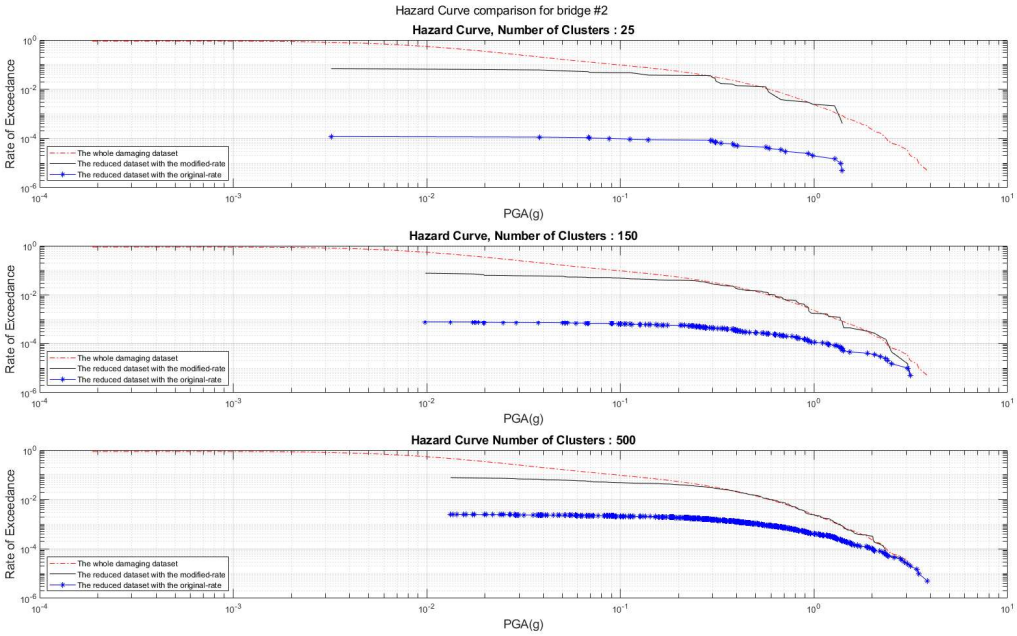


Figure 4-26 Hazard Curve Comparison for Bridge number 2 (for Case LDB and Hazard-Consistent), (unit of y axis is 1/year).

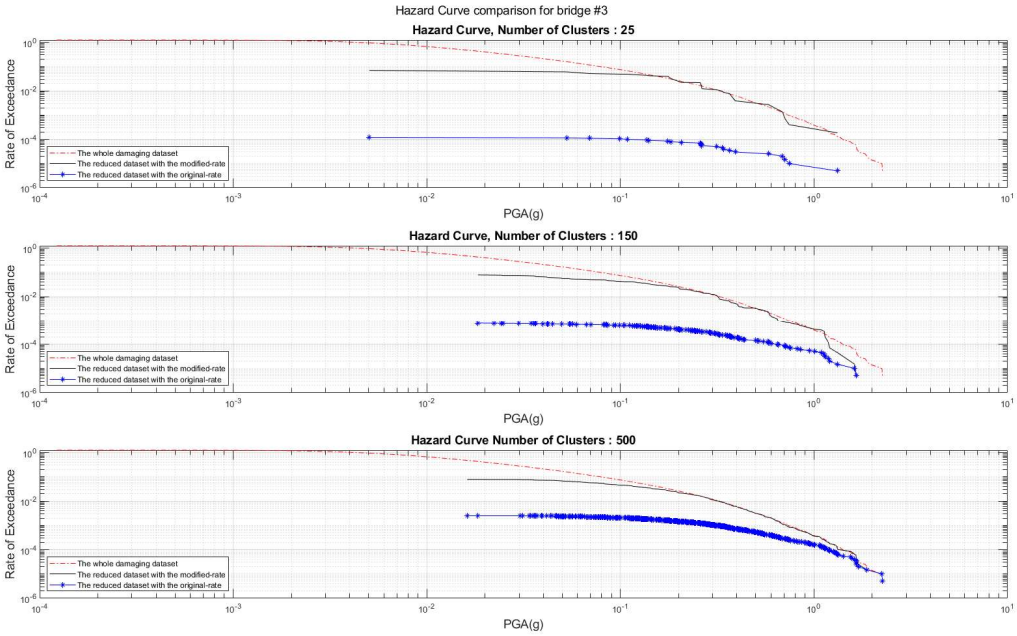


Figure 4-27 Hazard Curve Comparison for Bridge number 3 (for Case LDB and Hazard-Consistent), (unit of y axis is 1/year).

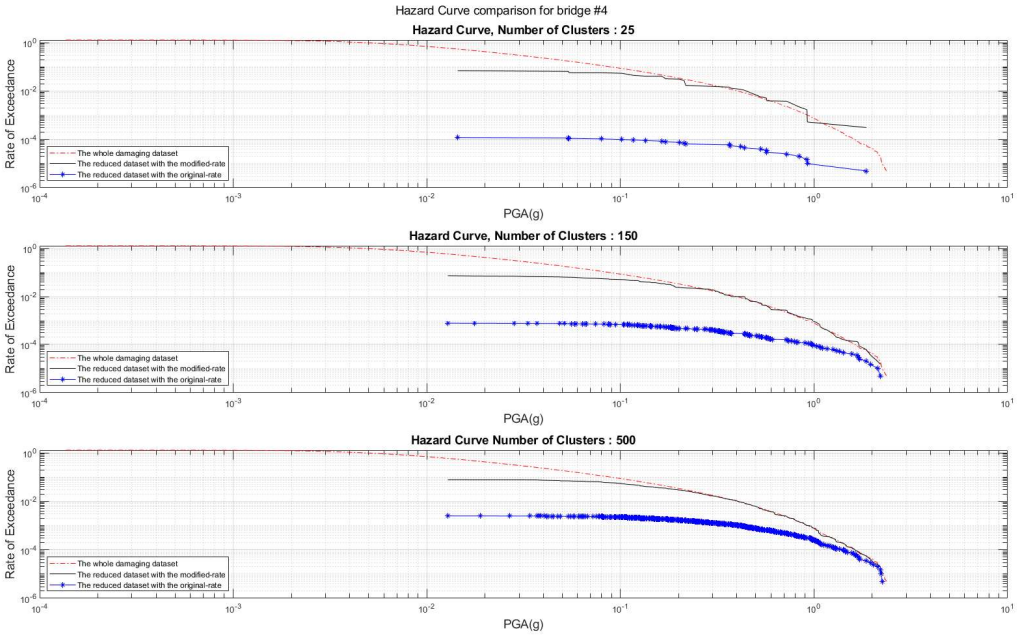


Figure 4-28 Hazard Curve Comparison for Bridge number 4 (for Case LDB and Hazard-Consistent), (unit of y axis is 1/year).

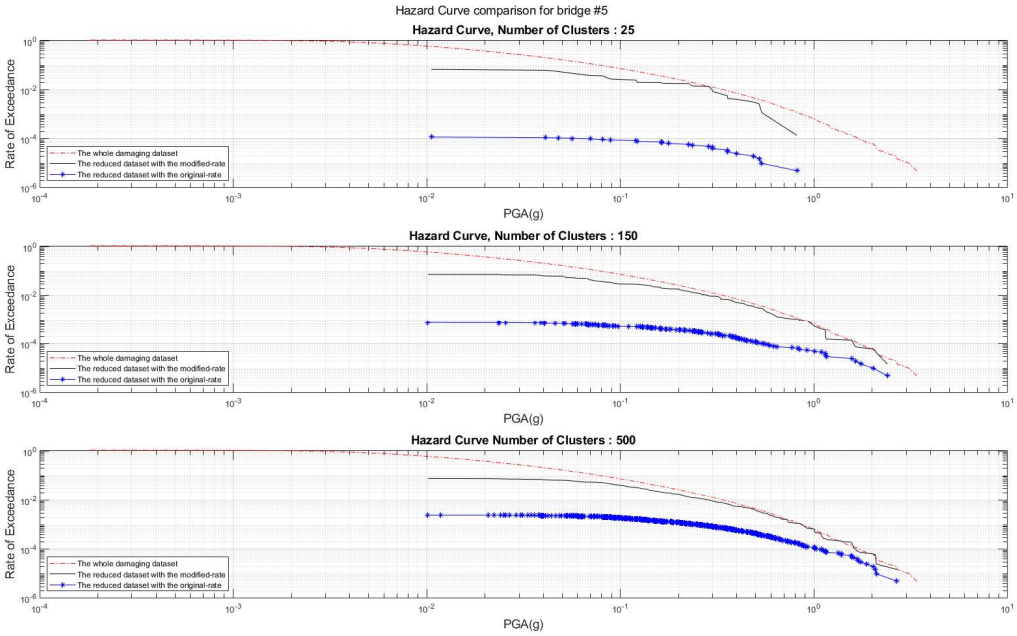


Figure 4-29 Hazard Curve Comparison for Bridge number 5 (for Case LDB and Hazard-Consistent), (unit of y axis is 1/year).

4.2.1.5 Discussions on the results of 4 different approaches in case of LDB

In this section the 4 different approaches with considering the case of direct damage loss metric were explicitly explained and the results were compared in detail in each subsection. Here we briefly review the observations of this sections. The comparison of the results of this section shows the expected pattern as by decreasing the effect of loss values in the clustering, the accuracy of the resulted risk curve is affected and to get better risk consistency greater number of events are needed in the reduced dataset.

As mentioned before, since the nondamaging earthquakes were removed from the event list at the first step of the reduction process, completely hazard consistent results are not achievable here. Based on the results of this section, it can be concluded that without including the effect of risk in clustering, risk errors are increased in fewer number of clustering like 150, and to get more accurate risk curves a greater reduced dataset, like 500 events is required.

4.2.2 Results for Total Time Delay (TD)

Time delay as a loss metric refers to the performance of the network and is a type of system level loss metric and is one of the applied loss metrics in this study. In this section to examine our approaches by different samples, the time delay loss metric is used as an effective loss value in the clustering. Again, 4 different approaches are adopted with the same order that were presented in the previous section as Risk- consistent, Weighted Risk-Hazard, Unweighted Risk-Hazard, and finally the Hazard-Consistent.

4.2.2.1 Risk-Consistent approach with TD

Like the same approach in the previous loss metric, here also 2 methods are applied to estimate the optimum number of clusters to get satisfactory results. One method is the common elbow point as shown in Figure 4-30 method in the k-means clustering and the other one is the risk error plotting that compare

the average errors in estimating the loss values associated with 475, 975, 2475 years return period are calculated and presented versus their pertinent clustering number as is shown in Figure 4-31.

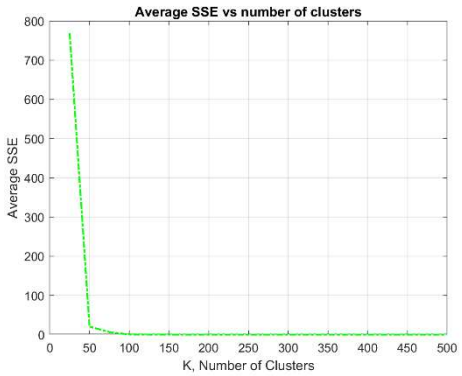


Figure 4-30 Elbow point for selecting the K (for Case TD and Risk-Consistent)

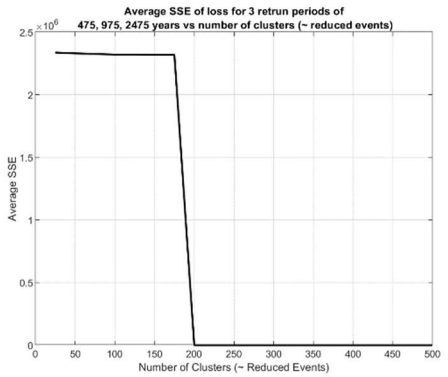


Figure 4-31 Average SSE of loss vs ClustNo for 475 ,975,2475 (for Case TD and Risk-Consistent)

Figure 4-32 shows the results of loss exceedance curve and compares the results of reduced dataset with the reference curve also compares the effect of increasing the number of clusters in the accuracy of our risk results. As shown here, like the previous section, in Risk-Consistent approach, even with a few numbers of clusters a complete range of loss values are remained in the reduced dataset. With the applied optimization risk error method in finding an optimum number of clustering, the clustering number of 200 shows satisfactory consistency with the reference risk curve.

The resulted hazard curve with the reduced dataset for 5 different bridges are shown in Figure 4-33 to Figure 4-37. In contrast to the accurate results for risk, the hazard results by this method for the time delay metric is not reliable due to the significant discrepancy even in the higher number of clustering like 500.

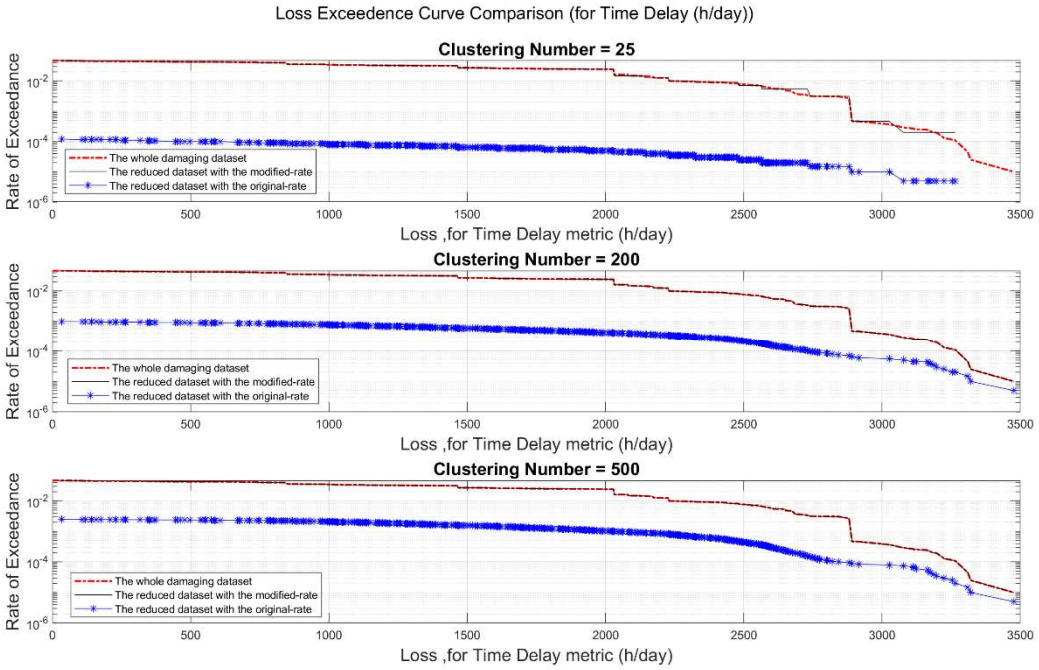


Figure 4-32 Comparison of LECs for reduced dataset with min, error based optimum, max number of clusters (for Case TD and Risk-Consistent), (unit of y axis is 1/year).

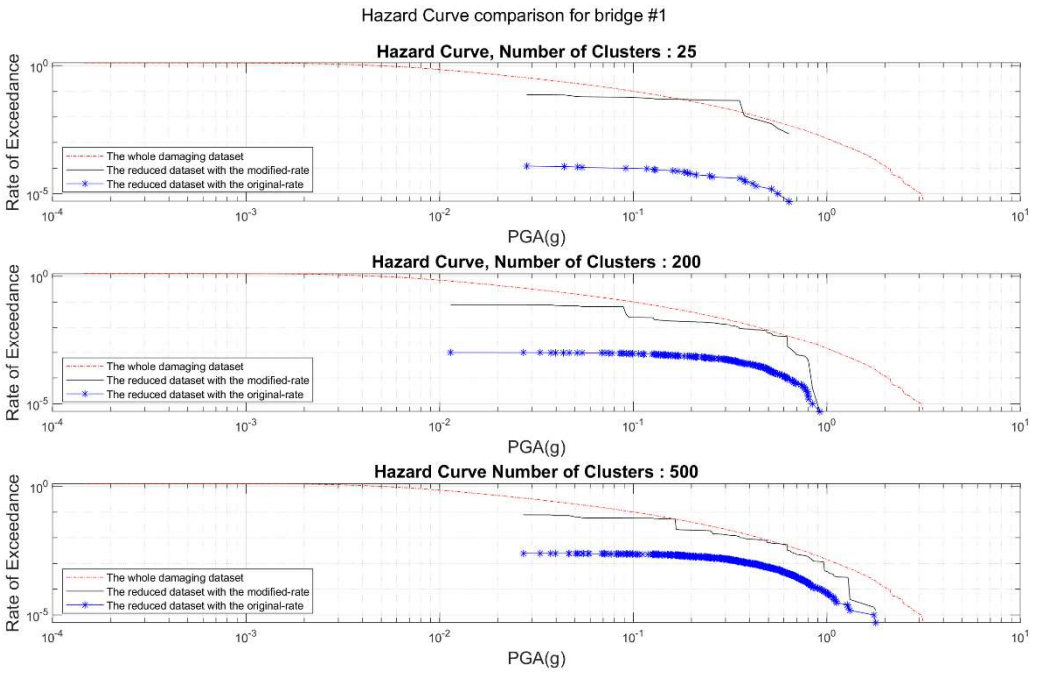


Figure 4-33 Hazard Curve Comparison for Bridge number 1 (for Case TD and Risk-Consistent), (unit of y axis is 1/year).

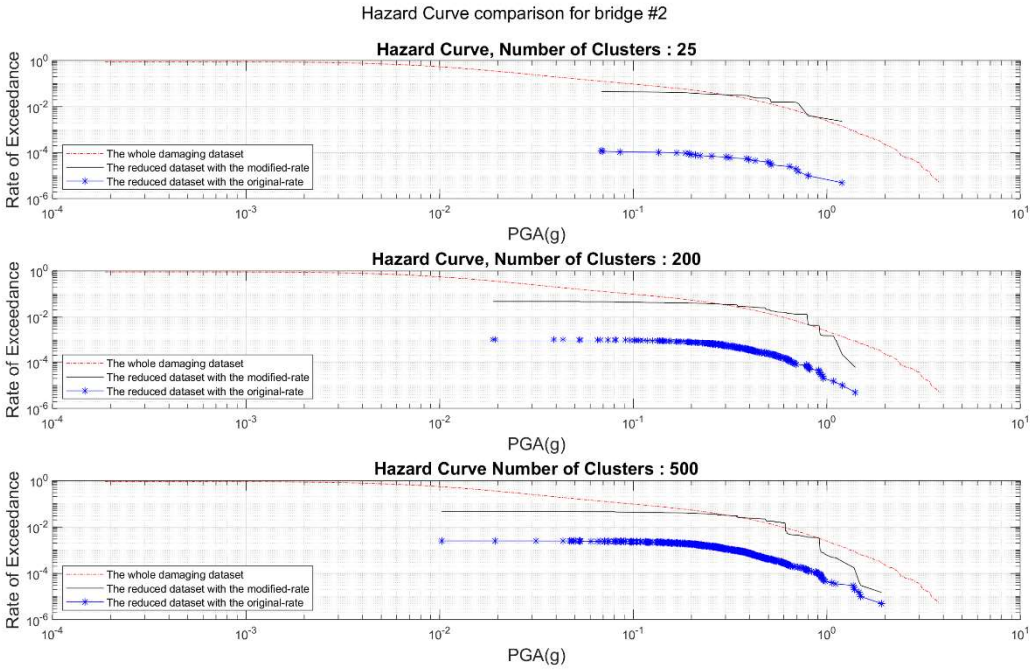


Figure 4-34 Hazard Curve Comparison for Bridge number 2 (for Case TD and Risk-Consistent), (unit of y axis is 1/year).

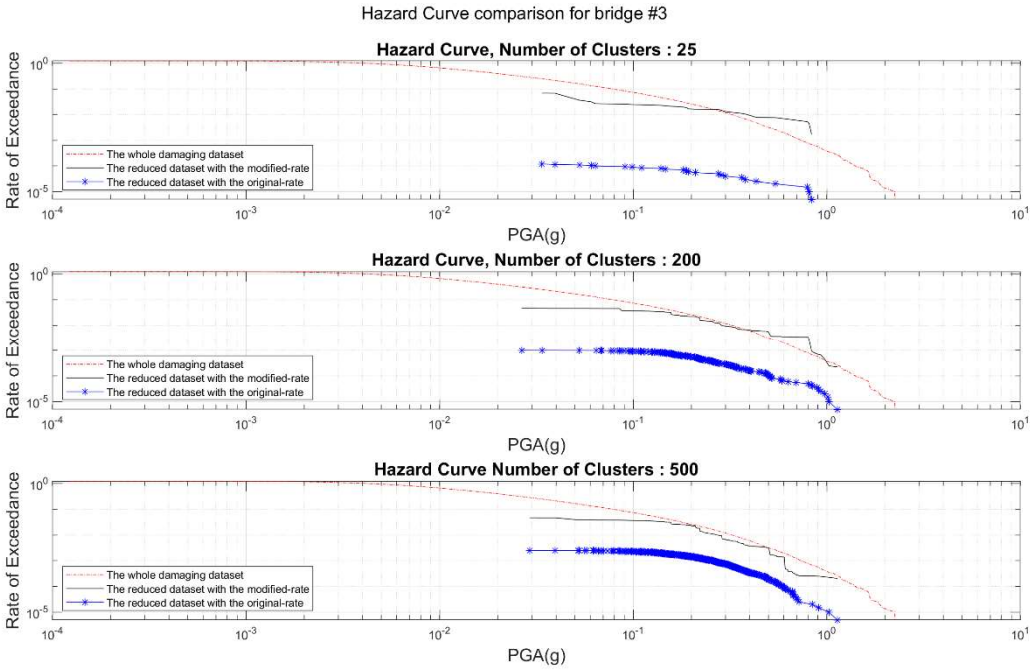


Figure 4-35 Hazard Curve Comparison for Bridge number 3 (for Case TD and Risk-Consistent), (unit of y axis is 1/year).

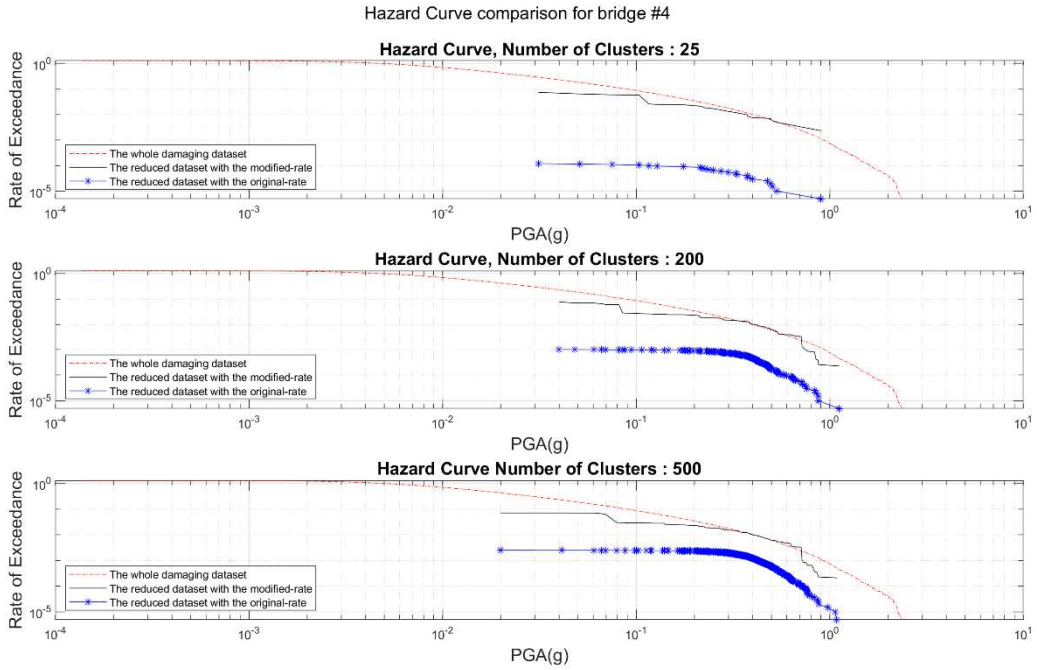


Figure 4-36 Hazard Curve Comparison for Bridge number 4 (for Case TD and Risk-Consistent), (unit of y axis is 1/year).

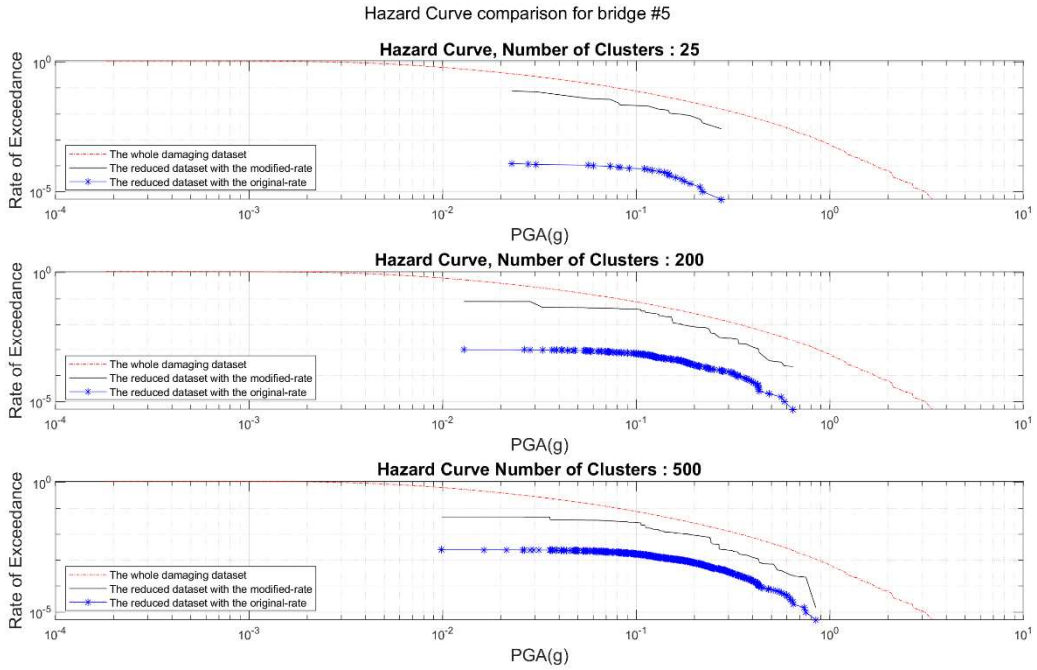


Figure 4-37 Hazard Curve Comparison for Bridge number 5 (for Case TD and Risk-Consistent), (unit of y axis is 1/year).

4.2.2.2 Weighted Risk-Hazard approach with TD

In this section like the previous loss metric case, a Weighted Risk-Hazard approach is applied to improve the hazard estimation. A weight of 0.5 is considered for the time delay loss values and 0.1 per each of 5 PGA values at the location of the selected bridges that hazard curve is also studied here.

Figure 4-38 shows the resulted risk curve with the reduced dataset by 3 different clustering number. As was expected and have seen in case of previous loss metric (LDB), The results are not as accurate as the Risk-Consistent approach and a greater number of events like 500 are required in the reduced dataset to get an almost accurate risk curve.

Figure 4-39 to Figure 4-43 shows the resulted hazard curve for 5 different locations with the reduced data set and compares it with the reference hazard curve also the effect of increasing number of clusters in accuracy of the results can be observed in these figures. As it shows in these curves, even by a slightly including the hazard effect, as the weight of them is less that the risk, good estimation for the less frequent hazards were resulted with using a large size of reduced dataset, like 500.

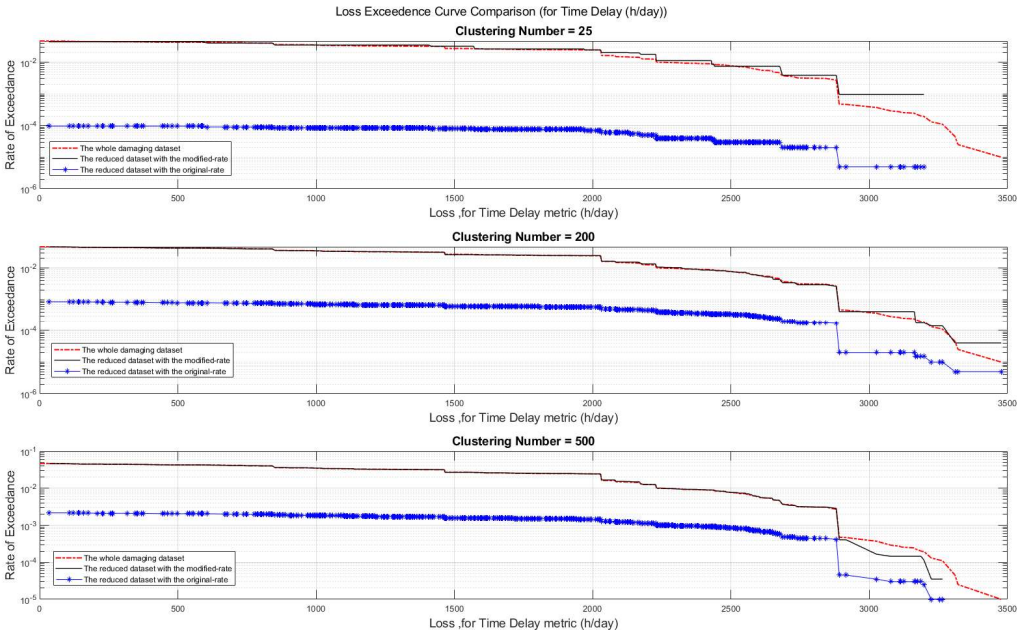


Figure 4-38 Comparison of LECs for reduced dataset with min, 100, and max number of clusters (for Case TD and Weighted Risk-Hazard), (unit of y axis is 1/year).

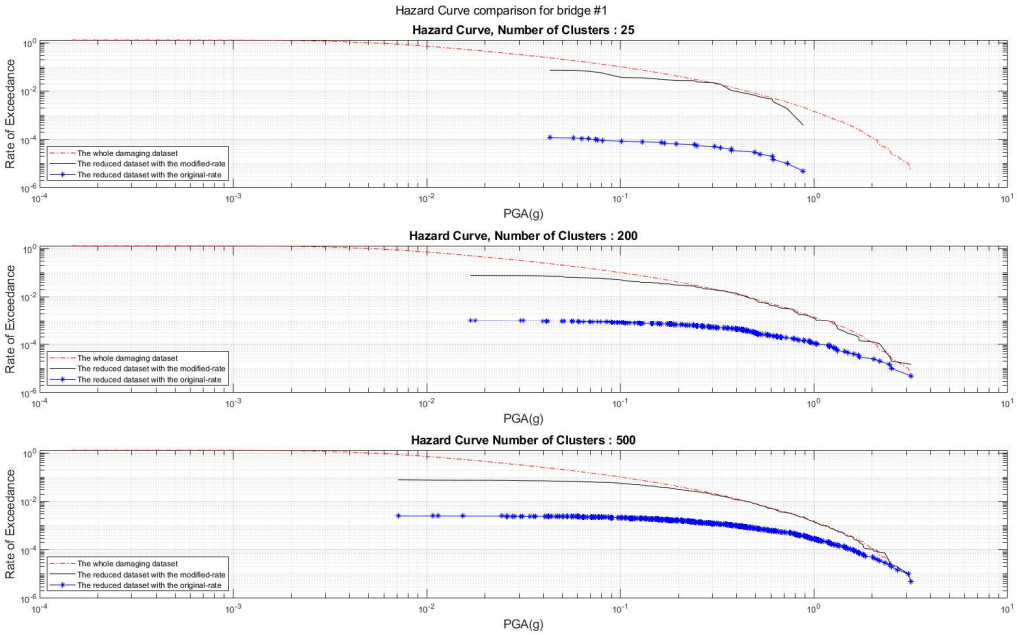


Figure 4-39 Hazard Curve Comparison for Bridge number 1 (for Case TD and Weighted Risk-Hazard), (unit of y axis is 1/year).

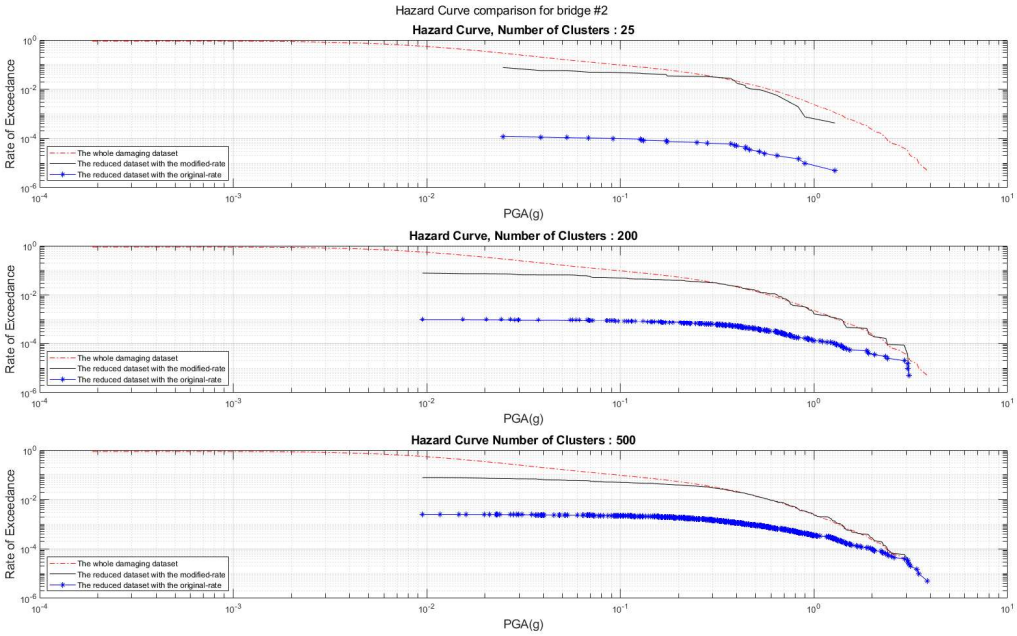


Figure 4-40 Hazard Curve Comparison for Bridge number 2 (for Case TD and Weighted Risk-Hazard), (unit of y axis is 1/year).

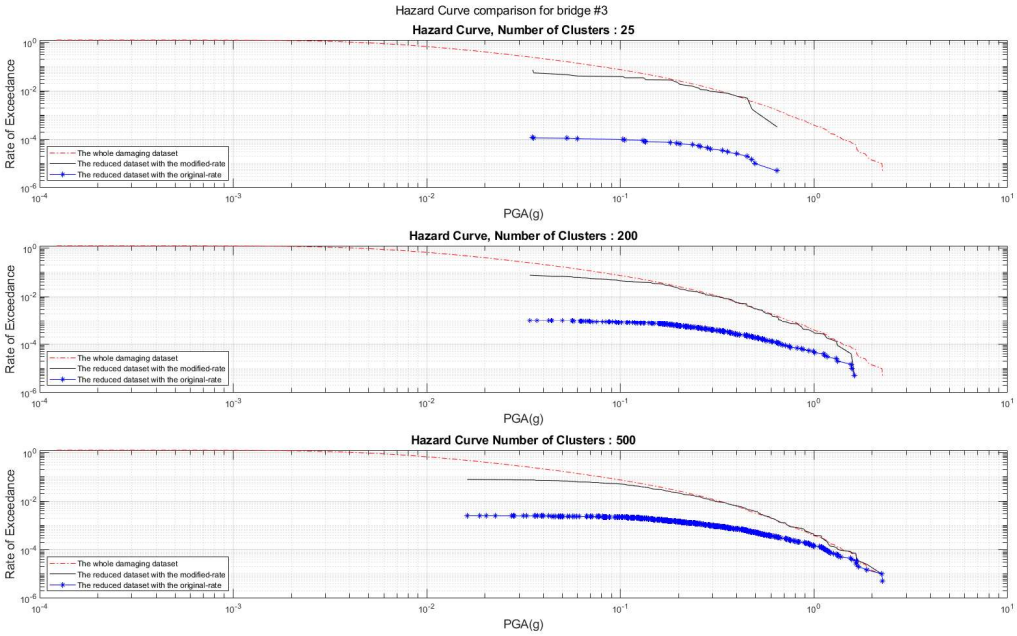


Figure 4-41 Hazard Curve Comparison for Bridge number 3 (for Case TD and Weighted Risk-Hazard), (unit of y axis is 1/year).

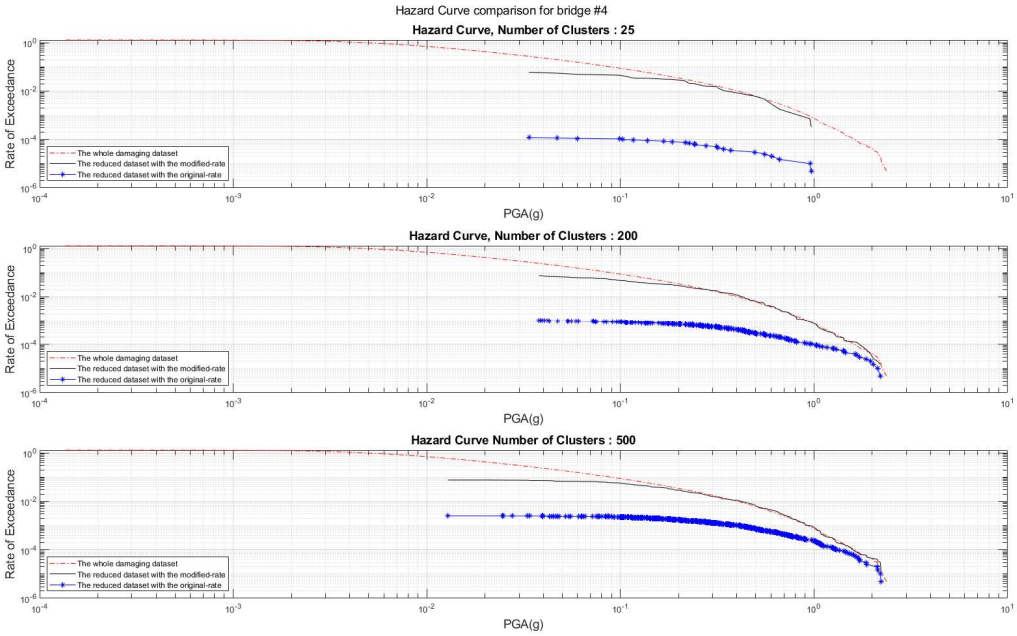


Figure 4-42 Hazard Curve Comparison for Bridge number 4 (for Case TD and Weighted Risk-Hazard), (unit of y axis is 1/year).

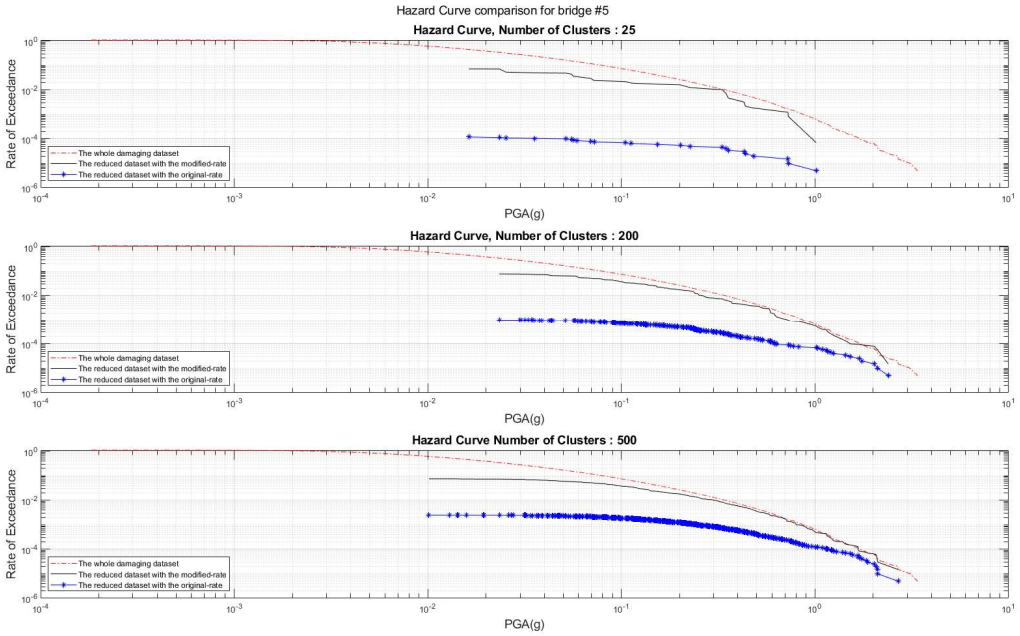


Figure 4-43 Hazard Curve Comparison for Bridge number 5 (for Case TD and Weighted Risk-Hazard), (unit of y axis is 1/year).

In the next part, the effect of equal dominancy of risk and hazard in clustering will be tested. A better hazard estimation is expected especially with fewer number of events in the dataset.

4.2.2.3 Unweighted Risk-Hazard approach TD

In this part, like the Weighted Risk-Hazard approach, 6 features are affecting the clustering, but without considering any weight for the features. Figure 4-44 shows the risk results with the reduced data set in applying 3 different clustering numbers. Again, risk curves fall short in estimating the less frequent risks with the smaller number of clusters. Some values of the loss are missing even in using 500 clusters.

Figure 4-45 to Figure 4-49 show the resulted hazard curve with the reduced dataset in 5 different locations and 3 different numbers of clustering. The hazard results are improved from the previous Weighted Risk-Hazard method and again with greater number of events a good consistency with the reference curve for the less frequent hazard can be observed.

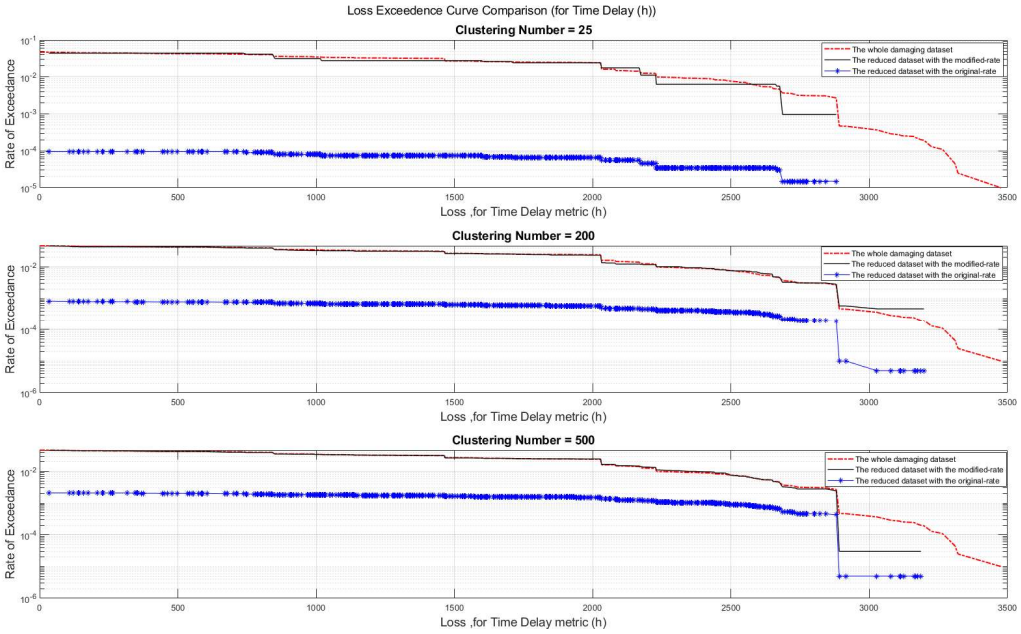


Figure 4-44 Comparison of LECs for reduced dataset with min, error based optimum, max number of clusters (for Case TD and Unweighted Risk-Hazard), (unit of y axis is 1/year).

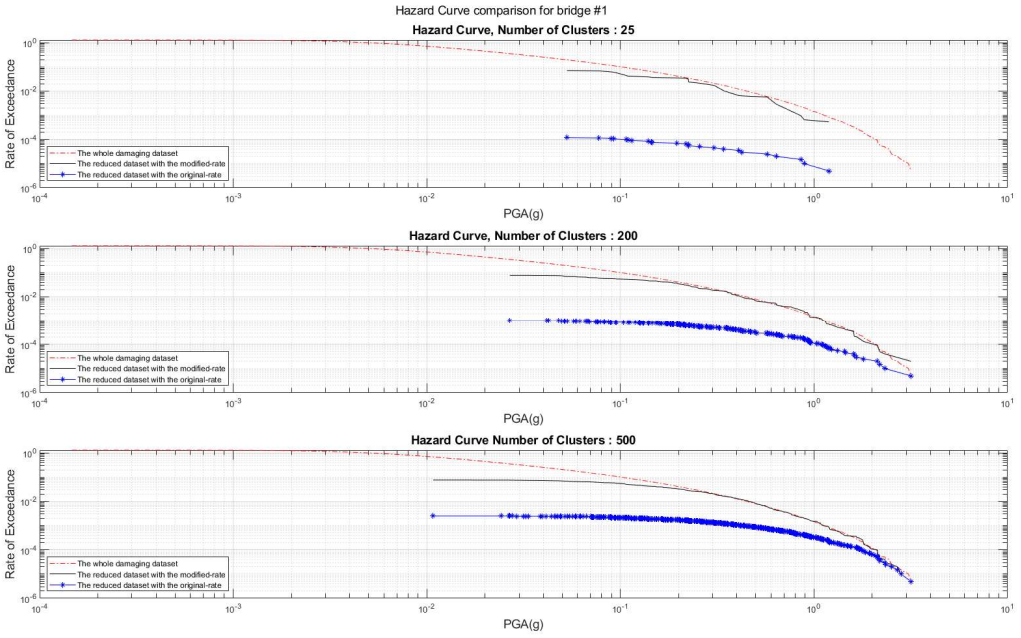


Figure 4-45 Hazard Curve Comparison for Bridge number 1 (for Case TD and Unweighted Risk-Hazard), (unit of y axis is 1/year).

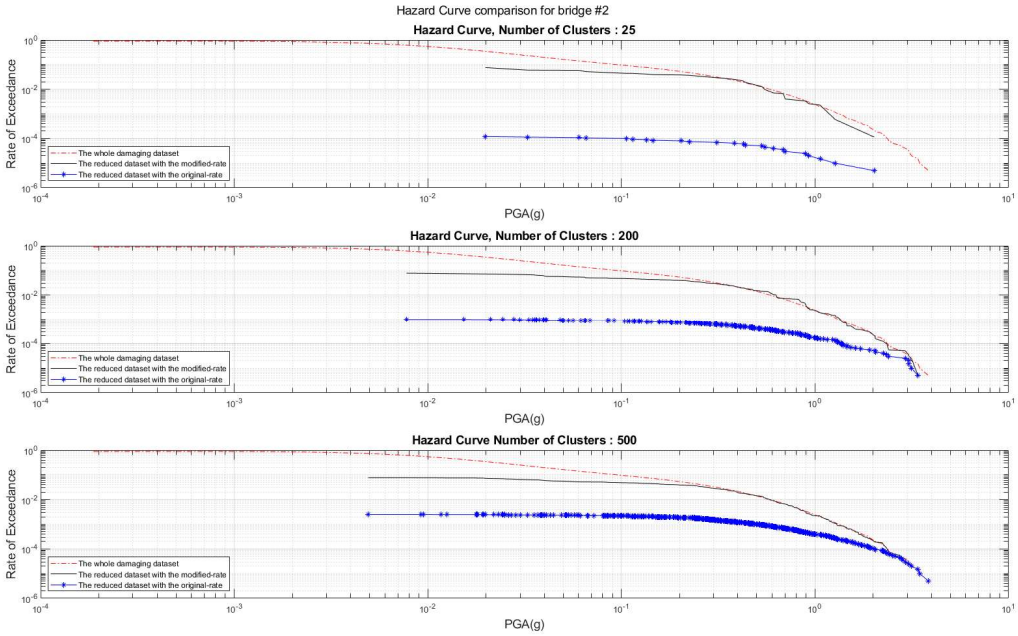


Figure 4-46 Hazard Curve Comparison for Bridge number 2 (for Case TD and Unweighted Risk-Hazard), (unit of y axis is 1/year).

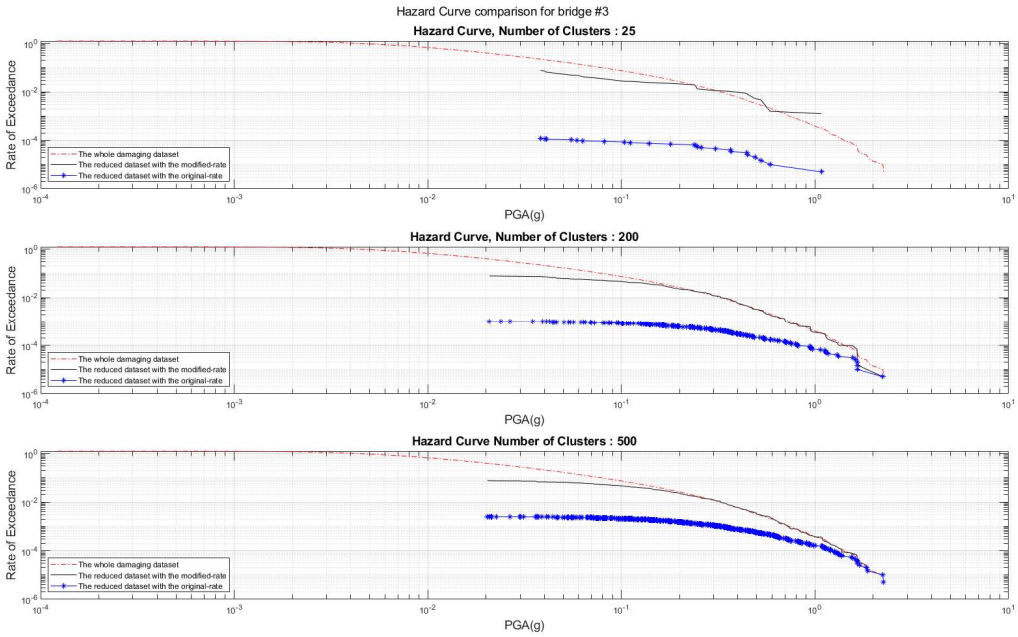


Figure 4-47 Hazard Curve Comparison for Bridge number 3 (for Case TD and Unweighted Risk-Hazard), (unit of y axis is 1/year).

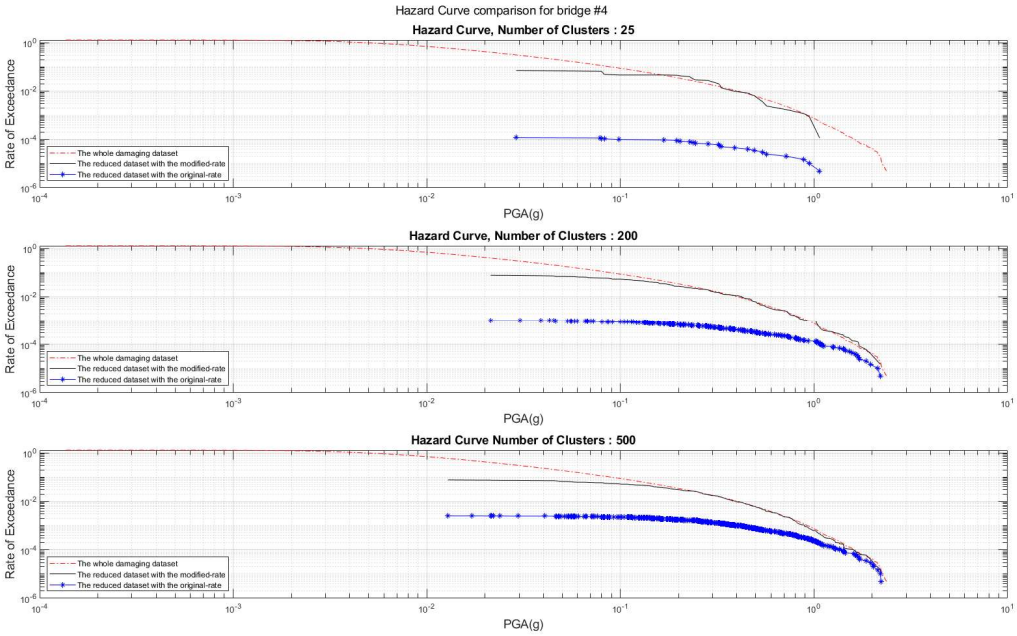


Figure 4-48 Hazard Curve Comparison for Bridge number 4 (for Case TD and Unweighted Risk-Hazard), (unit of y axis is 1/year).

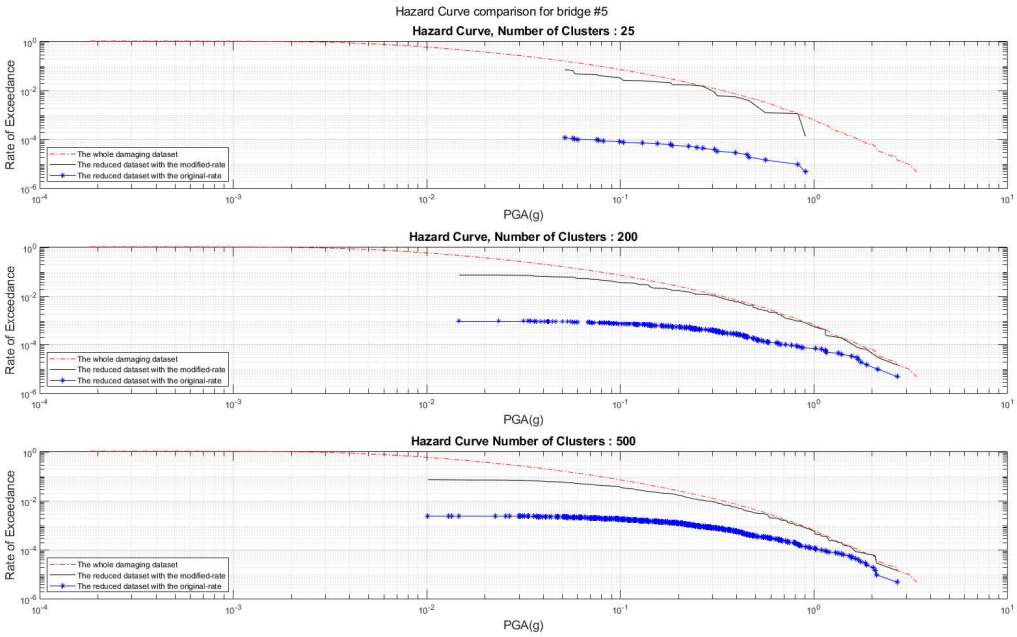


Figure 4-49 Hazard Curve Comparison for Bridge number 5 (for Case TD and Unweighted Risk-Hazard), (unit of y axis is 1/year).

4.2.2.4 Hazard-Consistent with the same clustering number with the case of TD

In this part, the only 5 PGA values of the bridges are considered as the features of clustering and to be compared with the other approaches, the results are shown with the same clustering number that were used in Risk-Consistent approach of this loss metric.

Figure 4-50 shows the resulted loss exceedance curve for TD using the reduced data set with this Hazard-Consistent approach for 3 different clustering numbers of min, optimum and maximum clustering numbers (all for risk-based apparat of TD). Like the two previous approaches, even for 500 number of clustering the error in risk curves is significant.

Figure 4-51 to Figure 4-55 show the resulted hazard curve by using Hazard-Consistent approach. again, it should be mentioned that like the method for LDB, in this Hazard-Consistent approach only damaging events are considered and clustered therefore, a wide range of low PGAs are missing even in greater clustering numbers. However, improvements in the resulted hazard curves especially in case of minimum number of clusters, 25, are observed.

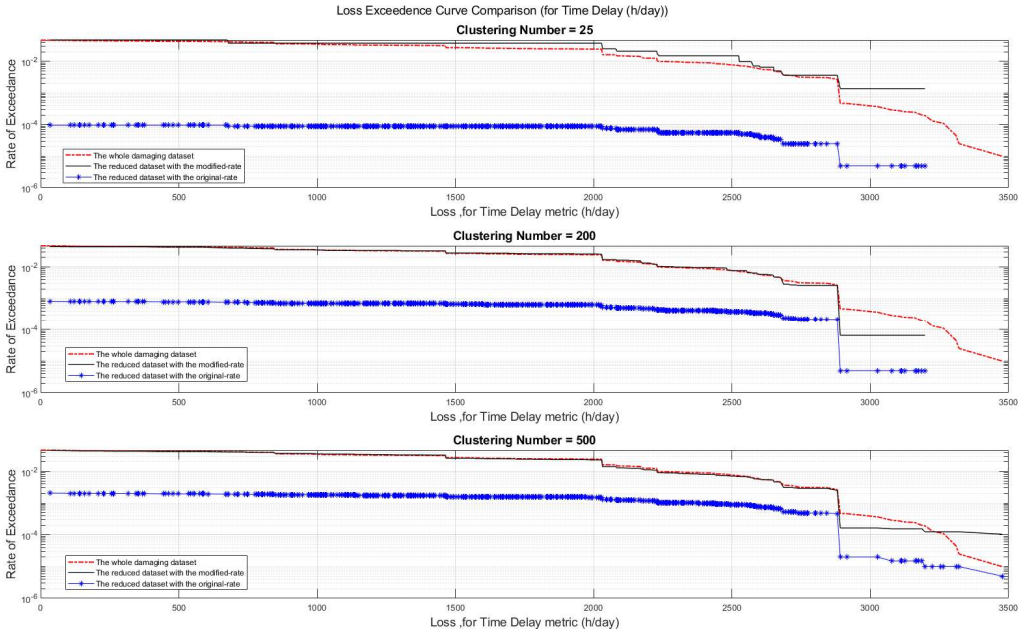


Figure 4-50 Comparison of LECs for reduced dataset with min, 200, max number of clusters (for Case TD and Hazard-Consistent), (unit of y axis is 1/year).

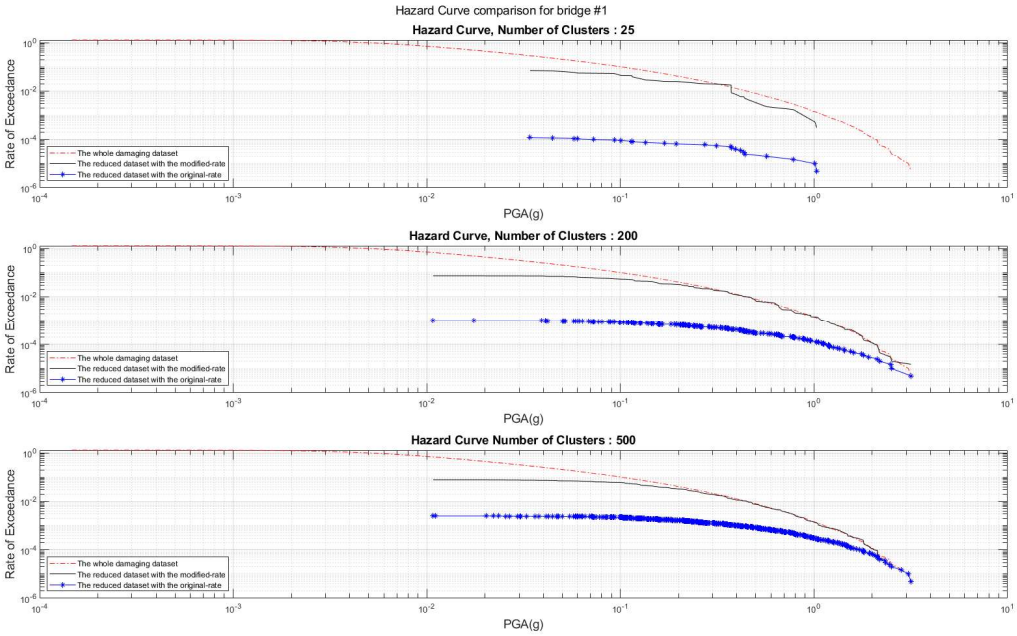


Figure 4-51 Hazard Curve Comparison for Bridge number 1 (for Case TD and Hazard-Consistent), (unit of y axis is 1/year).

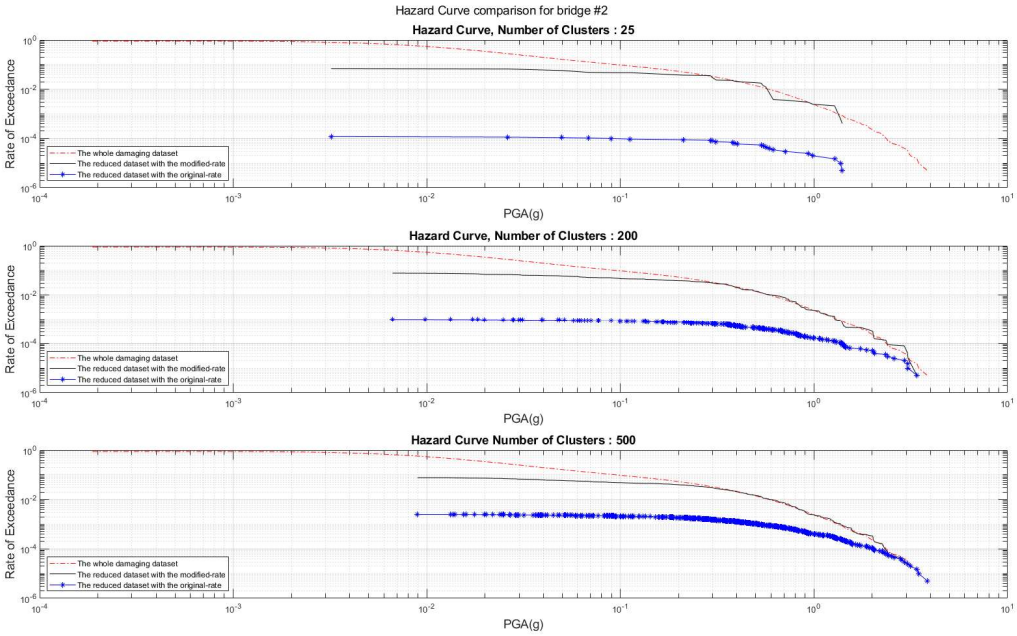


Figure 4-52 Hazard Curve Comparison for Bridge number 2 (for Case TD and Hazard-Consistent), (unit of y axis is 1/year).

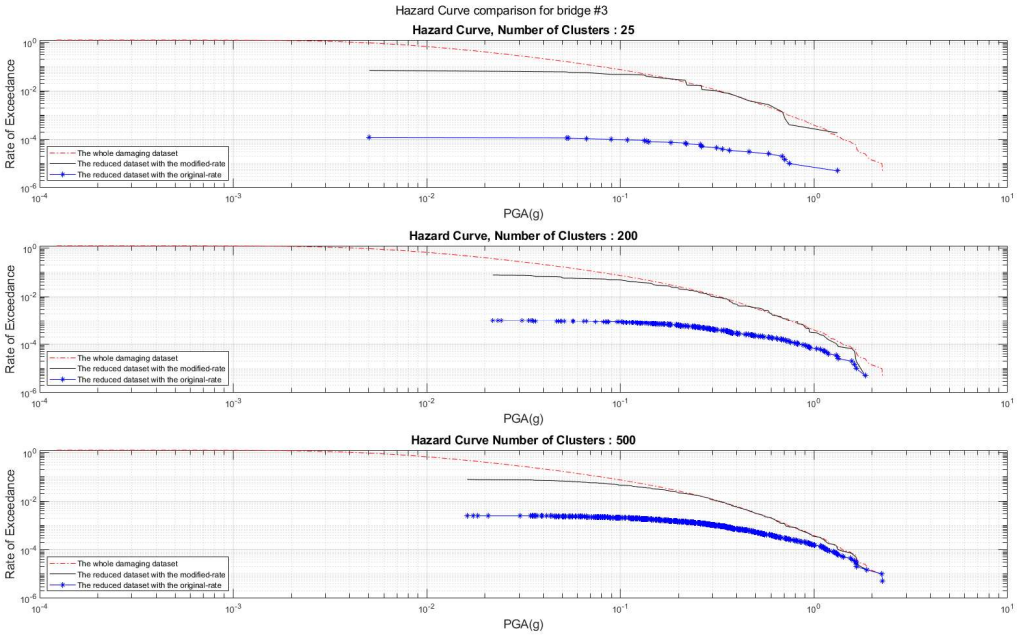


Figure 4-53 Hazard Curve Comparison for Bridge number 3 (for Case TD and Hazard-Consistent), (unit of y axis is 1/year).

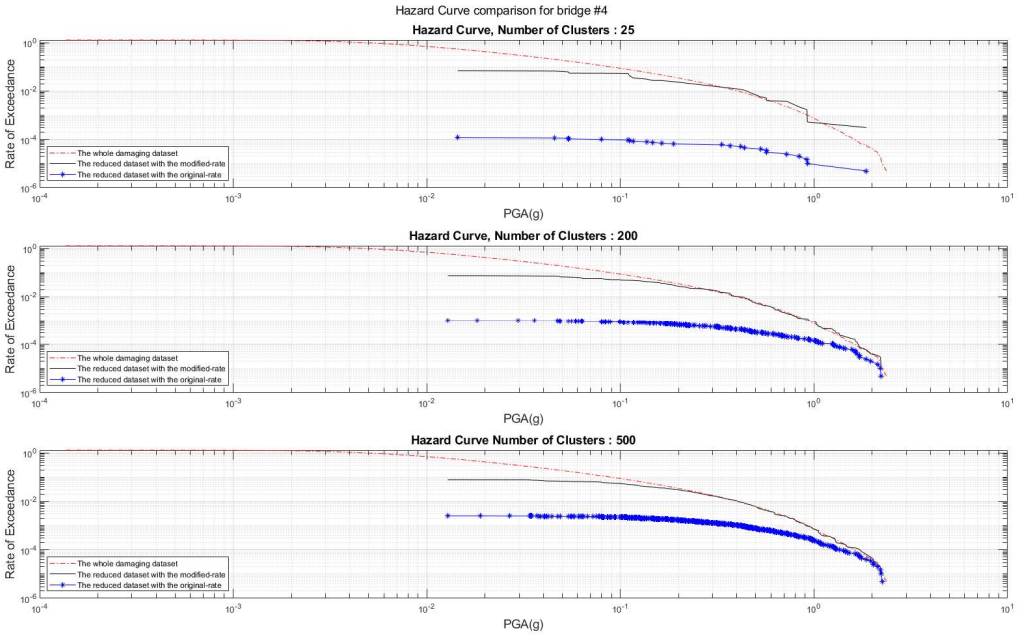


Figure 4-54 Hazard Curve Comparison for Bridge number 4 (for Case TD and Hazard-Consistent), (unit of y axis is 1/year).

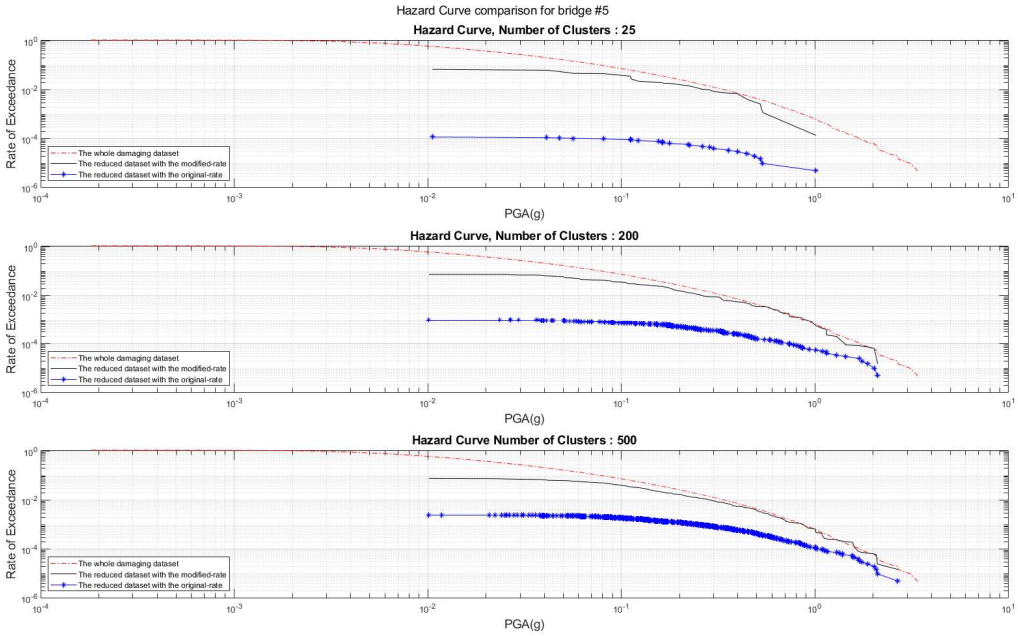


Figure 4-55 Hazard Curve Comparison for Bridge number 5 (for Case TD and Hazard-Consistent), (unit of y axis is 1/year).

4.2.2.5 Discussion on the results of different approaches with the case of TD loss metric

According to the comparison of the results of 4 different approaches in case of time delay loss metric, as it was observed in case of LDB, in Risk-Consistent approach, more satisfactory results can be obtained with the error risk optimum number of clusters and a wide range of loss values can be covered in the resulted risk curve even in the minimum number of clusters. On the other hand, the resulted hazard curves can be unreliable by Risk-Consistent method. By increasing the effect of hazard in the clustering, less accurate risk curves but improved hazard curves are derived and to get better risk results more clustering numbers (greater number of events in the reduced dataset) is needed.

4.2.3 Results for Simple Connectivity Loss (SCL)

Simple connectivity loss (SCL) is one of the loss metrics applied in this study that related to the connectivity of the network as the required feature of the network to be considered as functionable. Like 2 other loss metrics, 4 different approaches with the order of Risk-Consistent, Weighted Risk-Hazard consistent, Unweighted Risk-Hazard consistent and hazard consisted of are studied for the case of SCL as loss metric and illustrated as follows. The number of clusters are decided in Risk-Consistent approach and for the rest of approaches to be comparable the same clustering numbers are used.

4.2.3.1 Risk-Consistent approach with SCL loss metric

In Risk-Consistent approach for SCL, the only involved feature in calculating of distance metric for the clustering algorithm is SCL loss values. first the appropriate number of clusters are decided by two methods of elbow point and the previously mentioned risk error criteria which are shown in Figure 4-56 and Figure 4-57 respectively. Like what has been observed in simple spatial griding method, due to the a few numbers of distinct values of SCL in our results, the total number of clustering for this case in our study could not be greater than certain. Here based on the risk error plot, the optimum clustering number

15 number of clusters are assumed to be suitable and later in this section by comparing the resulted risk curve and hazard curve this assumption can be judged.

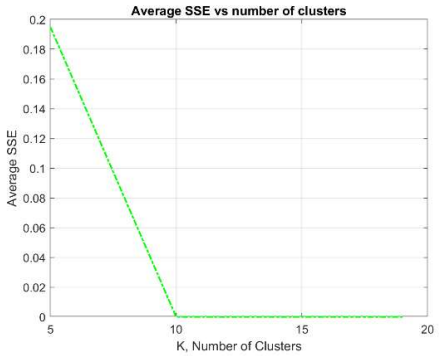


Figure 4-56 Elbow point for selecting the K (for Case SCL and Risk-Consistent)

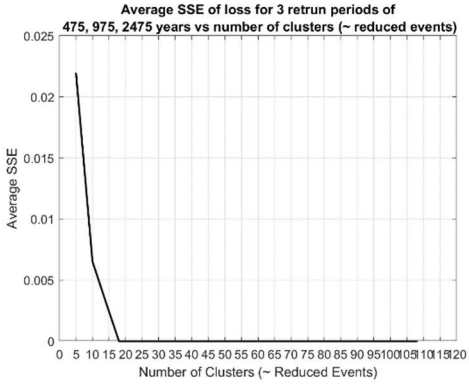


Figure 4-57 Average SSE of loss vs ClustNo for 475 ,975,2475 (for Case SCL and Risk-Consistent)

Figure 4-58 shows the resulted risk curve with the minimum number of clustering in this study, with the assumed optimum number of clustering in last paragraph, and with the maximum number of possible clustering of SCL loss metric in this study due to the limited number of distinct values of this loss in this study. Based on the comparison of the resulted risk curves, it can be concluded that the assumed optimum clustering number shows a satisfactory accuracy in comparison with the reference risk curve by the whole events dataset.

Since it is a Risk-Consistent approach, although the resulted risk curves are accurate this might not be the case for hazard curves. As is shown in Figure 4-59 Figure 4-63, hazard curves for 5 different bridge locations are not compatible with their pertinent reference hazard curve, that are generated with the whole event dataset and a wide range of the intensity measures (here PGA) is missing in the hazard curves resulted by the reduced dataset. however, it can be shown that by increasing the number of clusters the hazard curves can be improved.

In next approaches in following subsections, it is tried to decrease the dominancy of the risk (in 3 steps of weighted Risk-Consistent, unweighted Risk-Consistent and Hazard-Consistent) to include the effect of hazard in clustering and the resulted risk curves and hazard curves are compared. The applied loss clustering numbers in the following approaches will be identical with the Risk-Consistent approach to provide a better comparison of the effect of approaches on the results.

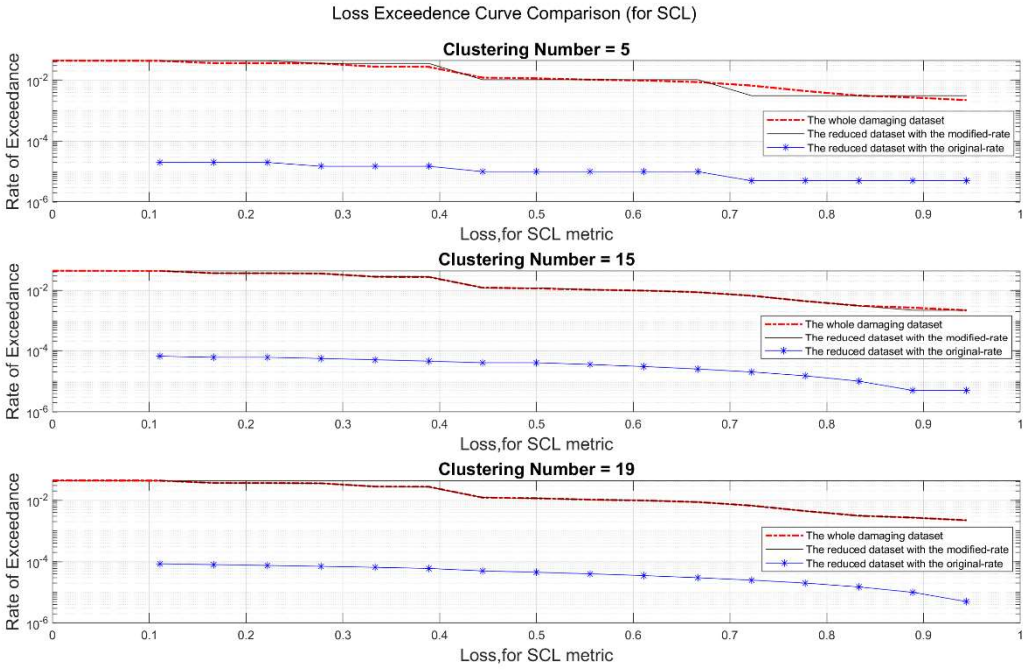


Figure 4-58 Comparison of LECs for reduced dataset with min, error based optimum, max number of clusters (for Case SCL and Risk-Consistent), (unit of y axis is 1/year).

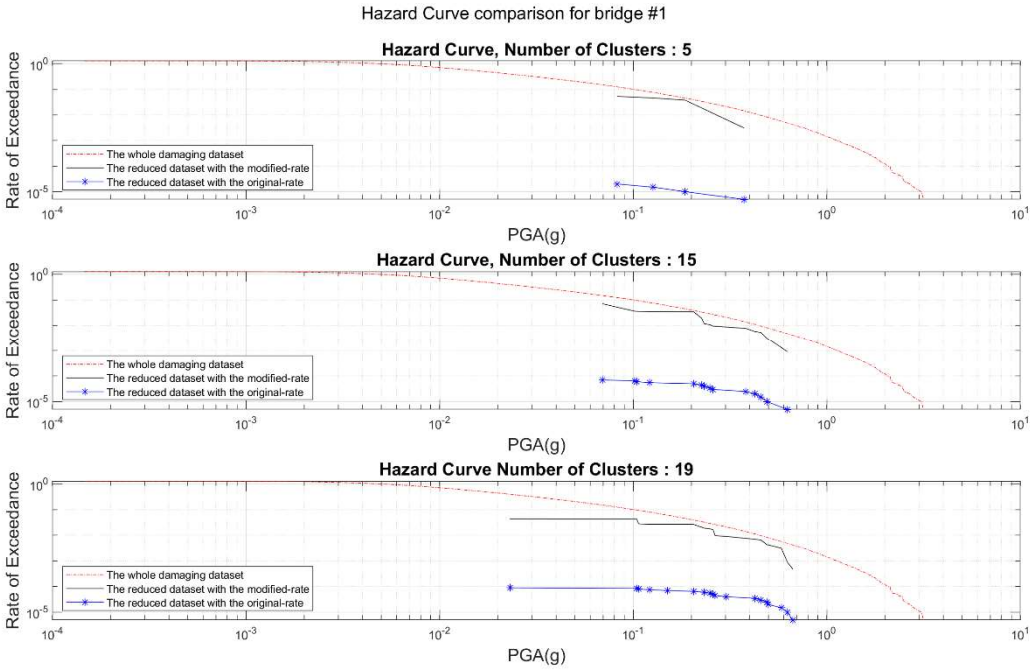


Figure 4-59 Hazard Curve Comparison for Bridge number 1 (for Case SCL and Risk-Consistent), (unit of y axis is 1/year).

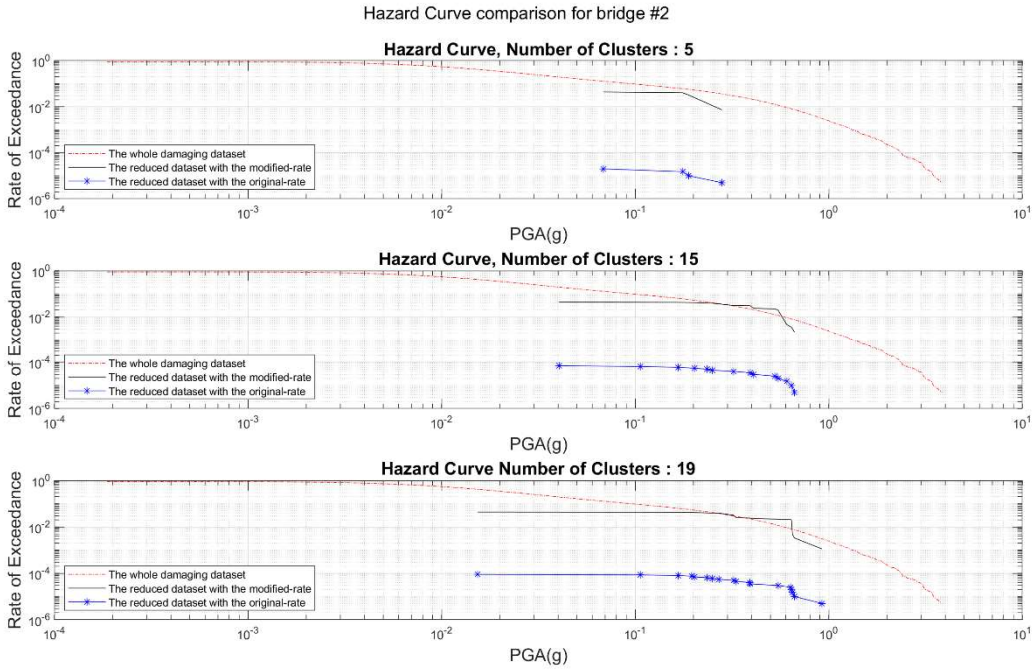


Figure 4-60 Hazard Curve Comparison for Bridge number 2 (for Case SCL and Risk-Consistent), (unit of y axis is 1/year).

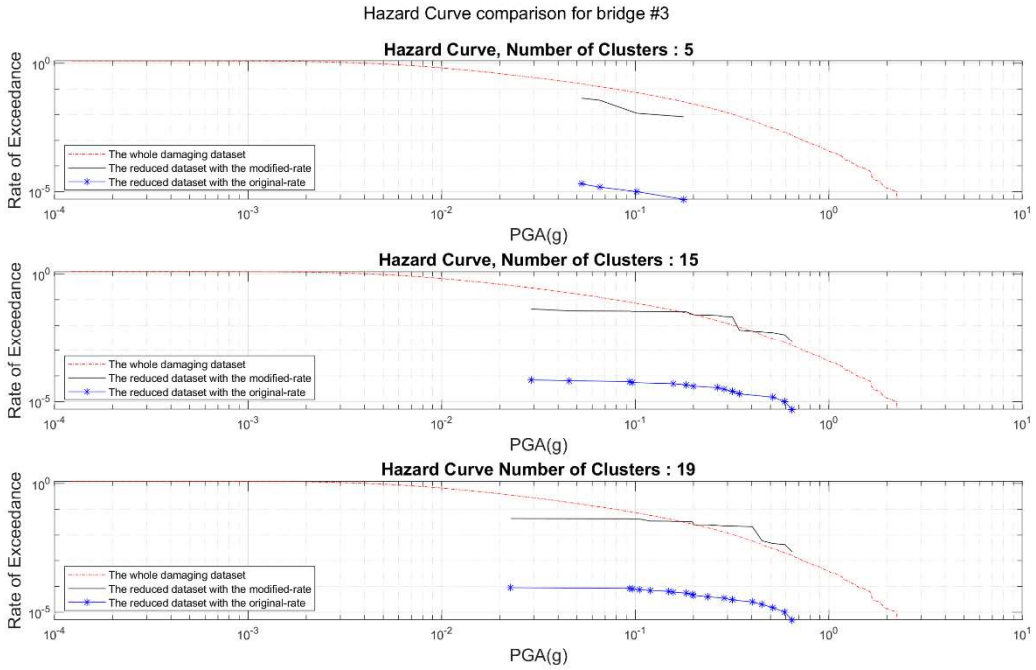


Figure 4-61 Hazard Curve Comparison for Bridge number 3 (for Case SCL and Risk-Consistent), (unit of y axis is 1/year).

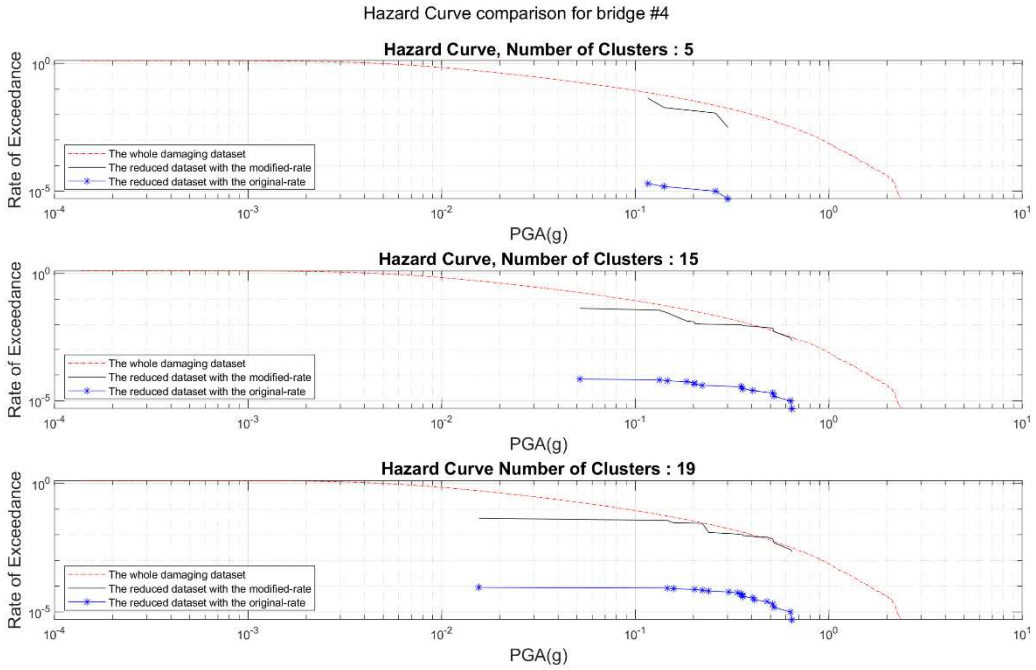


Figure 4-62 Hazard Curve Comparison for Bridge number 4 (for Case SCL and Risk-Consistent), (unit of y axis is 1/year).

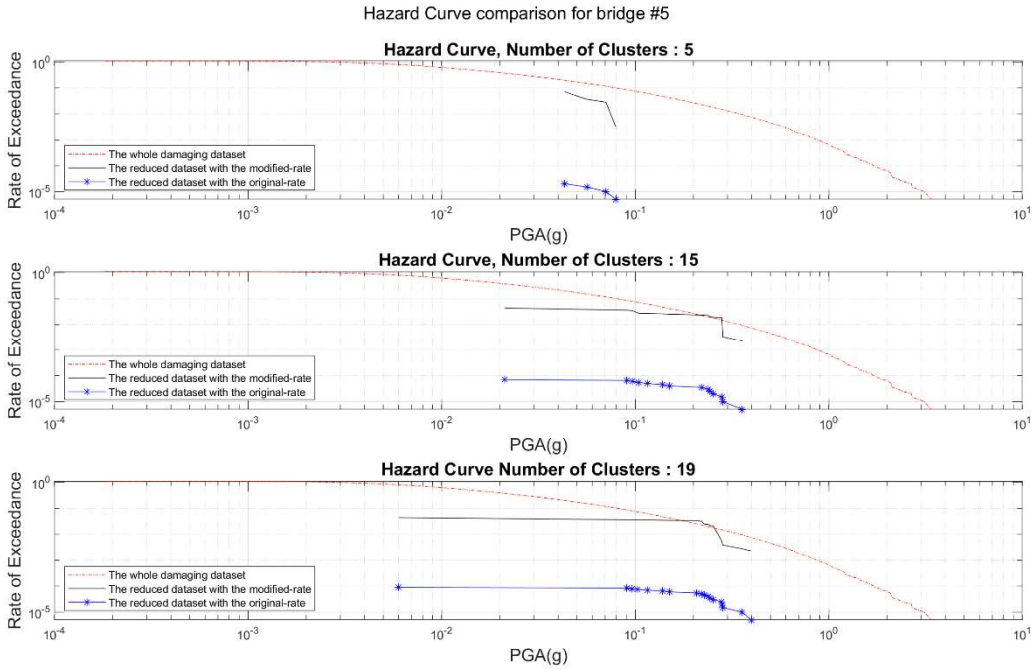


Figure 4-63 Hazard Curve Comparison for Bridge number 5 (for Case SCL and Risk-Consistent), (unit of y axis is 1/year).

4.2.3.2 Weighted Risk-Hazard approach with SCL loss metric

In this section the weighted Risk-Consistent approach is studied in case of SCL loss metric values. means that in calculation of distance metric to use in K-means clustering algorithm the value of loss has a weight

of 50% and the PGA values for each of 5 bridge locations has a weight of 10%. in this way the dominancy of the risk effect in Risk-Consistent approach is slightly decreased to check the possibility of getting a better compatibility in hazard curves with the resulted reduced dataset.

Figure 4-64 shows the comparison of the resulted loss exceedance curves with 3 different clustering numbers which are identical with those were used in the Risk-Consistent approach. As it can be observed in this figure, by involving the effect of hazard and decreasing the effect of risk in the clustering, the resulted risk curves are not as accurate as the ones resulted by the Risk-Consistent approach but still their errors are not significant. Which is what was expected and has been observed in the previous two loss metrics of LDB and TD.

Figure 4-65 to Figure 4-69 show the resulted hazard curves in the location of 5 bridges that in this approach has been involved in the clustering. The resulted hazard curves are showing an improvement in the compatibility of the hazard curves with the pertinent reference hazard curve for each bridge in comparison with the results of the Risk-Consistent approach for SCL.

In next approach, which is the Unweighted Risk-Hazard, all the features in calculating of distance metrics are treated equally and like two previous loss metrics, the effect of not using any weights in the resulted risk and hazard curves are studied for the case of SCL as well.

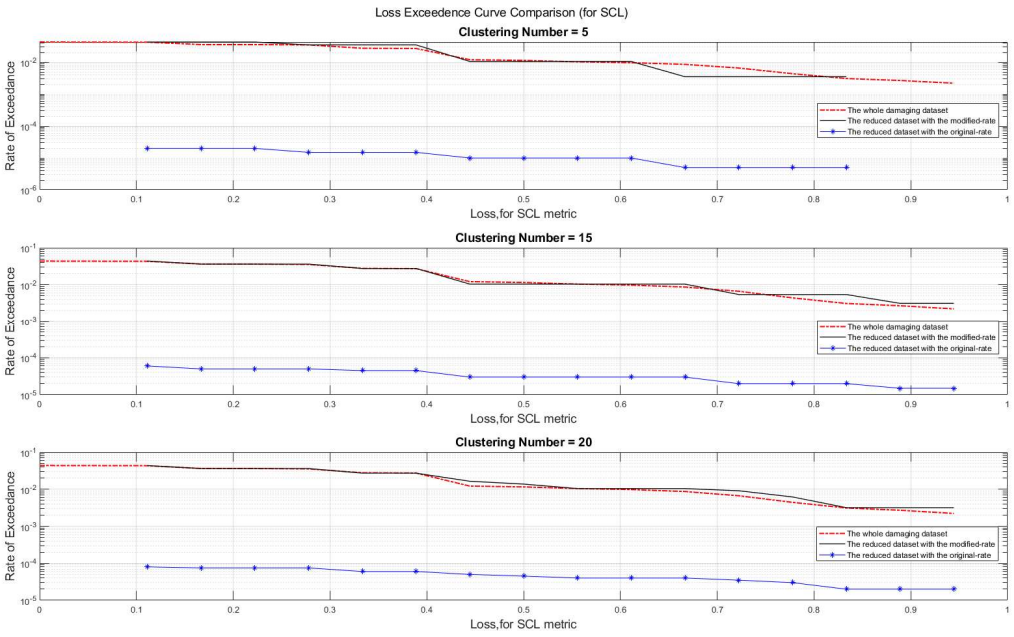


Figure 4-64 Comparison of LECs for reduced dataset with min, error based optimum, max number of clusters (for Case SCL and Weighted Risk-Hazard), (unit of y axis is 1/year).

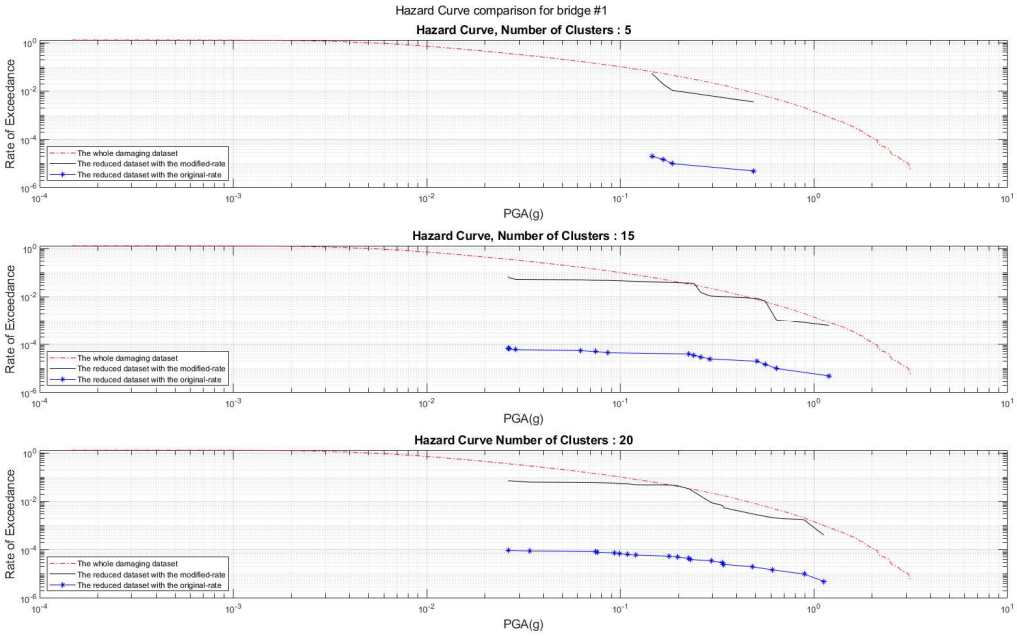


Figure 4-65 Hazard Curve Comparison for Bridge number 1 (for Case SCL and Weighted Risk-Hazard), (unit of y axis is 1/year).

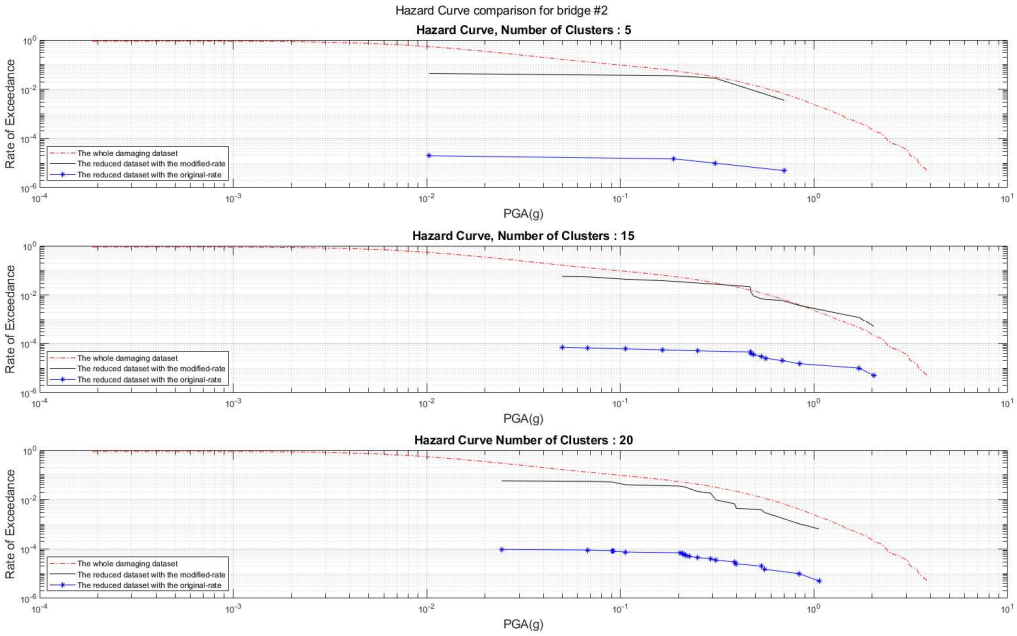


Figure 4-66 Hazard Curve Comparison for Bridge number 2 (for Case SCL and Weighted Risk-Hazard), (unit of y axis is 1/year).

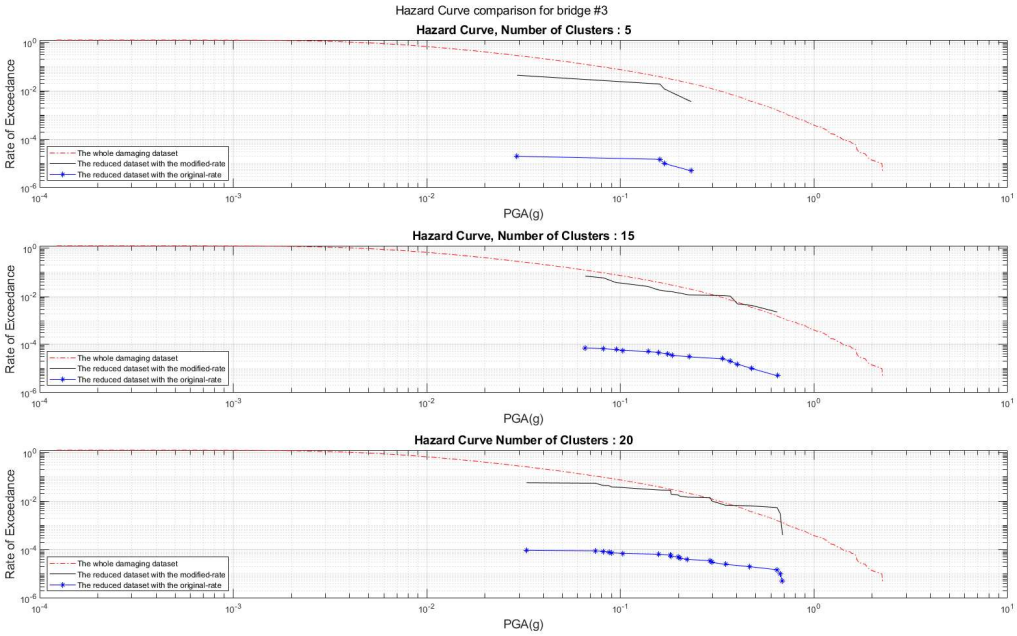


Figure 4-67 Hazard Curve Comparison for Bridge number 3 (for Case SCL and Weighted Risk-Hazard), (unit of y axis is 1/year).

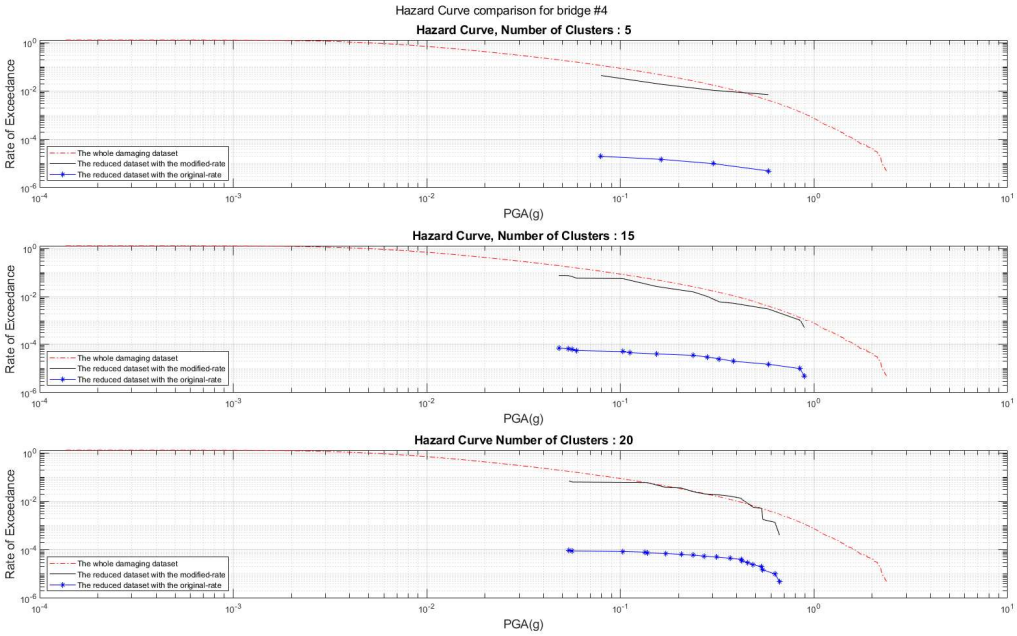


Figure 4-68 Hazard Curve Comparison for Bridge number 4 (for Case SCL and Weighted Risk-Hazard), (unit of y axis is 1/year).

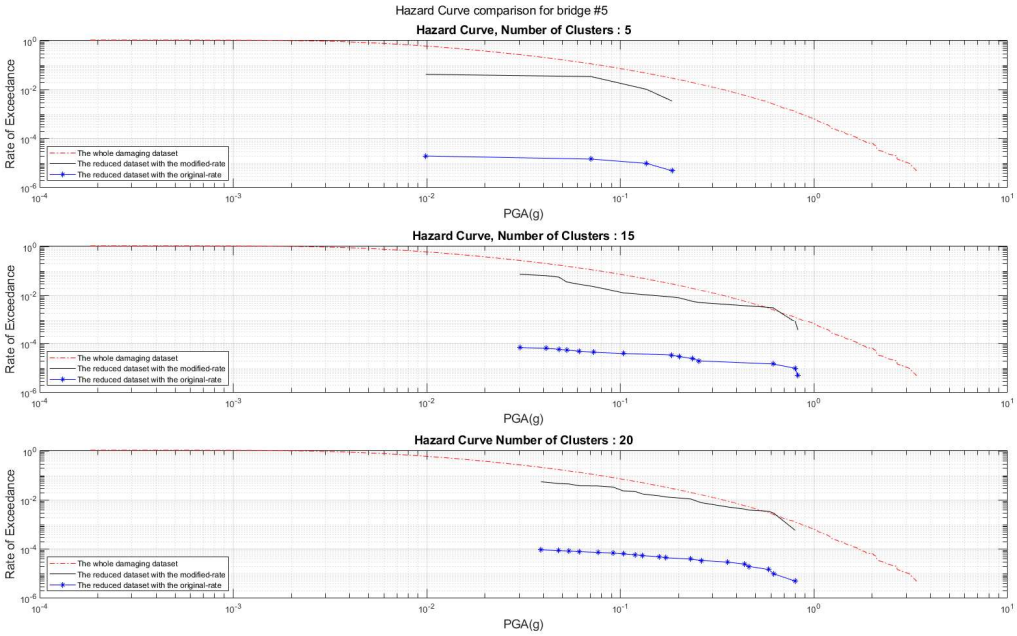


Figure 4-69 Hazard Curve Comparison for Bridge number 5 (for Case SCL and Weighted Risk-Hazard), (unit of y axis is 1/year).

4.2.3.3 Unweighted Risk-Hazard approach with SCL loss metric

In this part, the distance metric to be applied in k-means clustering is include unweighted features of SCL values of the network and 5 PGA values of bridges per event. In this way all features are treated equally, and the resulted risk and curves are shown in following.

Figure 4-70 shows the comparison of the resulted risk curve for 3 different clustering numbers identical to the applied clustering numbers in previous approaches for SCL. As it can be observed in this figure, the accuracy of the resulted Risk-Consistent curve with using the reduced dataset has been decreased in comparison with the two previous approaches of Risk-Consistent and weighted Risk-Consistent for the same clustering numbers. This trend is similar with what was resulted with two other loss metrics LDB and TD.

Figure 4-71 to Figure 4-75 show the resulted hazard curves with this Unweighted Risk-Hazard approach. some improvements in the accuracy of the curves are observed for example in the resulted hazard for bridge number 3, both for clustering number of 15 and 20 better compatibility is shown in comparison with the previous section, Weighted Risk-Hazard approach for the SCL. Again, it should be mentioned that due to the initial step in eliminating events that do not damage any bridges, a range of the PGAs are missing in the resulted hazard curves that are generated with the reduced dataset anyway and the goal here is to get better compatibility in the part of the hazard curve that matters in point of causing damage.

In the next approach, only hazard will be effective in clustering and the results loss curves for the SCL and hazard curves for 5 bridges are generated to be compared with 3 other approaches that has been studied for the SCL so far.

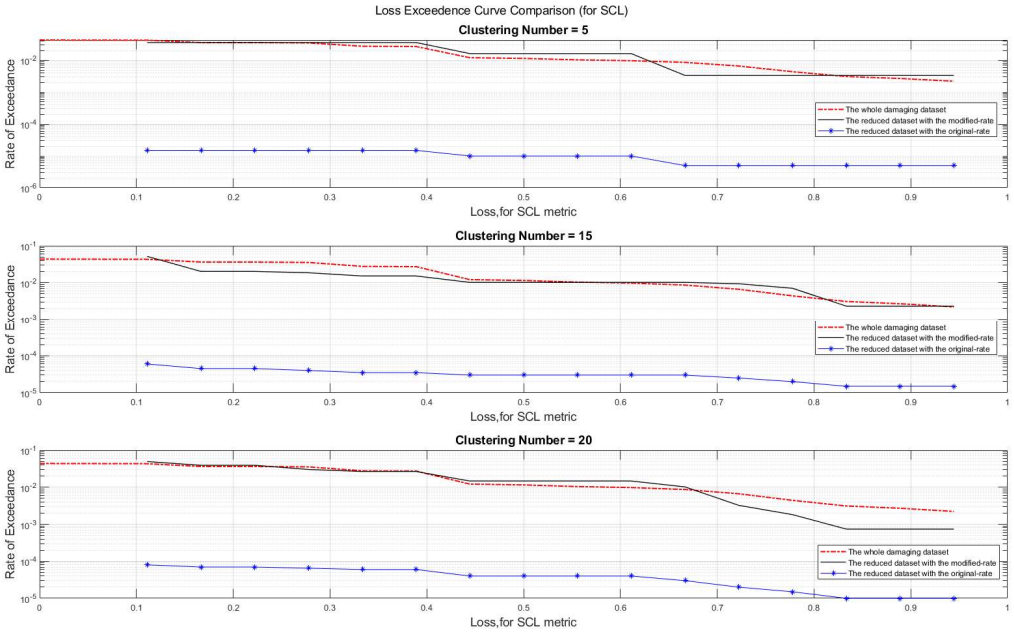


Figure 4-70 Comparison of LECs for reduced dataset with min, error based optimum, max number of clusters (for Case SCL and Unweighted Risk-Hazard), (unit of y axis is 1/year).

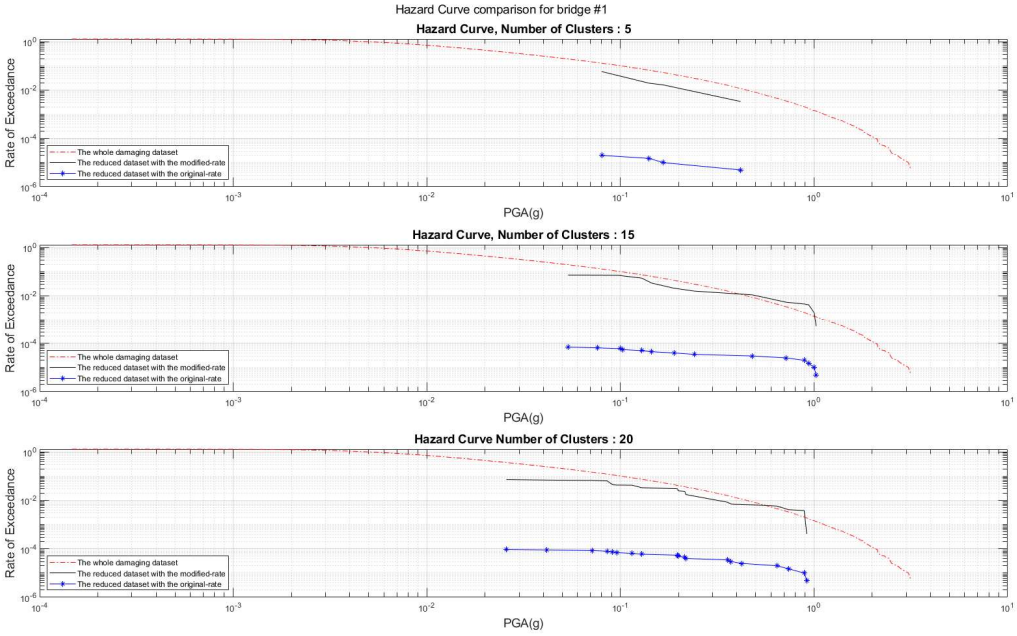


Figure 4-71 Hazard Curve Comparison for Bridge number 1 (for Case SCL and Unweighted Risk-Hazard), (unit of y axis is 1/year).

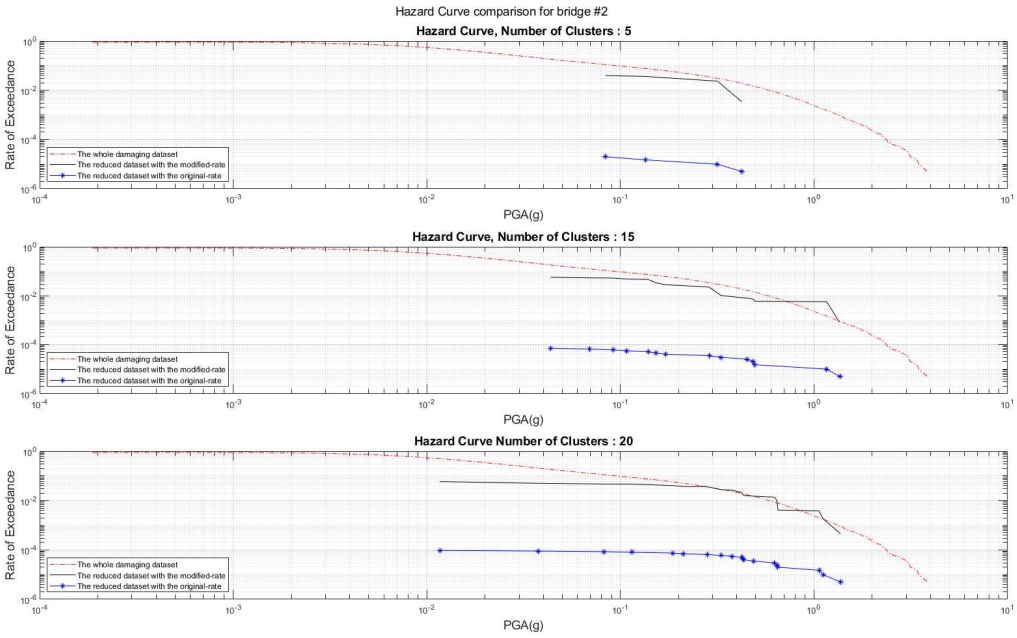


Figure 4-72 Hazard Curve Comparison for Bridge number 2 (for Case SCL and Unweighted Risk-Hazard), (unit of y axis is 1/year).

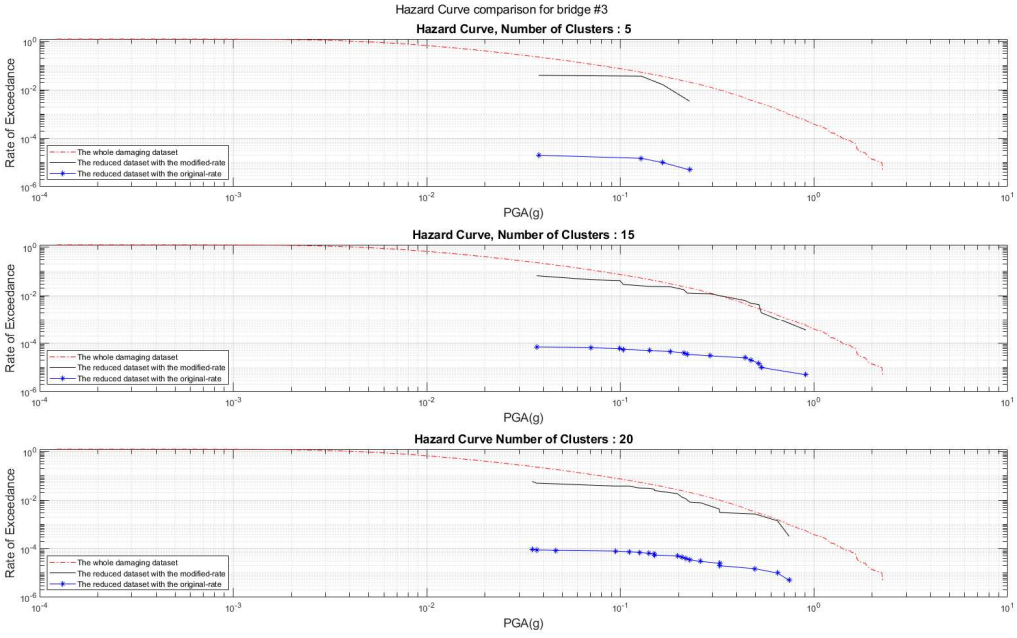


Figure 4-73 Hazard Curve Comparison for Bridge number 3 (for Case SCL and Unweighted Risk-Hazard), (unit of y axis is 1/year).

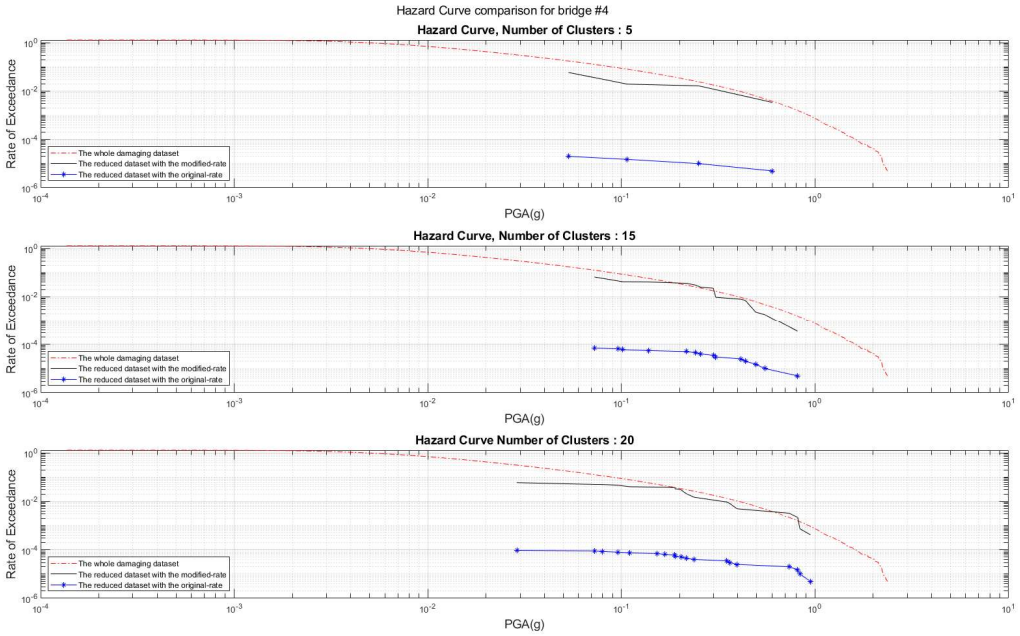


Figure 4-74 Hazard Curve Comparison for Bridge number 4 (for Case SCL and Unweighted Risk-Hazard), (unit of y axis is 1/year).

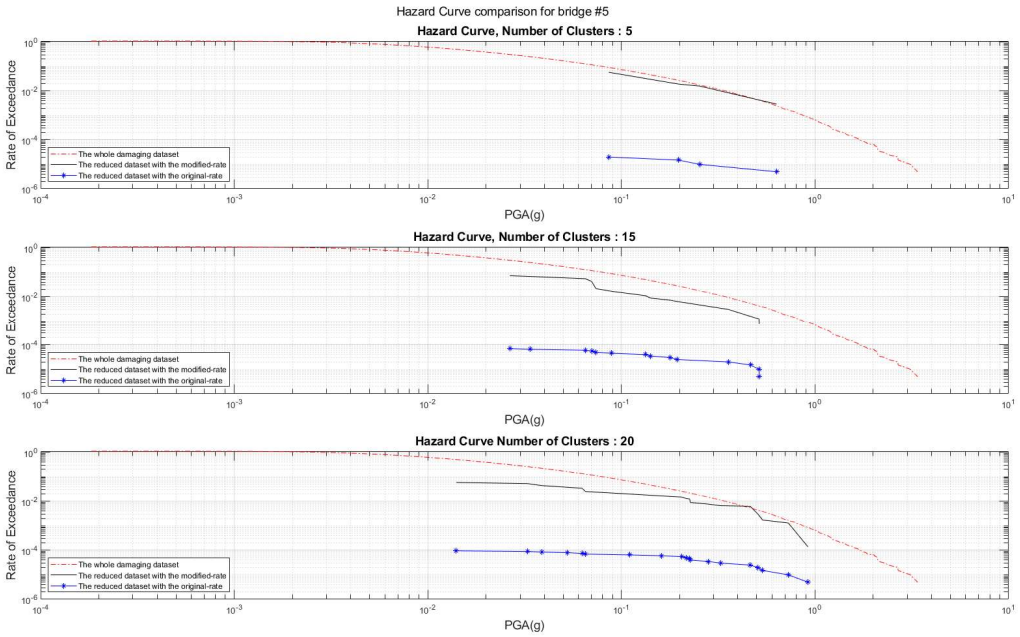


Figure 4-75 Hazard Curve Comparison for Bridge number 5 (for Case SCL and Unweighted Risk-Hazard), (unit of y axis is 1/year).

4.2.3.4 Hazard-Consistent comparison with the clustering number of SCL case

In this section only the 5 PGA values per 5 bridges per event is effective in calculation of the distance metric od the clustering algorithm. The resulted reduced dataset has been applied to generate loss exceedance curve for SCL as the distance metric. it should be mentioned that the clustering numbers are shown in the results are this section are tried to be identical with the clustering numbers that were applied in previous approaches for SCL to be compared with their results.

Figure 4-76 shows the comparison of the risk curve for those 3 different clustering numbers. As it can be observed, removing the effect of the loss value explicitly affected the resulted risk and resulted significant errors in comparison with the reference risk curve for example in cases with clustering numbers like 15 and 20 that a satisfactory accuracy was observed in the Risk-Consistent approach.

Figure 4-77 Figure 4-81 show the resulted hazard curves in the 5 bridges that the values of intensity measure in their location were included in calculation of distance metric of the clustering. a major change in the accuracy of hazard curve is not observed in comparison with the results of the unweighted approach and that might be because we use this reduction for the damaging events and many events that did not affect the network has been eliminated from the dataset that we implement the Hazard-Consistent clustering on them, therefore.

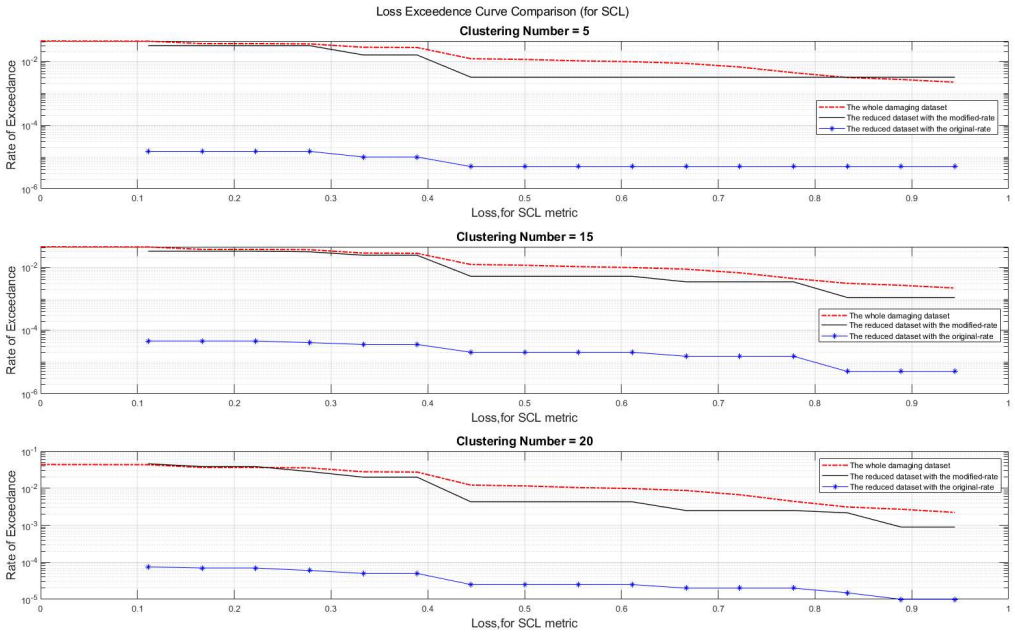


Figure 4-76 Comparison of LECs for reduced dataset with min, 15, max number of clusters (for Case SCL and Hazard-Consistent), (unit of y axis is 1/year).

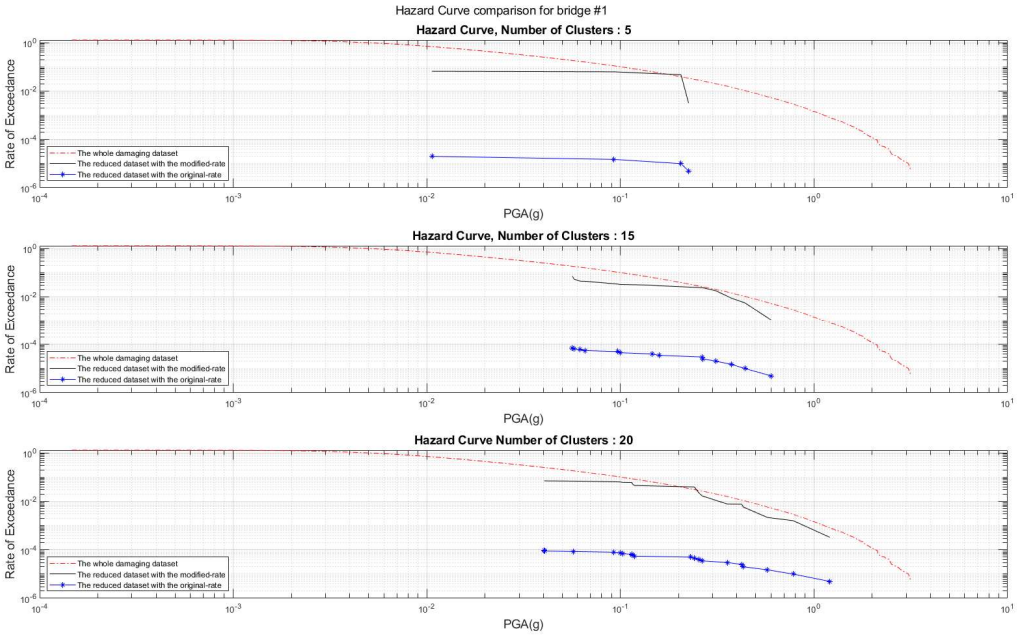


Figure 4-77 Hazard Curve Comparison for Bridge number 1 (for Case SCL and Hazard-Consistent), (unit of y axis is 1/year).

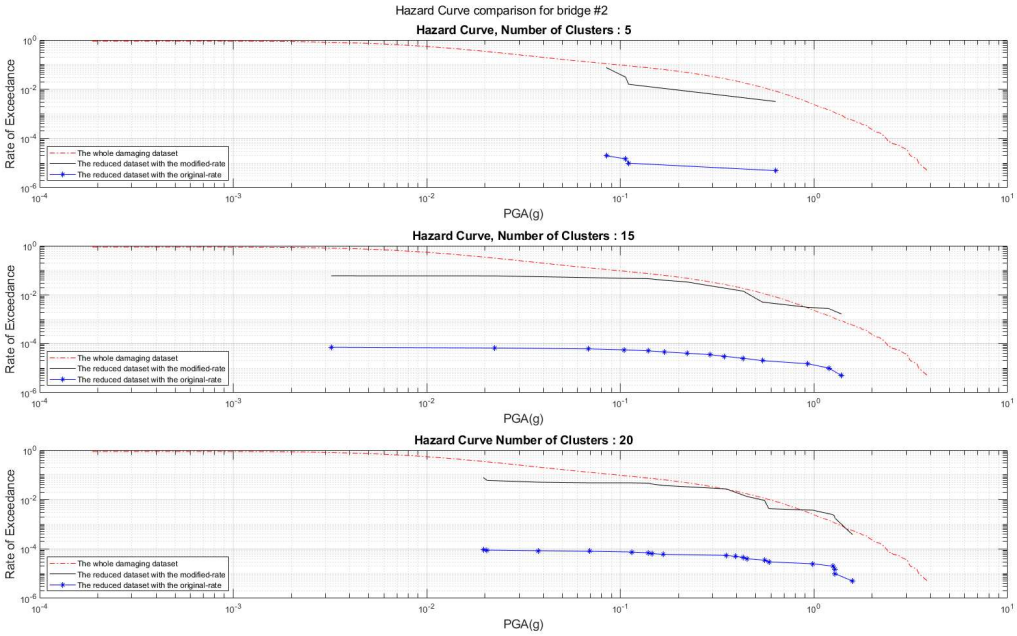


Figure 4-78 Hazard Curve Comparison for Bridge number 2 (for Case SCL and Hazard-Consistent), (unit of y axis is 1/year).

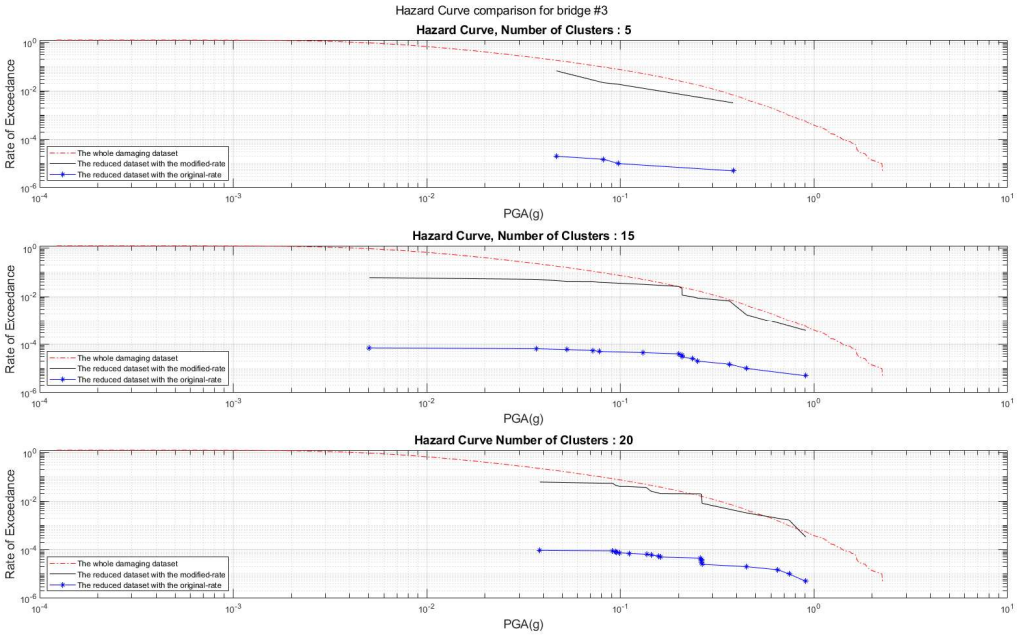


Figure 4-79 Hazard Curve Comparison for Bridge number 3 (for Case SCL and Hazard-Consistent), (unit of y axis is 1/year).

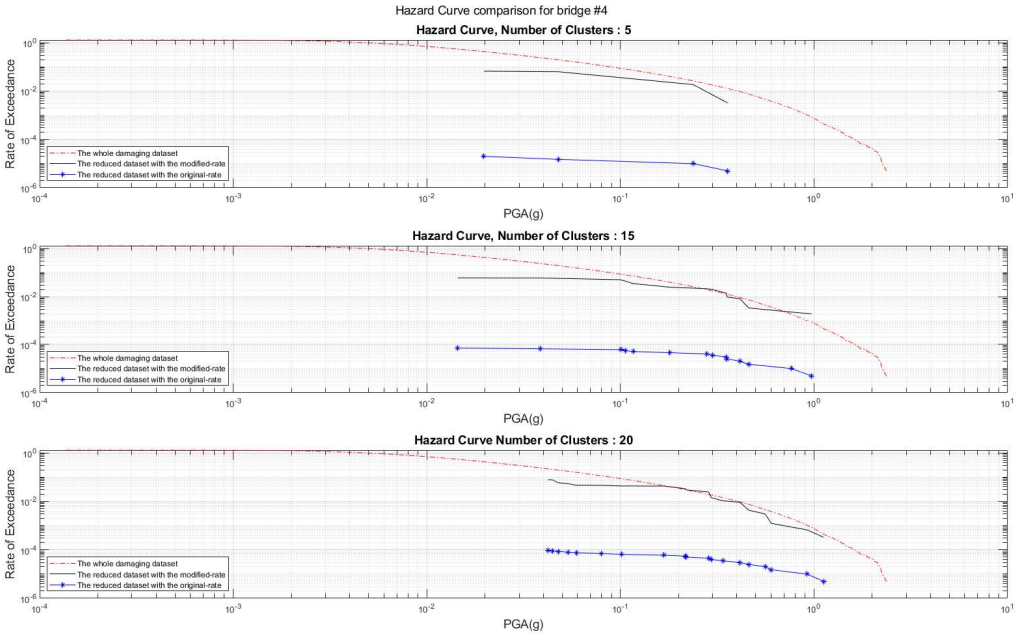


Figure 4-80 Hazard Curve Comparison for Bridge number 4 (for Case SCL and Hazard-Consistent), (unit of y axis is 1/year).

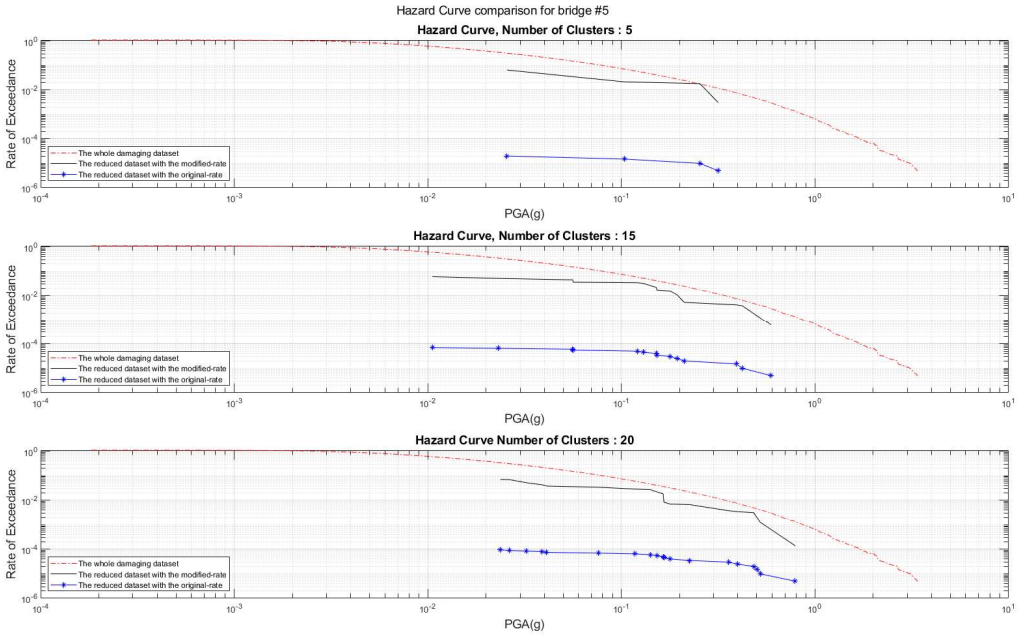


Figure 4-81 Hazard Curve Comparison for Bridge number 5 (for Case SCL and Hazard-Consistent), (unit of y axis is 1/year).

4.2.3.5 Discussion on the results for SCL

In this section the 4 different approaches of Risk-Consistent, Weighted Risk-Hazard, unweighted risk hazard and Hazard-Consistent approaches have been taking in clustering with K-means algorithm for the case of Simple Connectivity Loss (SCL) as the applied loss metric. like the trend that has been observed for two other loss metrics LDB and TD, the accuracy of the loss exceedance curve in comparison with the reference loss exceedance curve generated with the whole package of damaging events, is satisfactory in the Risk-Consistent approach. this accuracy has been affected by decreasing the dominancy of the loss values in the clustering of the events. however, in Risk-Consistent approach, the discrepancy of resulted hazard curves with the pertinent reference hazard curve for each bridge site that generated with the whole package of damaging earthquake events is highest among three other approaches. Since only the damaging events are used in clustering events in Hazard-Consistent approach, a large range of lower intensity measures are missing in the resulted hazard curve.

4.2.4 Results for Weighted Connectivity Loss (WCL)

Once again, the effect of 4 different approaches on the resulted reduced dataset have been compared and this time the weighted connectivity loss is used as a loss metric. following are the results of applying different approaches by using k-means clustering and the order of presenting is again in Risk-Consistent, Weighted Risk-Hazard, Unweighted Risk-Hazard, and the Hazard-Consistent approach. the 3 different number of clustering that are used in comparing the risk and hazard results of each approach are identical to the clustering numbers that was applied in the Risk-Consistent approach.

4.2.4.1 *Risk-Consistent approach with WCL loss metric*

Here, the only effective feature in calculation of distance metric, is the values of WCL. Again, Figure 4-82 and Figure 4-83 are to decide about the optimum clustering number which is assumed to be 100 here.

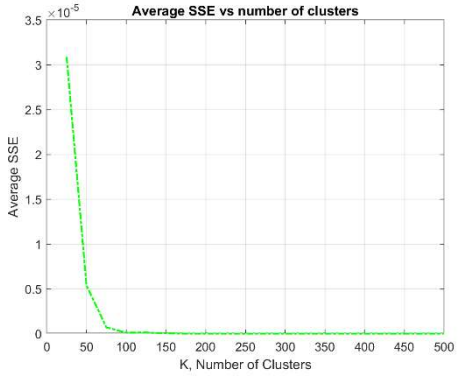


Figure 4-82 Elbow point for selecting the K (for Case WCL and Risk-Consistent)

Figure 4-83 Average SSE of loss vs ClustNo for 475 ,975,2475 (for Case WCL and Risk-Consistent)

Figure 4-84 shows the comparison of the resulted loss exceedance curves by applying Risk-Consistent approach for WCL loss metric in 3 different clustering number in K-means as minimum number of clustering in this study, optimum which has been chosen 100 for WCL based on error risk plot in above, and the maximum number of clustering that was implemented in this study. Like Risk-Consistent approach for 3 other loss metric LDB, TD and SCL, the results show satisfactory consistency with the reference risk curve that has been generated using the whole package of damaging events even in case of the minimum number of clustering 25.

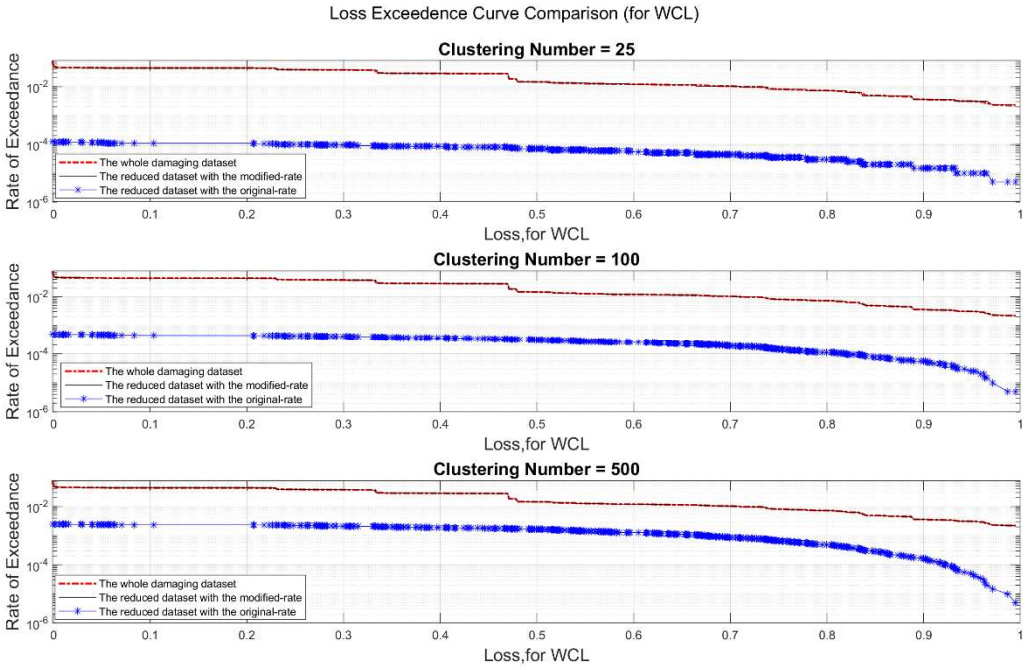


Figure 4-84 Comparison of LECs for reduced dataset with min, error based optimum, max number of clusters (for Case WCL and Risk-Consistent), (unit of y axis is 1/year).

Figure 4-85 to Figure 4-89 show the resulted hazard curve generated with he reduced dataset in this approach for the same 3 different clustering number that has been mentioned above. As it was expected, the resulted hazards are not compatible with the reference hazard curve because of the Risk-Consistent approach and not including the effect of hazard in clustering. Also, using only the damaging earthquake events in clustering cases of not having a range of lower intensity measures that do not affect the network in the resulted hazard curves. However, with more clustering numbers, like 500, the accuracy of resulted hazard curves can be improved to some extent.

again, like what has been examined for the previous 3 loss metrics, in the next section using the Weighted Risk-Hazard approach the effect of decreasing the dominancy of the risk in clustering and the resulted risk and hazard curves will be studied.

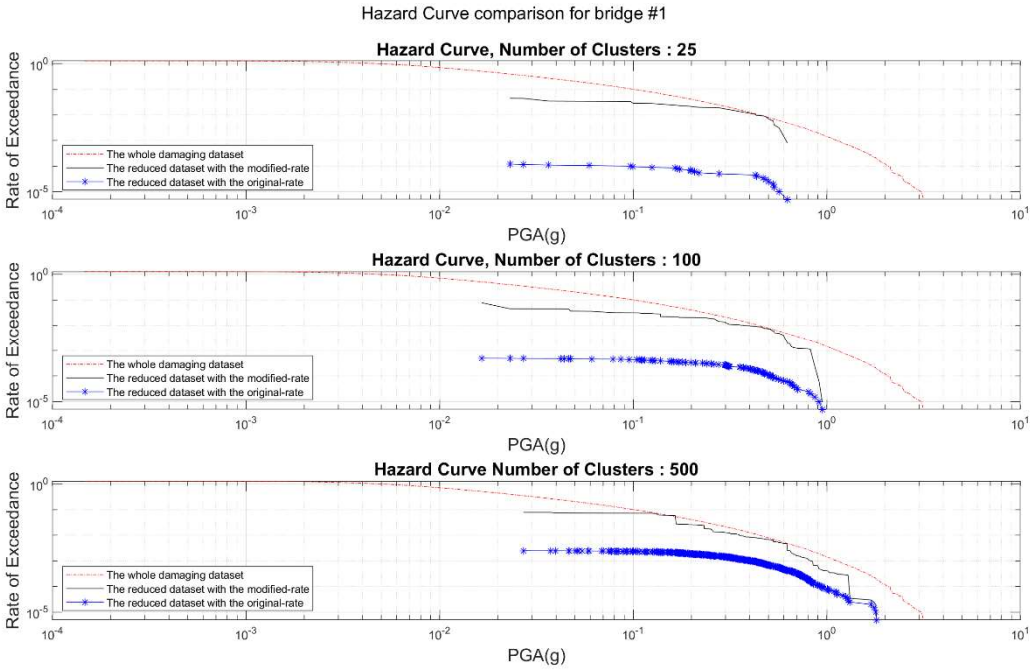


Figure 4-85 Hazard Curve Comparison for Bridge number 1 (for Case WCL and Risk-Consistent), (unit of y axis is 1/year).

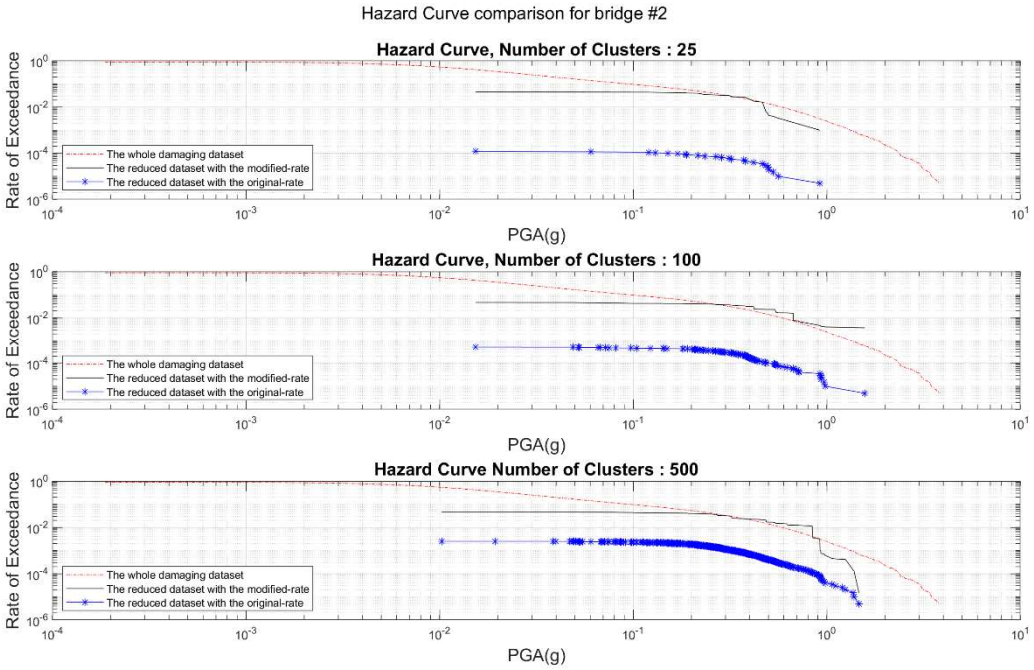


Figure 4-86 Hazard Curve Comparison for Bridge number 2 (for Case WCL and Risk-Consistent), (unit of y axis is 1/year).

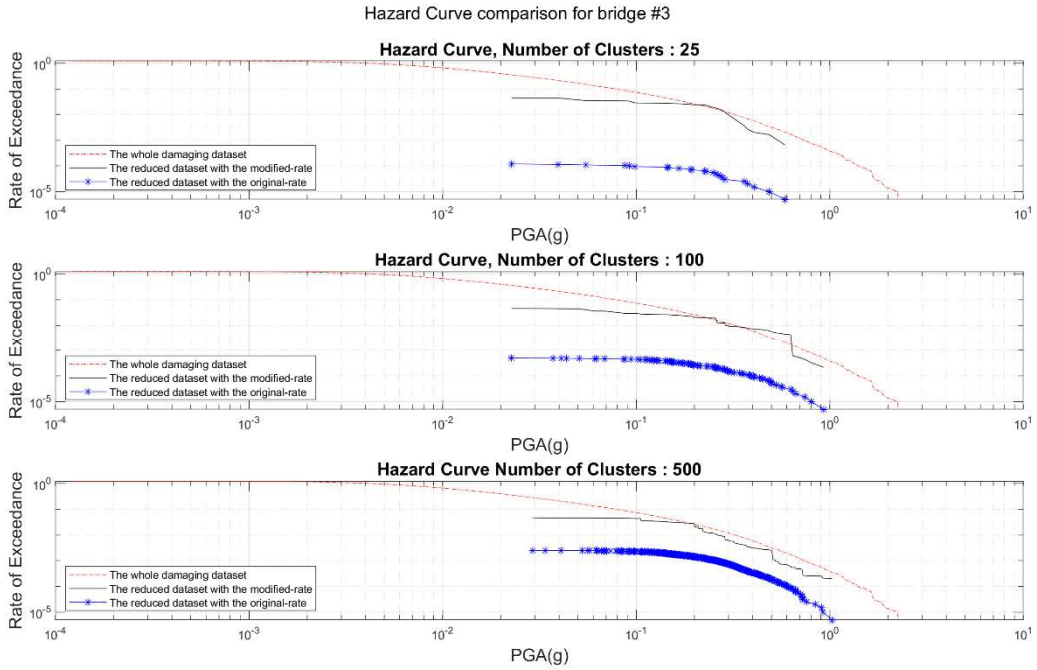


Figure 4-87 Hazard Curve Comparison for Bridge number 3 (for Case WCL and Risk-Consistent), (unit of y axis is 1/year).

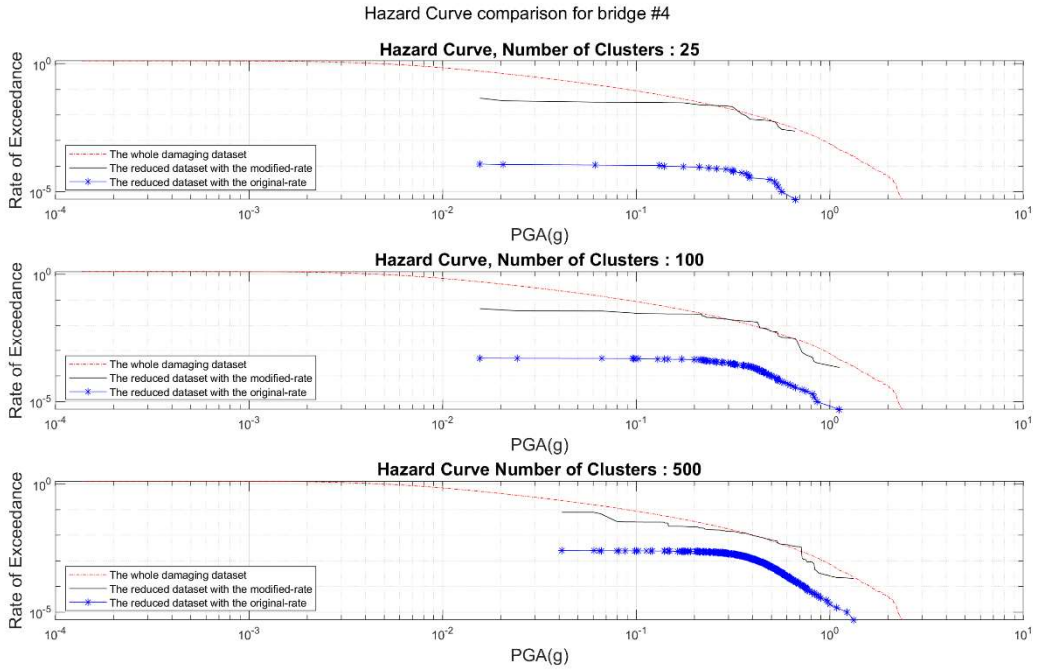


Figure 4-88 Hazard Curve Comparison for Bridge number 4 (for Case WCL and Risk-Consistent), (unit of y axis is 1/year).

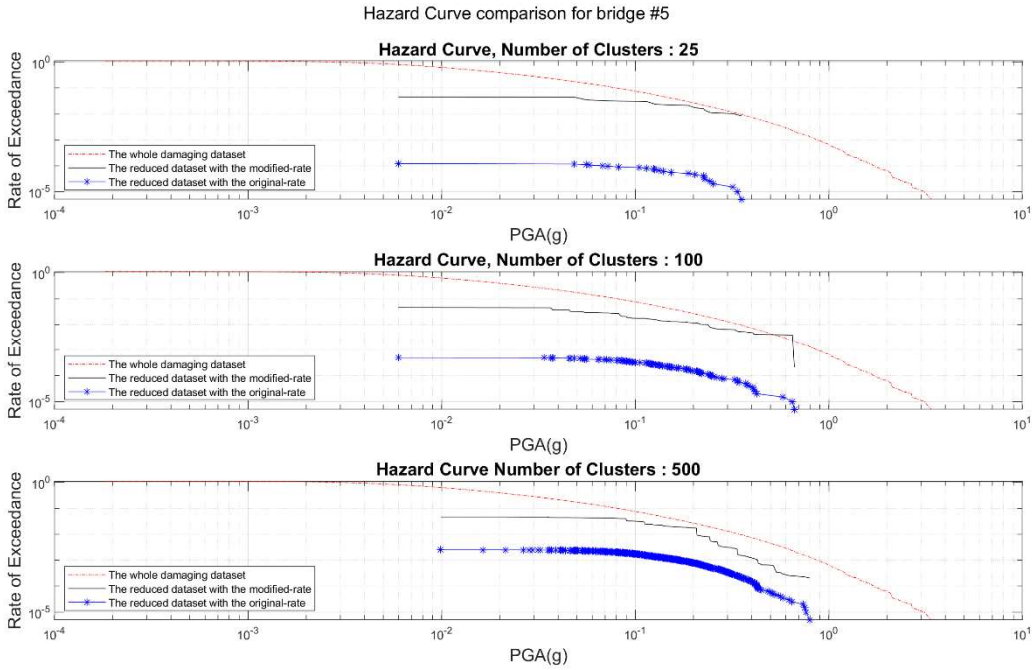


Figure 4-89 Hazard Curve Comparison for Bridge number 5 (for Case WCL and Risk-Consistent), (unit of y axis is 1/year).

4.2.4.2 Weighted Risk-Hazard approach with WCL loss metric

In this approach, like what has been tested for previous loss metrics, in calculation of distance metric as one of the key parameters in clustering, a weight of 50% has been allocated for the loss value (here is WCL) and 10% for the PGA values of each of 5 bridge locations. In this way, the effect of hazard is slightly included to the clustering process. The resulted risk and hazard curves that are shown in this section are by using same clustering numbers that was applied in the previous approach for WCL which was the Risk-consistent one and the clustering numbers were of 25, 100, and 500. In this way, a better comparison of the results of approaches is provided.

Figure 4-90 shows the resulted loss exceedance curve of reduced datasets with taking this approach in k-means clustering. although the dominancy of the risk in the clustering has been decreased by 50% in comparison with the Risk-Consistent approach, still the consistency of the risk curve of the reduced dataset with the reference risk curve that generated using the whole damaging package is satisfactory for the Risk-Consistent optimum number of clustering which is 100 here also for the maximum number of clustering as well.

Figure 4-91 to Figure 4-95 show the generated hazard curves with the Weighted Risk-Hazard approach using WCL as a loss metric. the improvement in hazard curves is observed for all 5 bridges in comparison with the results of Risk-Consistent approach. Especially in case of using the maximum number of clustering which is 500 in this study.

In the following section, the effect of Unweighted Risk-Consistent approach in case of using WCL as a loss metric is studied.

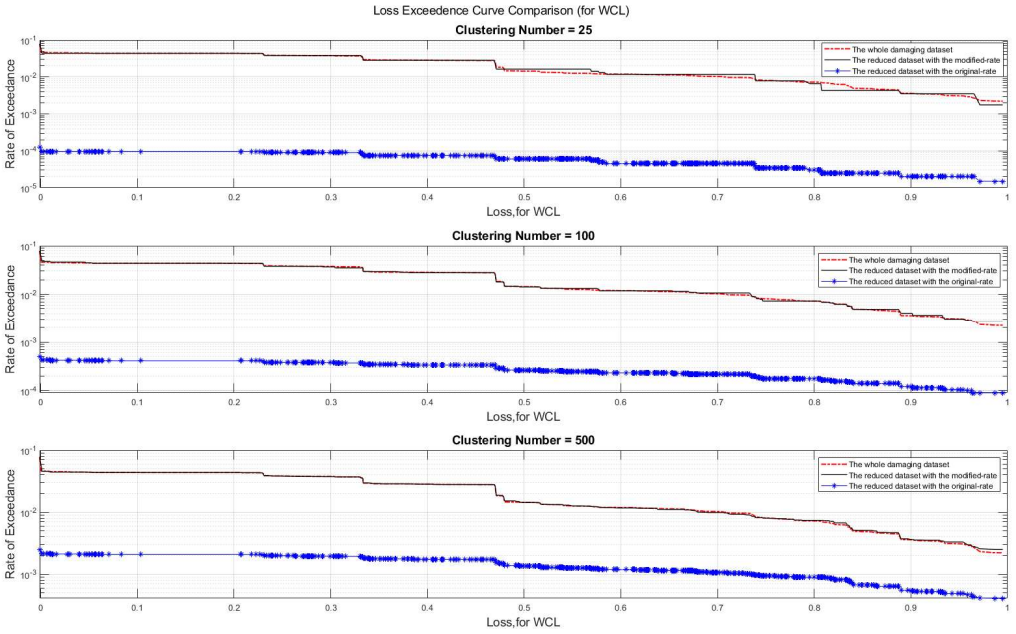


Figure 4-90 Comparison of LECs for reduced dataset with min, error based optimum, max number of clusters (for Case WCL and Weighted Risk-Hazard), (unit of y axis is 1/year).

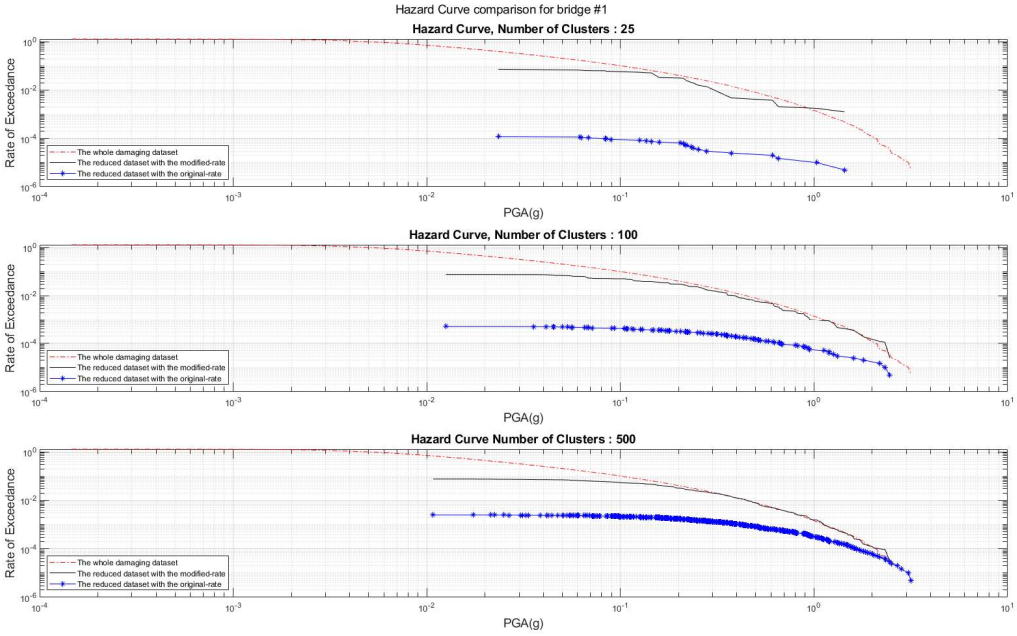


Figure 4-91 Hazard Curve Comparison for Bridge number 1 (for Case WCL and Weighted Risk-Hazard), (unit of y axis is 1/year).

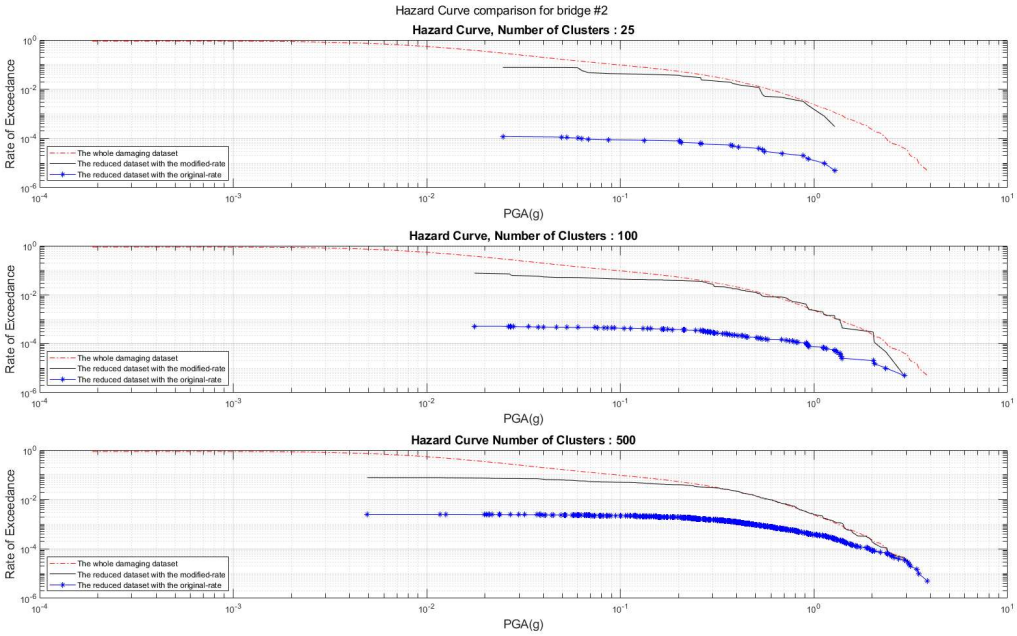


Figure 4-92 Hazard Curve Comparison for Bridge number 2 (for Case WCL and Weighted Risk-Hazard), (unit of y axis is 1/year).

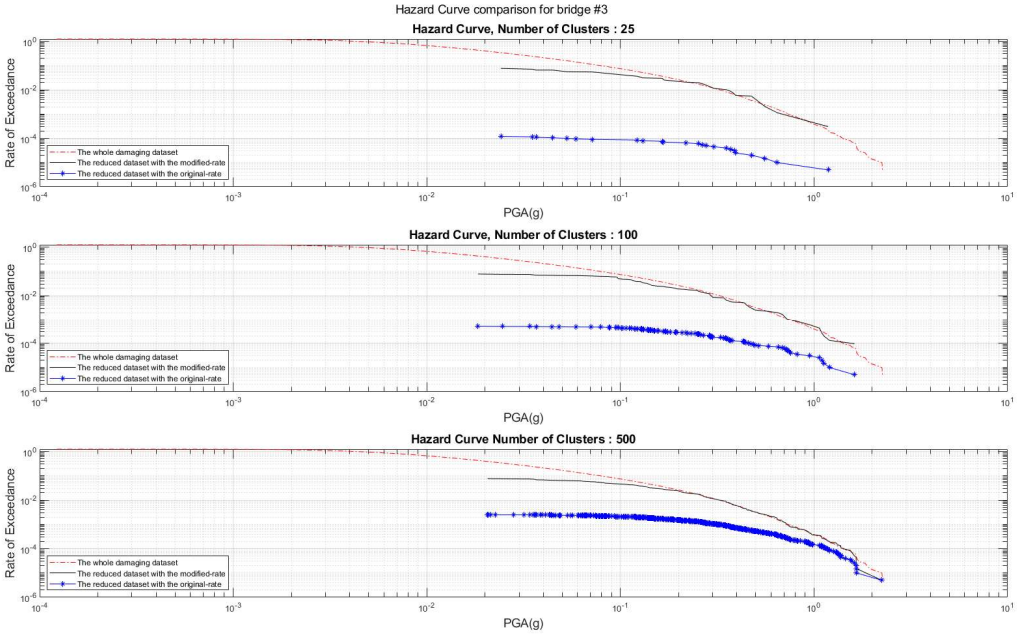


Figure 4-93 Hazard Curve Comparison for Bridge number 3 (for Case WCL and Weighted Risk-Hazard), (unit of y axis is 1/year).

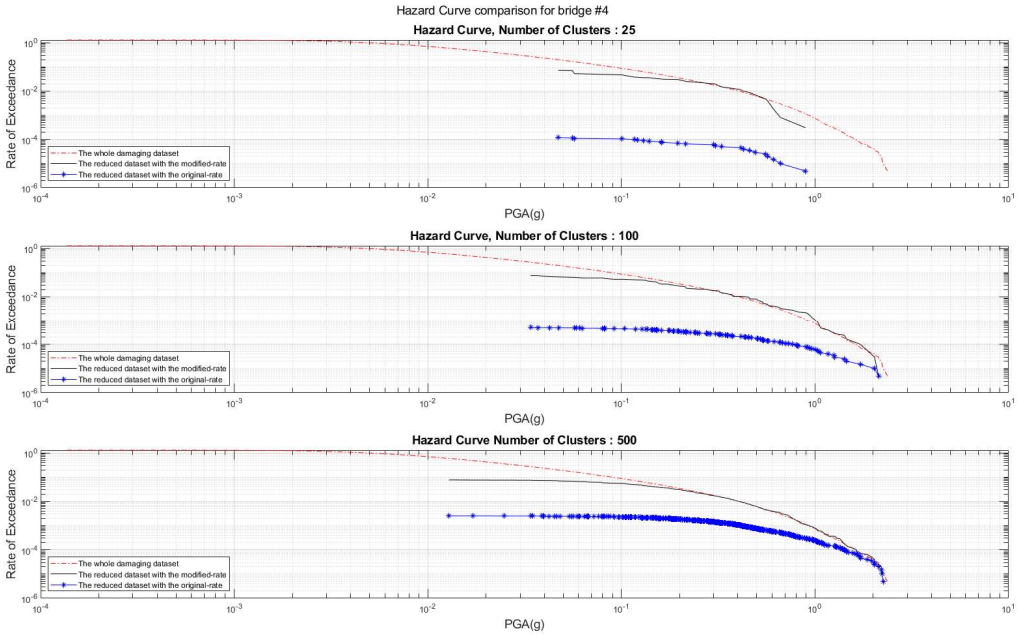


Figure 4-94 Hazard Curve Comparison for Bridge number 4 (for Case WCL and Weighted Risk-Hazard), (unit of y axis is 1/year).

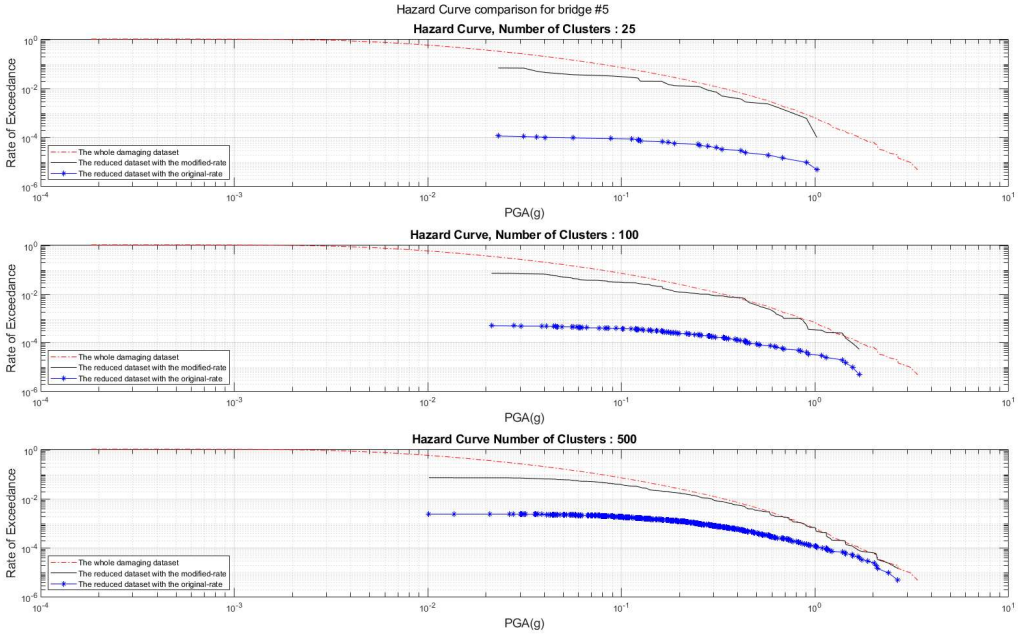


Figure 4-95 Hazard Curve Comparison for Bridge number 5 (for Case WCL and Weighted Risk-Hazard), (unit of y axis is 1/year).

4.2.4.3 Unweighted Risk-Hazard approach with WCL loss metric

In this approach like the previous section, 6 values are used in calculation of distance metric in clustering algorithm including the WCL value, and the PGA vales in 5 different bridge location. But unlike the Weighted Risk-Hazard approach, no weight has been allocated to any of these parameters. The results that have been shown in this approach are for results of clustering numbers identical to the previous approaches for WCL (25, 100, 500) to be comparable with each other.

Figure 4-96 shows the comparison of the resulted risk curve with the reduced dataset in 3 different sizes with the reference risk curve that has been generated using the whole damaging events package. As it can be observed in this figure, for clustering number of 25 and 100 the accuracy is affected in this approach especially for 25 clustering number. However, the consistency of the risk curve with the reference risk curve is satisfactory for the larger sizes of reduced dataset which is 500 here.

Figure 4-97 to Figure 4-101 show the resulted hazard curve with the reduced dataset in this approach and for the case of using WCL as a loss metric and the results of same mentioned 3 different clustering numbers has been compared. Slight improvements in the resulted hazard curves can be observed in comparison with the Weighted Risk-Hazard approach. For example, in case of bridge 3 and clustering numbers of 100, the error for higher PGAs has been slightly improved.

In the next section, the only effective feature in clustering will be the hazard, similar what has been studied for other 3 loss metrics in this chapter. The resulted risk curve for WCL and the hazard curves are generated and compared.

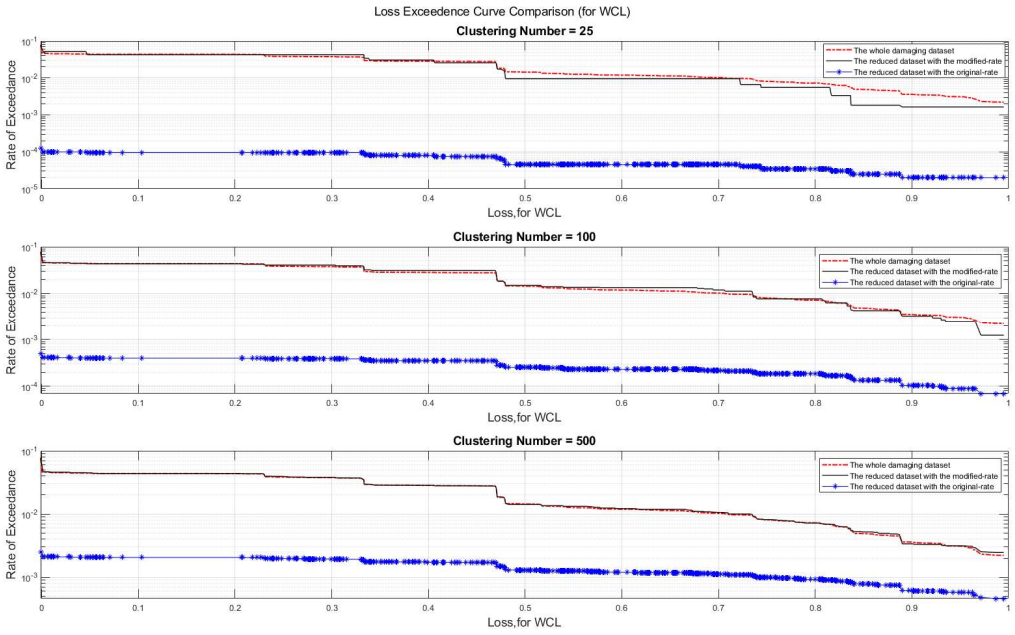


Figure 4-96 Comparison of LECs for reduced dataset with min, error based optimum, max number of clusters (for Case WCL and Unweighted Risk-Hazard), (unit of y axis is 1/year).

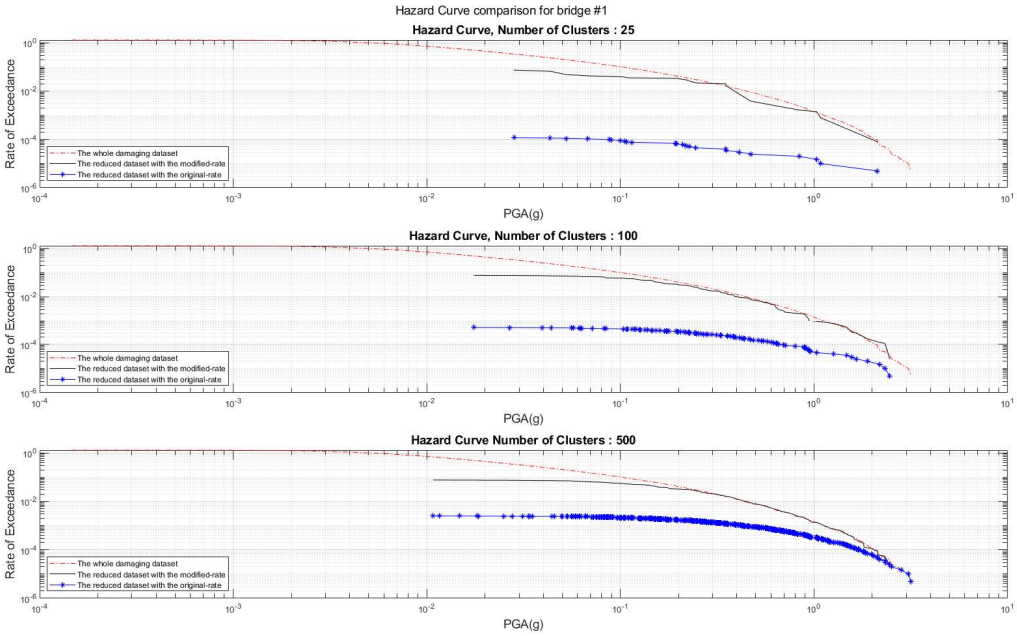


Figure 4-97 Hazard Curve Comparison for Bridge number 1 (for Case WCL and Unweighted Risk-Hazard), (unit of y axis is 1/year).

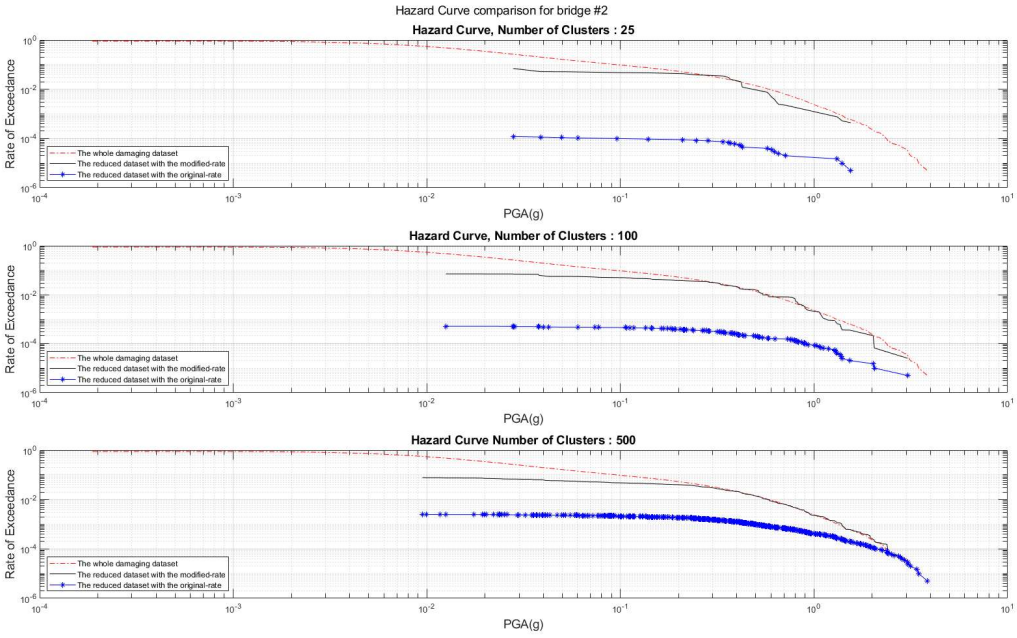


Figure 4-98 Hazard Curve Comparison for Bridge number 2 (for Case WCL and Unweighted Risk-Hazard), (unit of y axis is 1/year).

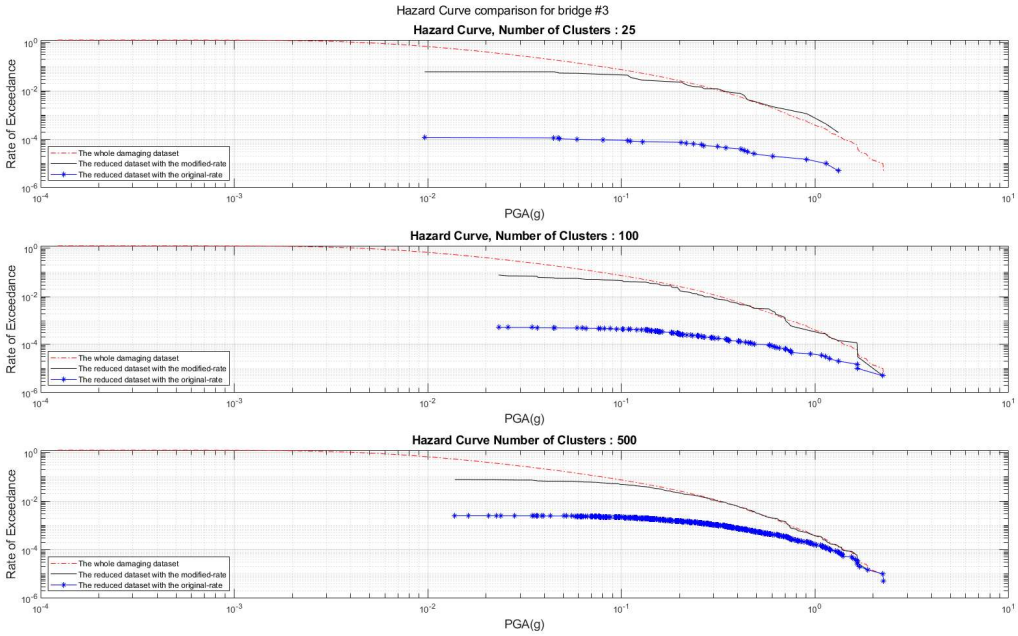


Figure 4-99 Hazard Curve Comparison for Bridge number 3 (for Case WCL and Unweighted Risk-Hazard), (unit of y axis is 1/year).

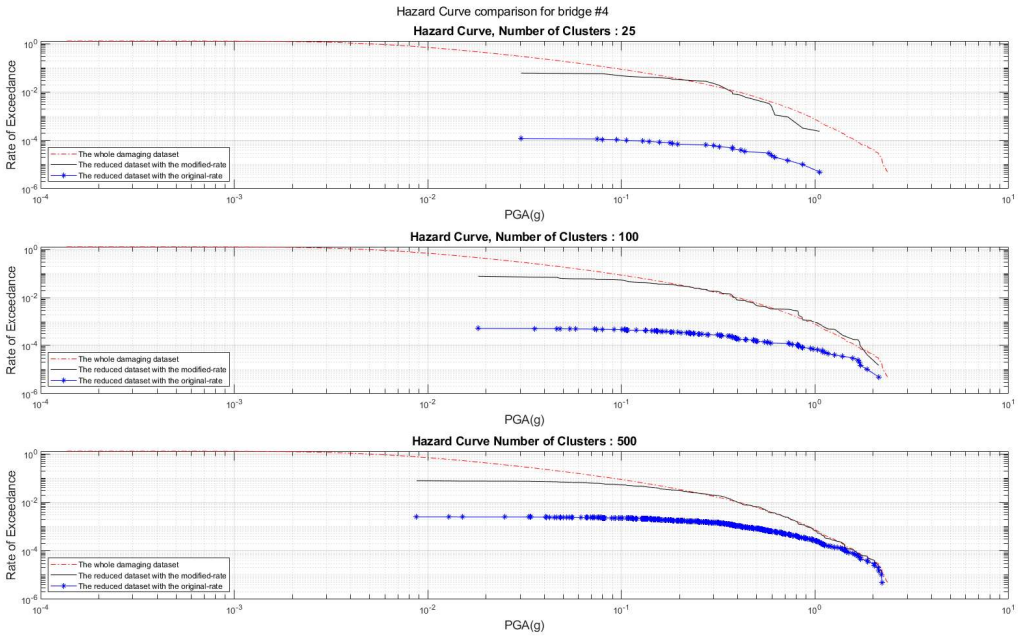


Figure 4-100 Hazard Curve Comparison for Bridge number 4 (for Case WCL and Unweighted Risk-Hazard), (unit of y axis is 1/year).

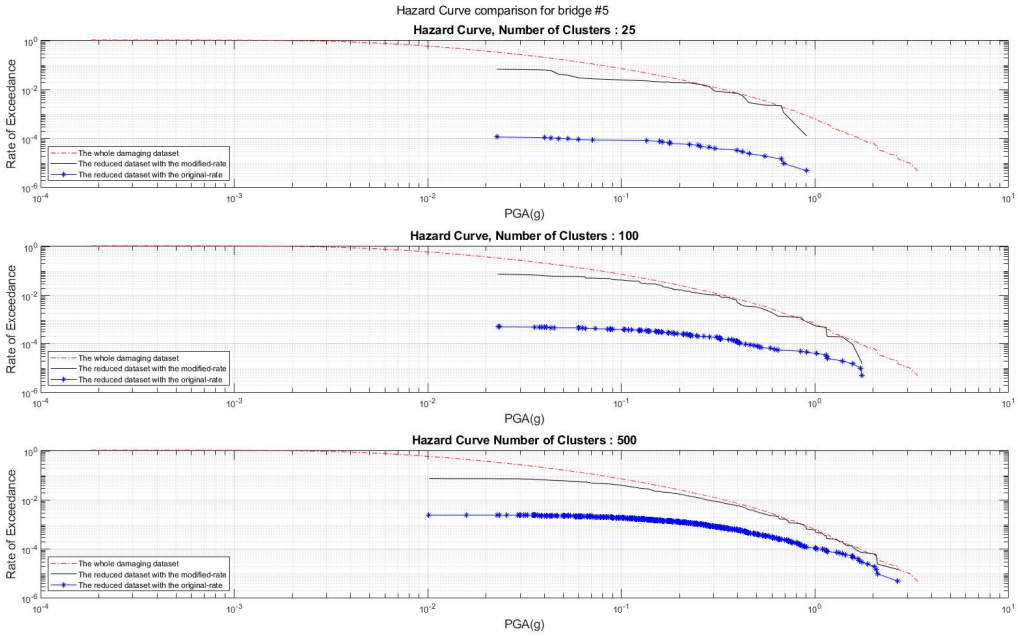


Figure 4-101 Hazard Curve Comparison for Bridge number 5 (for Case WCL and Unweighted Risk-Hazard), (unit of y axis is 1/year).

4.2.4.4 Hazard-Consistent approach with damaging events dataset using same clustering numbers with WCL Risk-Consistent approach

In this approach, only 5 PGA values of the bridge locations are used in calculation of distance metric of clustering with K-means and the resulted reduced dataset are shown in identical sized with the previous approaches for WCL as the applied loss metric to be compared with the results of previous 3 approaches.

Figure 4-102 shows the resulted risk curve for WCL with the reduced dataset using this approach and are in different sizes identical with the previous approaches for the case of WCL. As it can be observed, the compatibility of the reduced risk curves has been significantly affected especially in comparison with the results of Risk-Consistent approach. Still, by using larger number of clustering, like 500 here, this compatibility can be improved.

Figure 4-103 to Figure 4-107 show the resulted hazard curves with the reduced dataset using this approach in different clustering number identical to previous applied approaches for WCL. The compatibility of the resulted hazard curves with the reference hazard curve for each bridge has been slightly improved in comparison with the unweighted approach; however, it is not a major improvement. And a reason for it is that only the damaging event package is used in clustering which inherently does not include many events with resulting lower PGAs in the selected bridge locations.

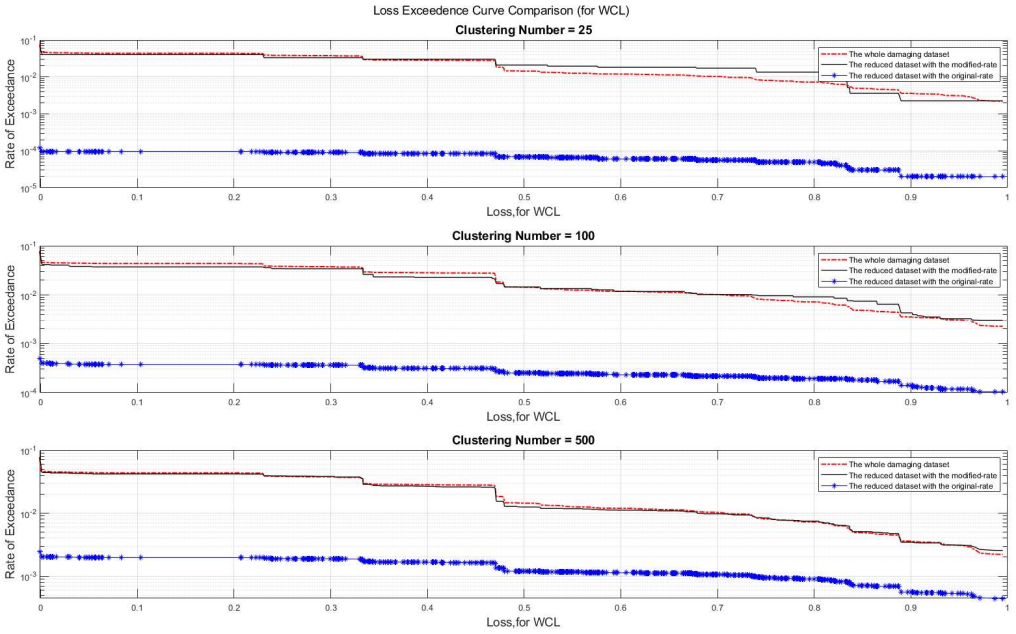


Figure 4-102 Comparison of LECs for reduced dataset with min, 100, max number of clusters (for Case WCL and Hazard-Consistent), (unit of y axis is 1/year).

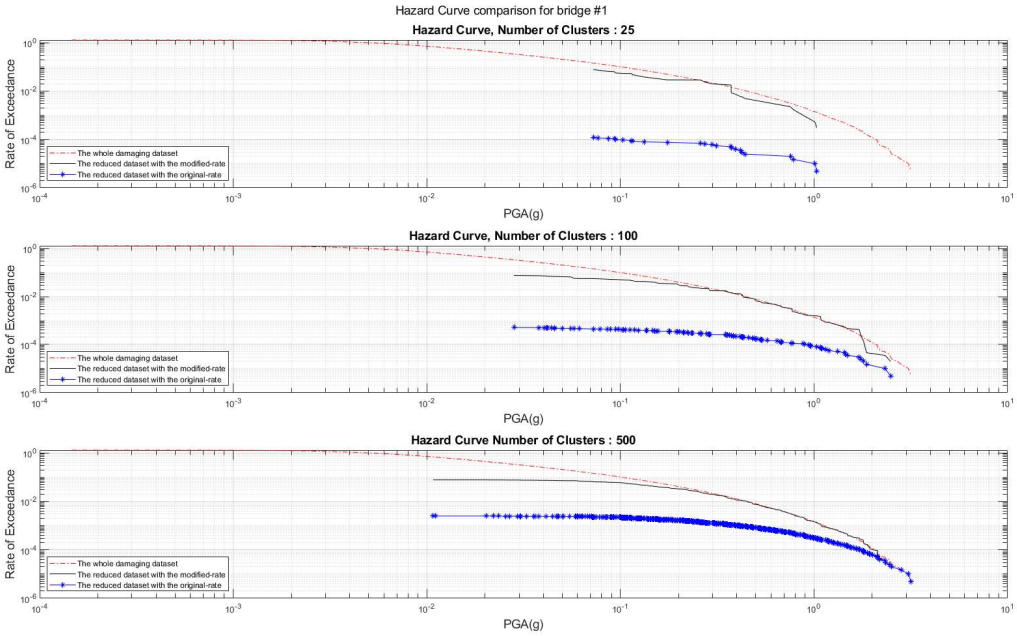


Figure 4-103 Hazard Curve Comparison for Bridge number 1 (to be compared with Case WCL and Hazard-Consistent), (unit of y axis is 1/year).

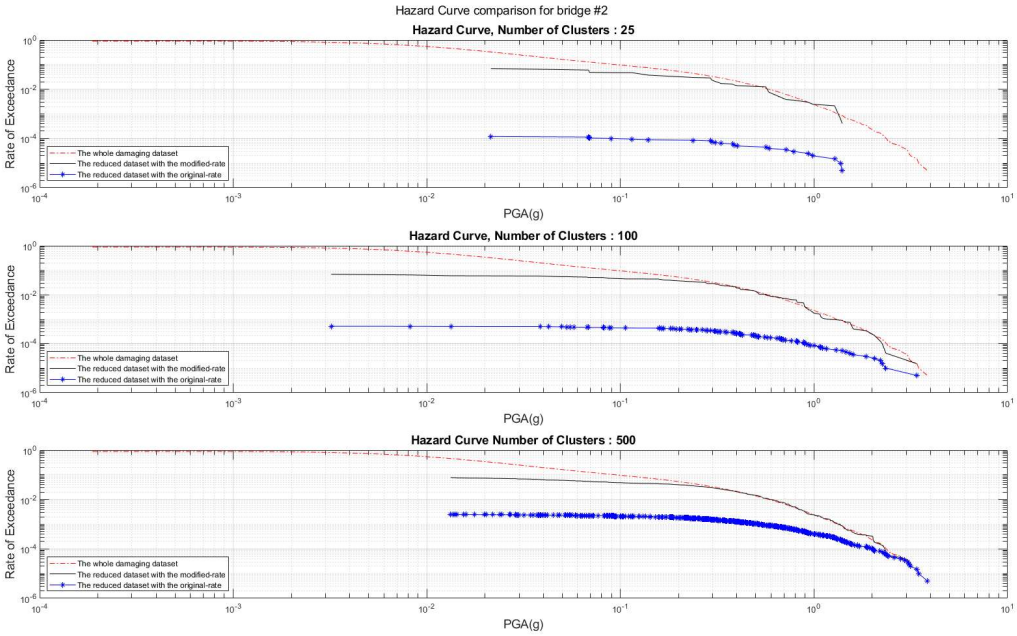


Figure 4-104 Hazard Curve Comparison for Bridge number 2 (to be compared with Case WCL and Hazard-Consistent), (unit of y axis is 1/year).

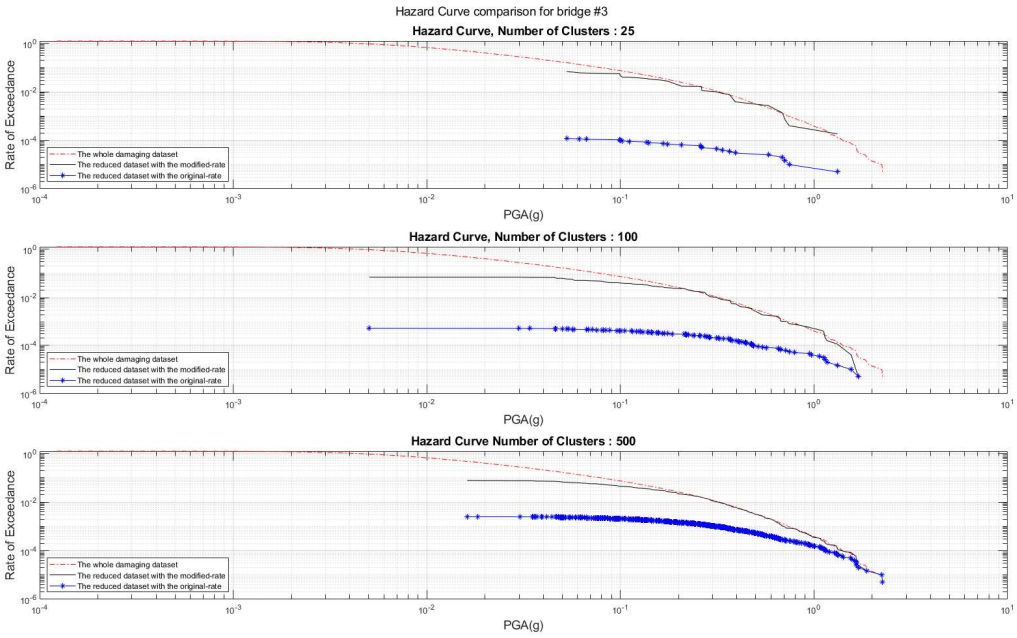


Figure 4-105 Hazard Curve Comparison for Bridge number 3 (to be compared with Case WCL and Hazard-Consistent), (unit of y axis is 1/year).

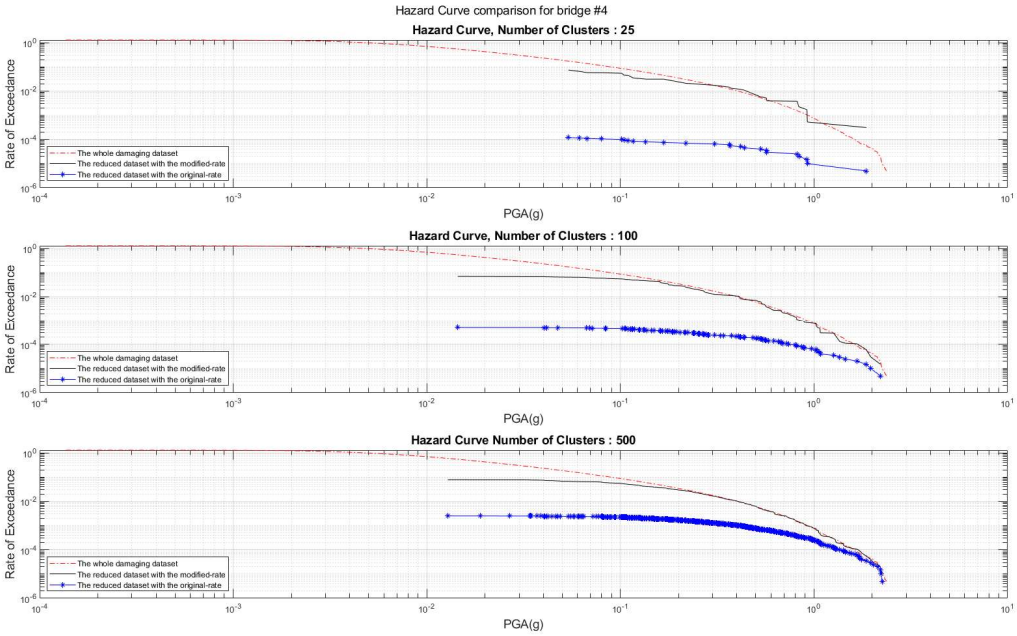


Figure 4-106 Hazard Curve Comparison for Bridge number 4 (to be compared with Case WCL and Hazard-Consistent), (unit of y axis is 1/year).

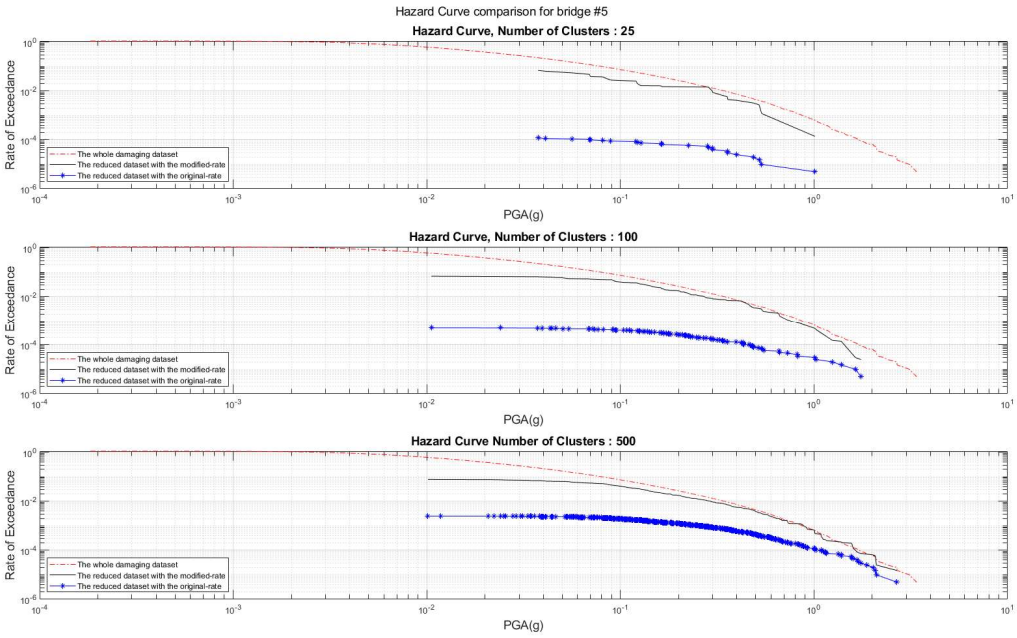


Figure 4-107 Hazard Curve Comparison for Bridge number 5 (to be compared with Case WCL and Hazard-Consistent), (unit of y axis is 1/year).

4.2.4.5 Discussions on results for WCL

4 different approaches of Risk-Consistent, Weighted Risk-Hazard, Unweighted Risk-Hazard and Hazard-Consistent have been studied for the case of using WCL as the loss metric. the trend of the effect of these approaches on the results were similar with what have been observed in previse 3 loss metrics LDB, TD and SCL. Means that, by decreasing dominancy of risk from in 4 different approaches from the Risk-Consistent to the Hazard-Consistent, the less accuracy in the resulted risk curves were observed and to get better compatibility, greater number of reduced datasets were required.

4.2.5 Results for Distance-based Weighted Connectivity Loss (DWCL)

The 4 different approaches of Risk-Consistent, Weighted Risk-Hazard, Unweighted Risk-Hazard, and Hazard-Consistent are studied here for the case of DWCL as the applied loss metric. the results of all approaches are presented for 3 different clustering numbers that has been chosen in the Risk-Consistent approach.

4.2.5.1 Risk-Consistent approach with DWCL loss metric

In this approach the only effective feature in clustering is the value of DWCL and to choose the appropriate number of clustering Figure 4-108 and Figure 4-109 show the elbow point and the risk-error-based plot. Accordingly, the appropriate number of clusters are assumed to be 100 here.

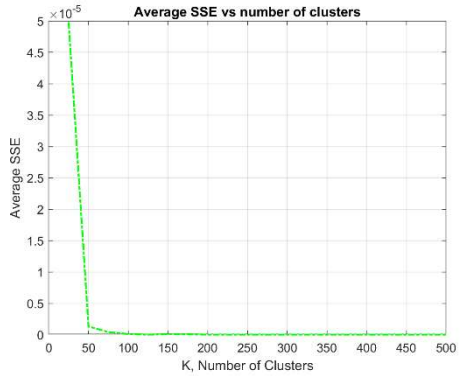


Figure 4-108 Elbow point for selecting the K (for Case DWCL and Risk-Consistent)

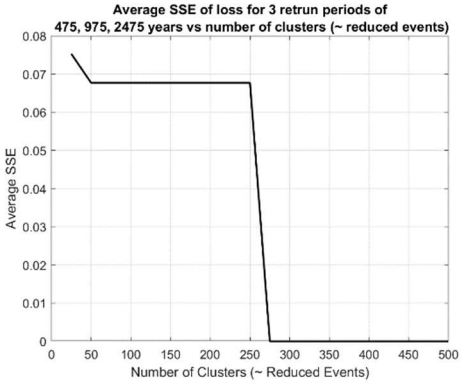


Figure 4-109 Average SSE of loss vs ClustNo for 475 ,975,2475 (for Case DWCL and Risk-Consistent)

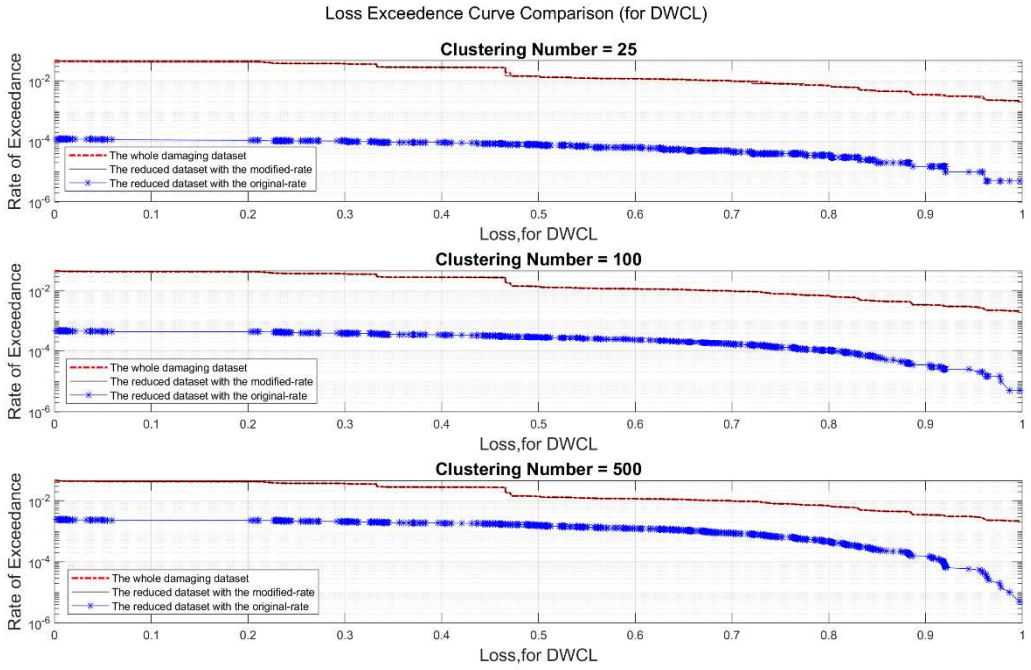


Figure 4-110 Comparison of LECs for reduced dataset with min, error based optimum, max number of clusters (for Case DWCL and Risk-Consistent), (unit of y axis is 1/year).

Figure 4-110 shows the resulted loss exceedance curve with the reduced dataset in this approach in 3 different sizes of 25, 100 and 500. As it can be observed in this figure, like results of Risk-Consistent approach in other loss metrics, the compatibility of the resulted risk curve with reduced dataset with the reference risk curve using the whole damaging events package is satisfactory event for the minimum number of clustering which is 25 here. Figure 4-111 to Figure 4-115 show the resulted hazard curves with the reduced data set in 3 different sizes that were mentioned above. They results are not consistent with the reference hazard curve for each bridge location yet by increasing the number of clusters can improve in the resulted hazard curve.

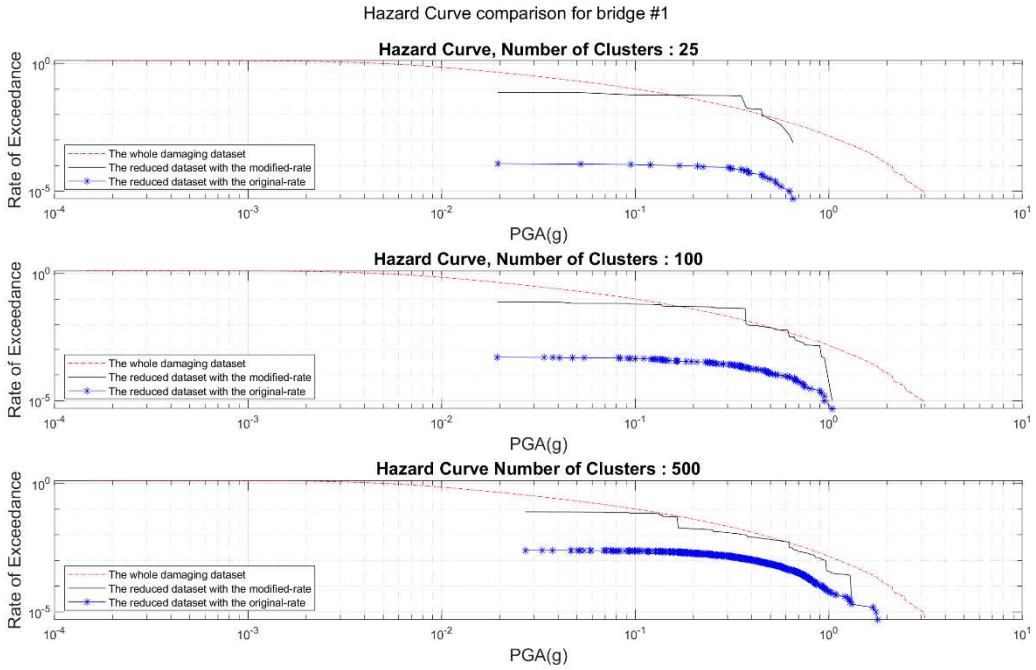


Figure 4-111 Hazard Curve Comparison for Bridge number 1 (for Case DWCL and Risk-Consistent), (unit of y axis is 1/year).

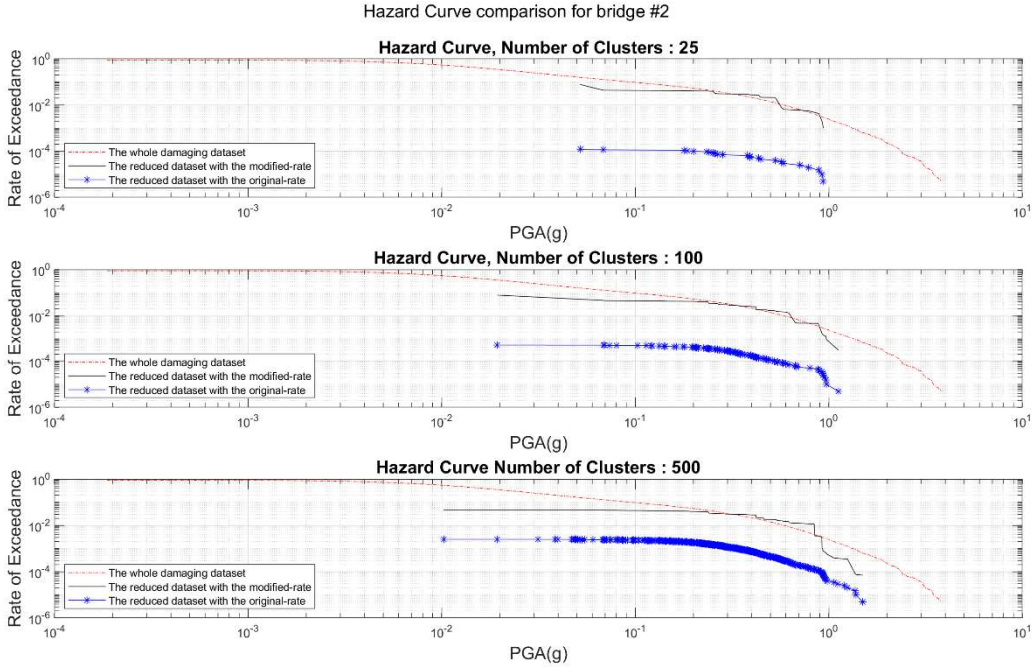


Figure 4-112 Hazard Curve Comparison for Bridge number 2 (for Case DWCL and Risk-Consistent), (unit of y axis is 1/year).

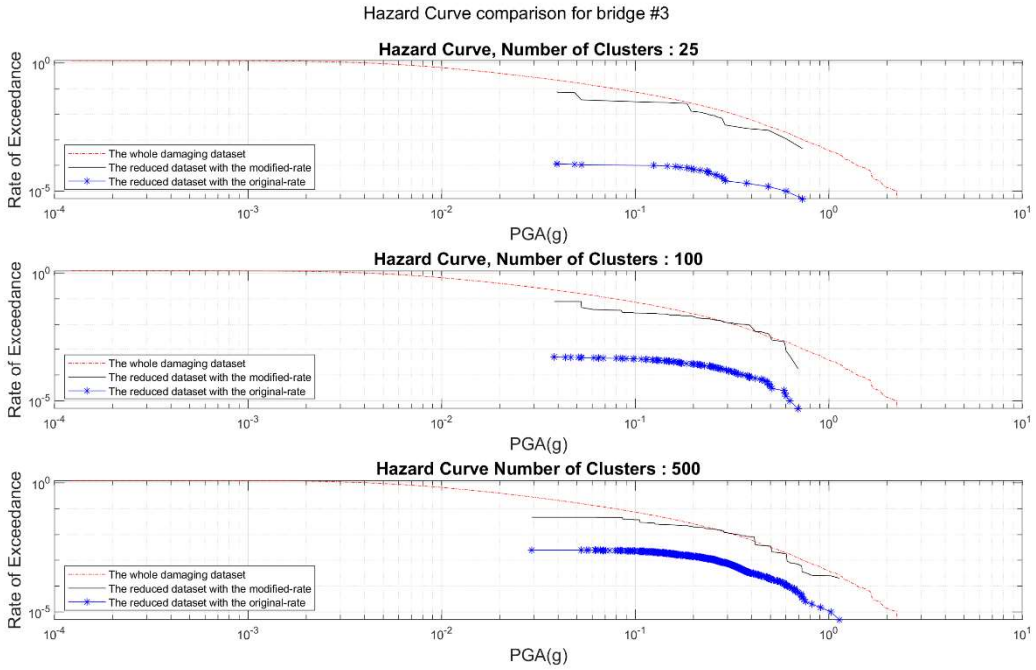


Figure 4-113 Hazard Curve Comparison for Bridge number 3 (for Case DWCL and Risk-Consistent), (unit of y axis is 1/year).

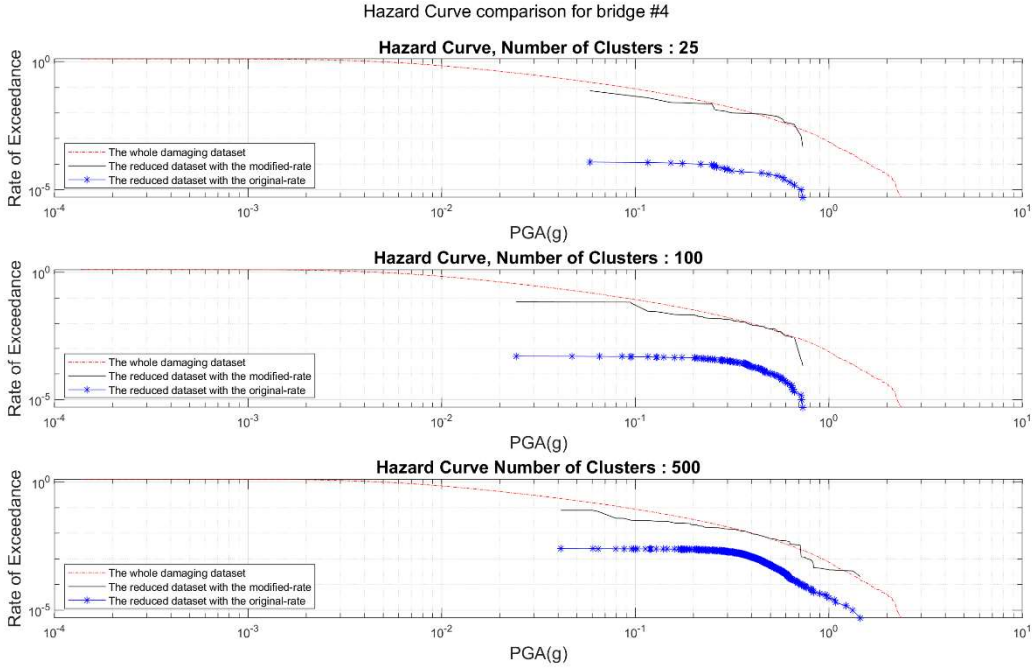


Figure 4-114 Hazard Curve Comparison for Bridge number 4 (for Case DWCL and Risk-Consistent), (unit of y axis is 1/year).

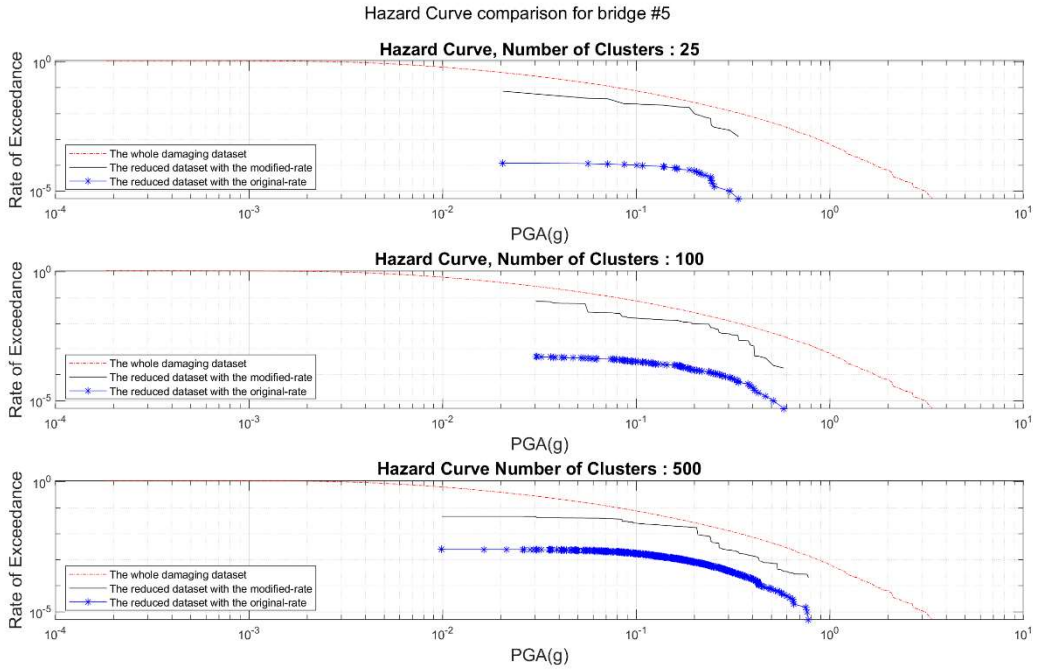


Figure 4-115 Hazard Curve Comparison for Bridge number 5 (for Case DWCL and Risk-Consistent), (unit of y axis is 1/year).

4.2.5.2 Weighted Risk-Hazard approach with DWCL loss metric

In this approach, like what has been implemented for the other 4 loss metrics, a weight of 50% is allocated to the Loss value which is DWCL here and 10% for PGA values of each bridge location in calculations of distance metric as a required parameter in clustering which is the K-means clustering method in this section. The results of this approach are shown and compared in 3 different sizes of clustering that were used in the previous Risk-Consistent approach for DWCL and the numbers were 25, 100, and 500.

Figure 4-116 shows the comparison of the resulted risk curve with the reduced dataset in this approach and the reference risk curve that generated by using the whole package of damaging events. some errors can be observed in smaller datasets like 25 and 100 but still the error in 100 clustering number can be considered as a negligible error.

Figure 4-117 to Figure 4-121 show the resulted hazard curve using reduced dataset in this approach in 3 different sizes of 25, 100, 500 number of events in datasets. Some improvement in the resulted hazard curves can be observed in comparison with the Risk-Consistent approach for WCL. For example, the improvement in hazard curve of bridge number 1 with clustering number of 100 in comparison with the same clustering number of bridge 1 in Risk-Consistent approach. more satisfactory hazard curves for less frequent intensity measure are obtainable with larger sizes of dataset like 500 in these figures.

In next section, using the Unweighted Risk-Hazard approach, the effect of decreasing the dominancy of the risk in clustering is studied and compared.

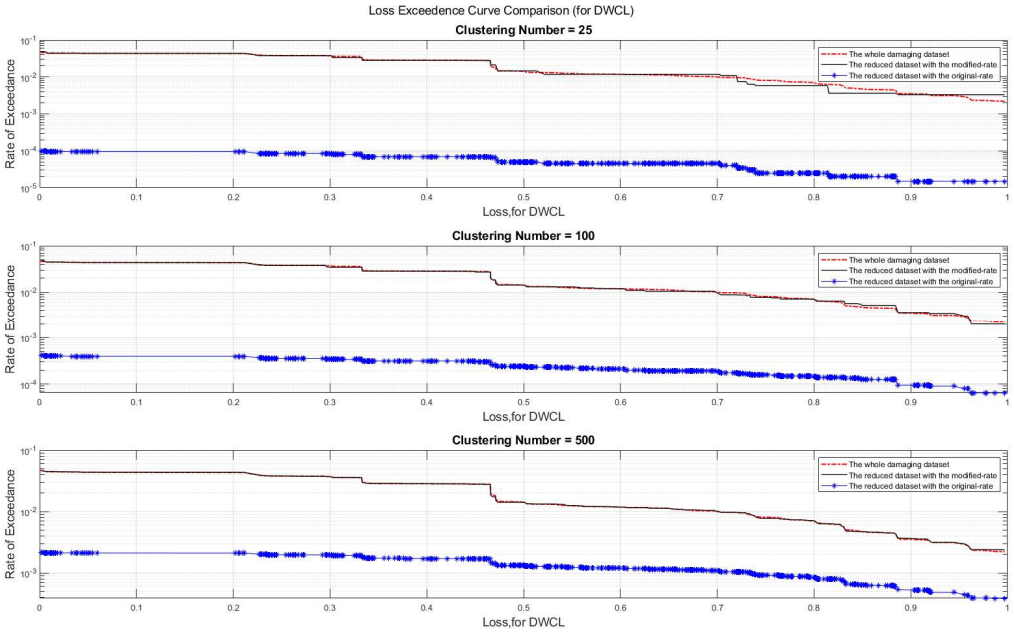


Figure 4-116 Comparison of LECs for reduced dataset with min, error based optimum, max number of clusters (for Case DWCL and Weighted Risk-Hazard), (unit of y axis is 1/year).

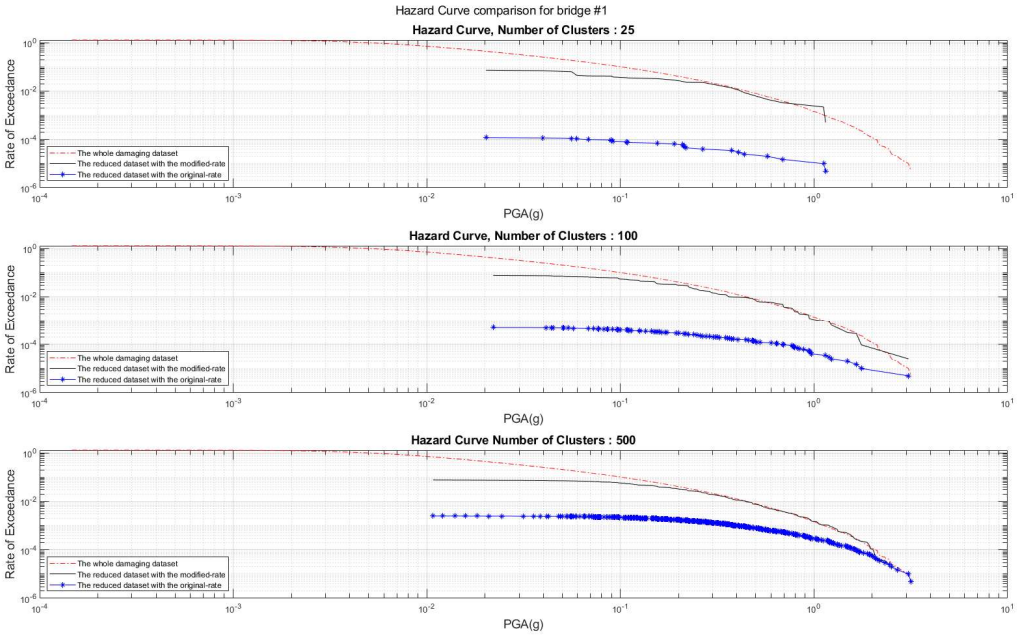


Figure 4-117 Hazard Curve Comparison for Bridge number 1 (for Case DWCL and Weighted Risk-Hazard), (unit of y axis is 1/year).

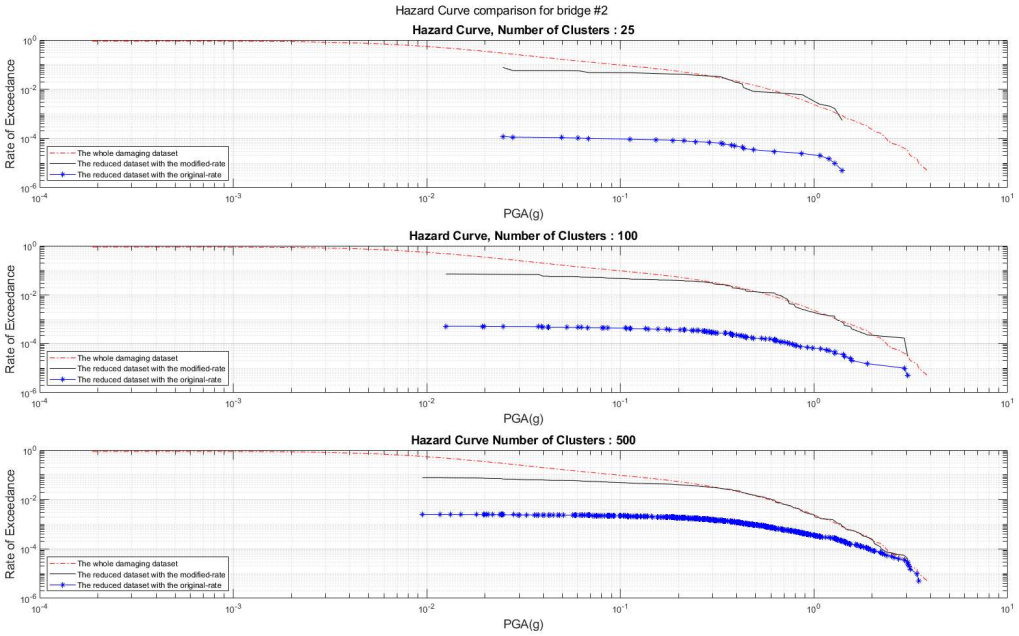


Figure 4-118 Hazard Curve Comparison for Bridge number 2 (for Case DWCL and Weighted Risk-Hazard), (unit of y axis is 1/year).

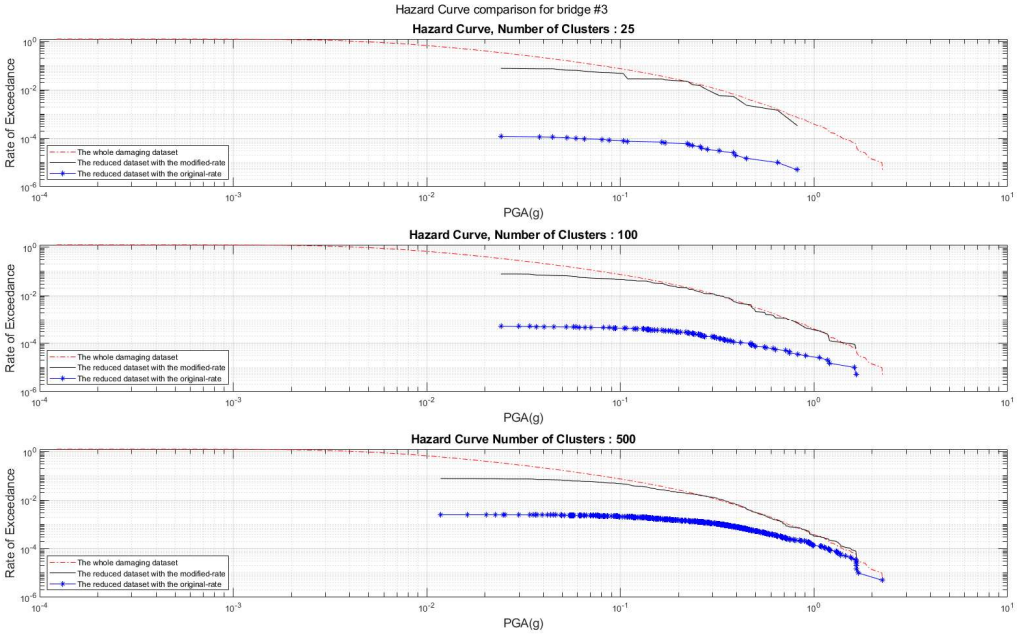


Figure 4-119 Hazard Curve Comparison for Bridge number 3 (for Case DWCL and Weighted Risk-Hazard), (unit of y axis is 1/year).

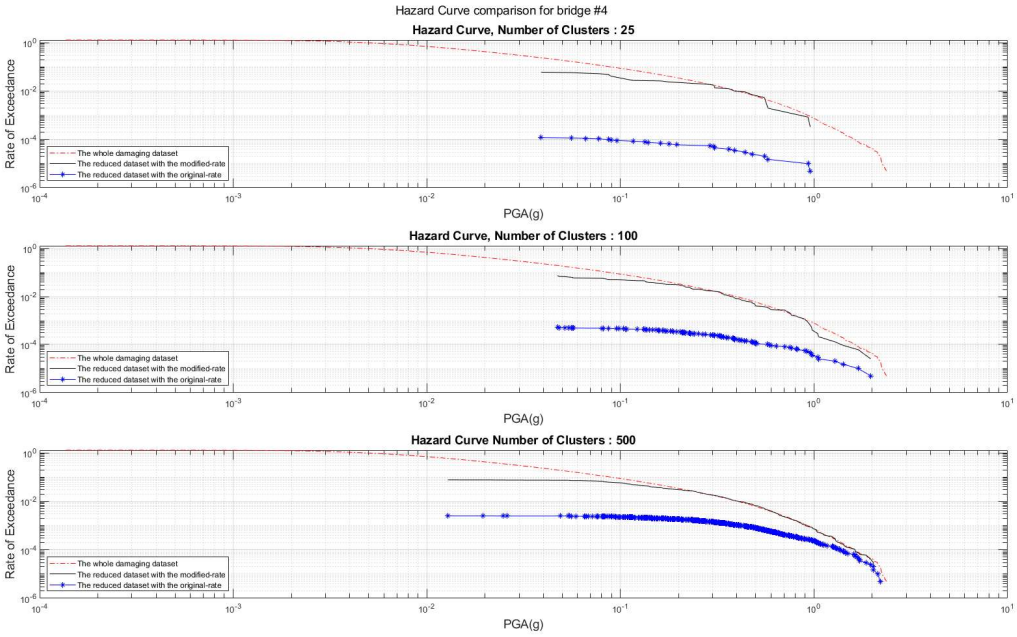


Figure 4-120 Hazard Curve Comparison for Bridge number 4 (for Case DWCL and Weighted Risk-Hazard), (unit of y axis is 1/year).

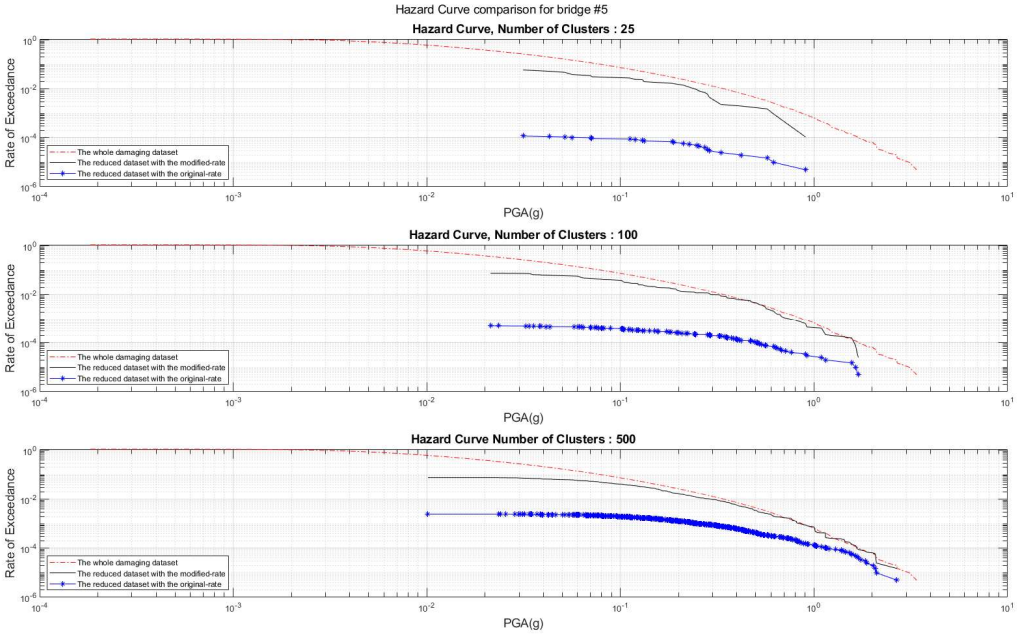


Figure 4-121 Hazard Curve Comparison for Bridge number 5 (for Case DWCL and Weighted Risk-Hazard), (unit of y axis is 1/year).

4.2.5.3 Unweighted Risk-Hazard approach with DWCL loss metric

Here the Unweighted Risk-Hazard approach is studied for the case of DWCL which is similar previous approach but without allocating any weight for the features. Figure 4-122 shows the resulted risk curves with the reduced datasets in this approach.

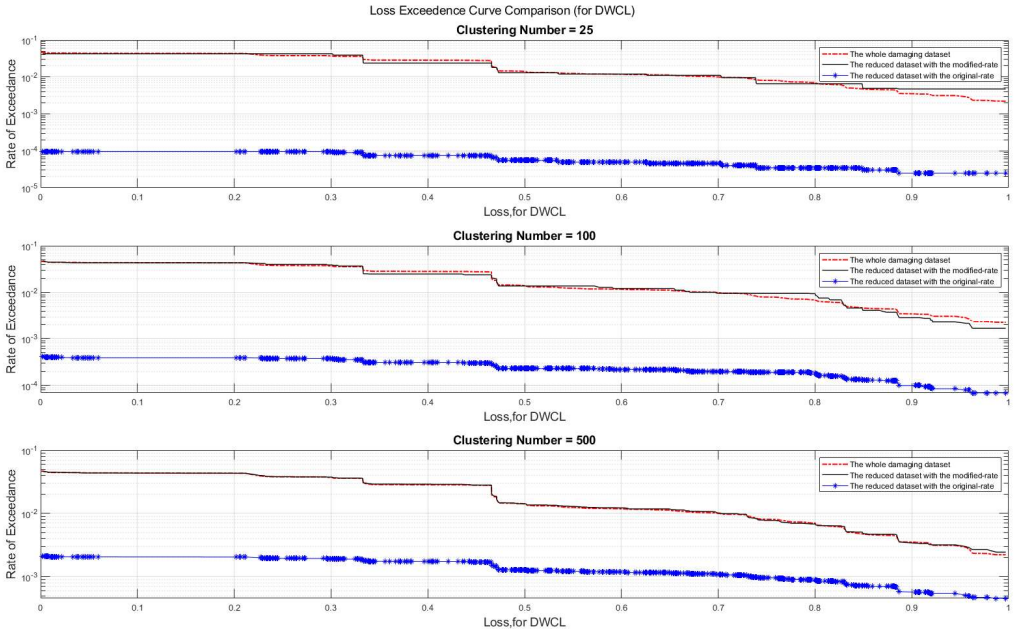


Figure 4-122 Comparison of LECs for reduced dataset with min, error based optimum, max number of clusters (for Case DWCL and Unweighted Risk-Hazard), (unit of y axis is 1/year).

As it can be observed in Figure 4-122, less accurate risk curves are resulted especially for clustering numbers of 25 and 100. However, satisfactory curves are possible by increasing the number of clusters like 500 in this figure. Figure 4-123 to Figure 4-127 show the resulted hazard curves using the reduced dataset in this approach in 3 different sizes of dataset of 25, 100, and 500.

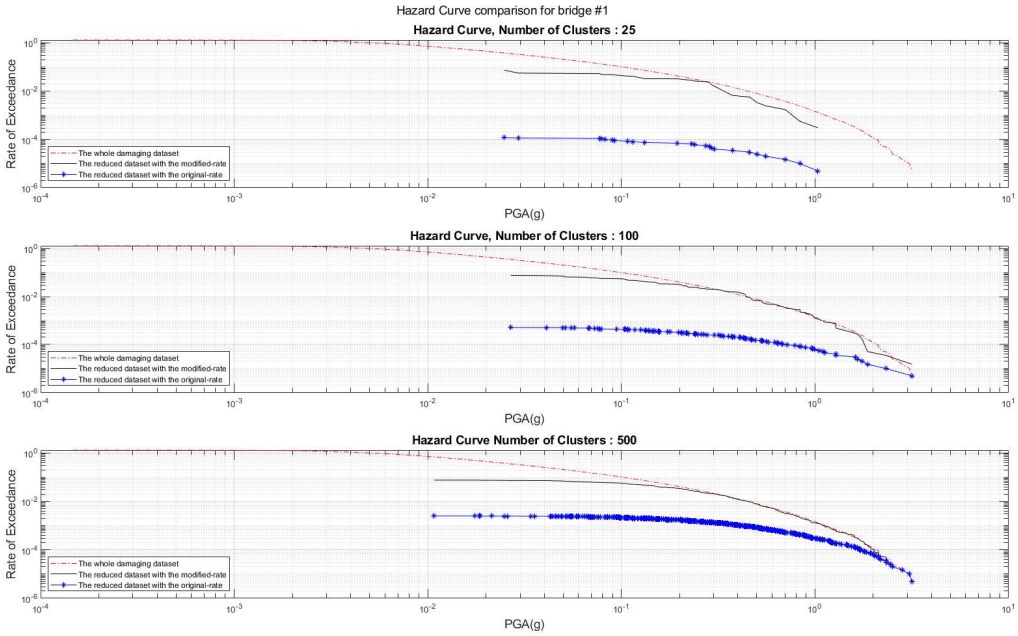


Figure 4-123 Hazard Curve Comparison for Bridge number 1 (for Case DWCL and Unweighted Risk-Hazard), (unit of y axis is 1/year).

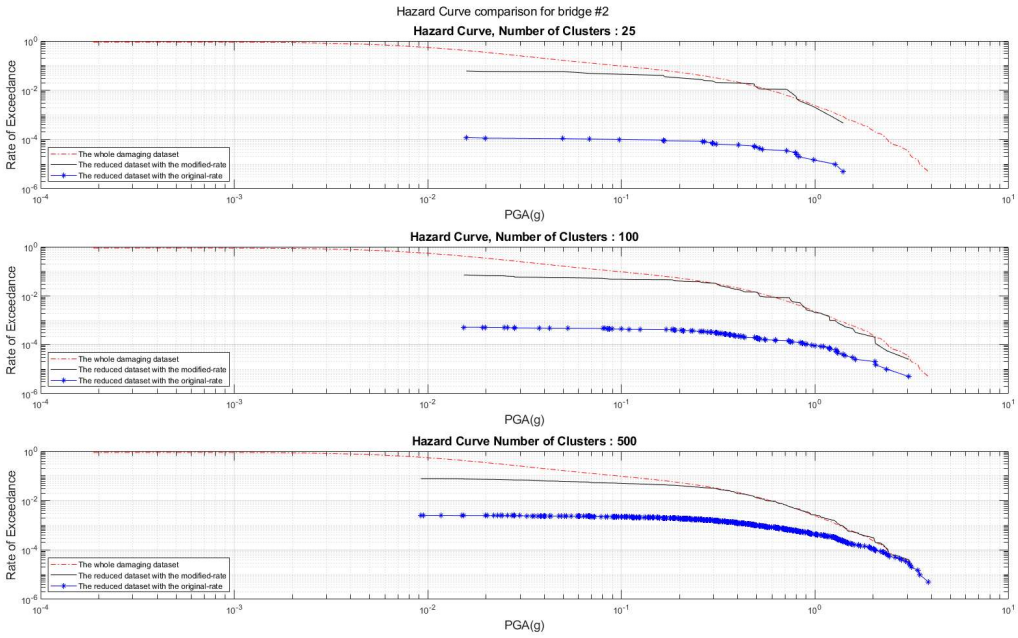


Figure 4-124 Hazard Curve Comparison for Bridge number 2 (for Case DWCL and Unweighted Risk-Hazard), (unit of y axis is 1/year).

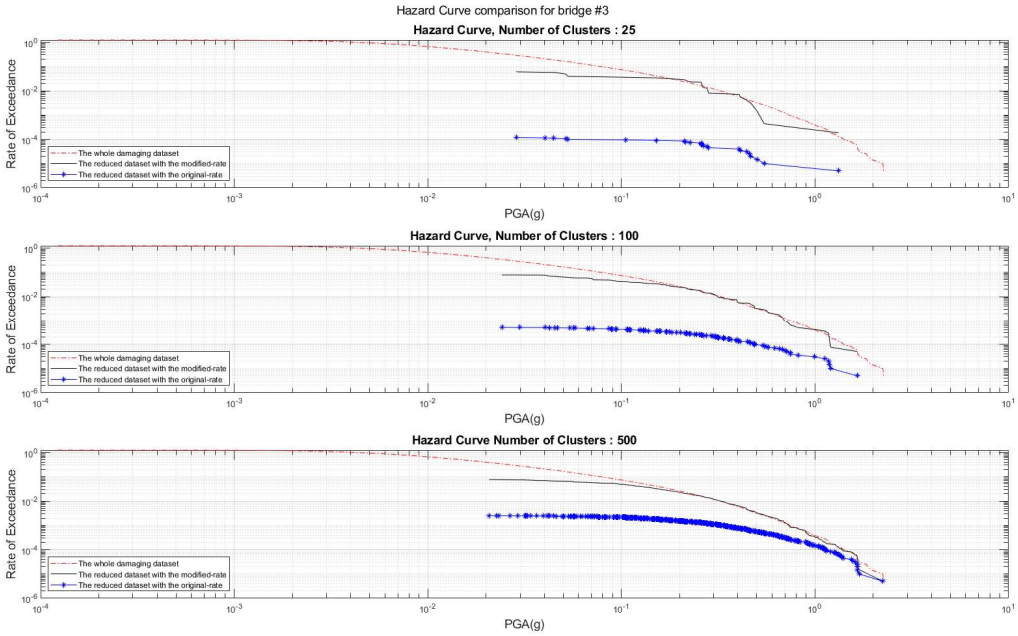


Figure 4-125 Hazard Curve Comparison for Bridge number 3 (for Case DWCL and Unweighted Risk-Hazard), (unit of y axis is 1/year).

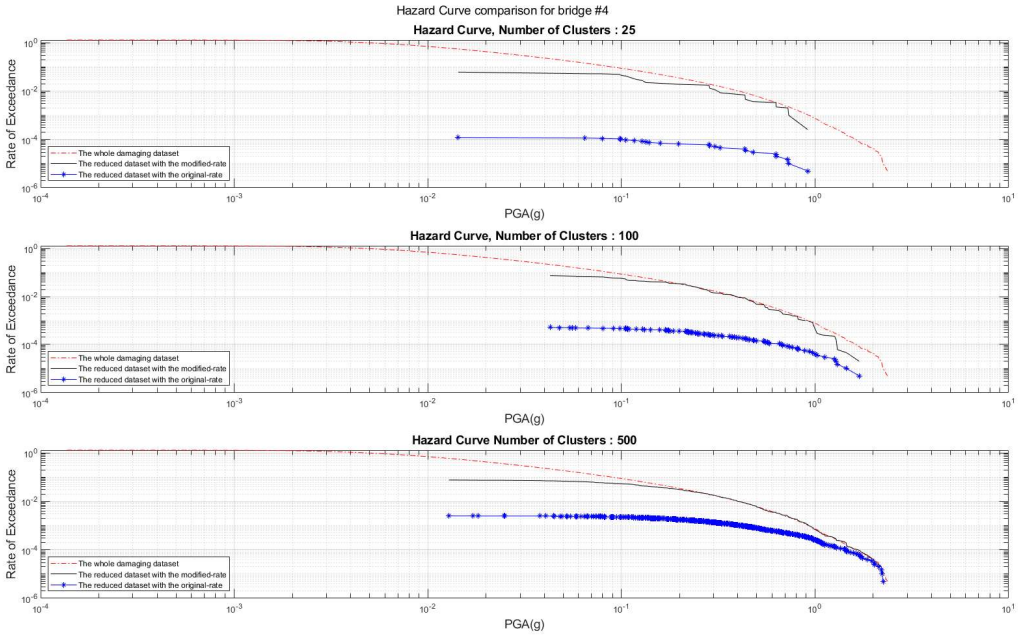


Figure 4-126 Hazard Curve Comparison for Bridge number 4 (for Case DWCL and Unweighted Risk-Hazard), (unit of y axis is 1/year).

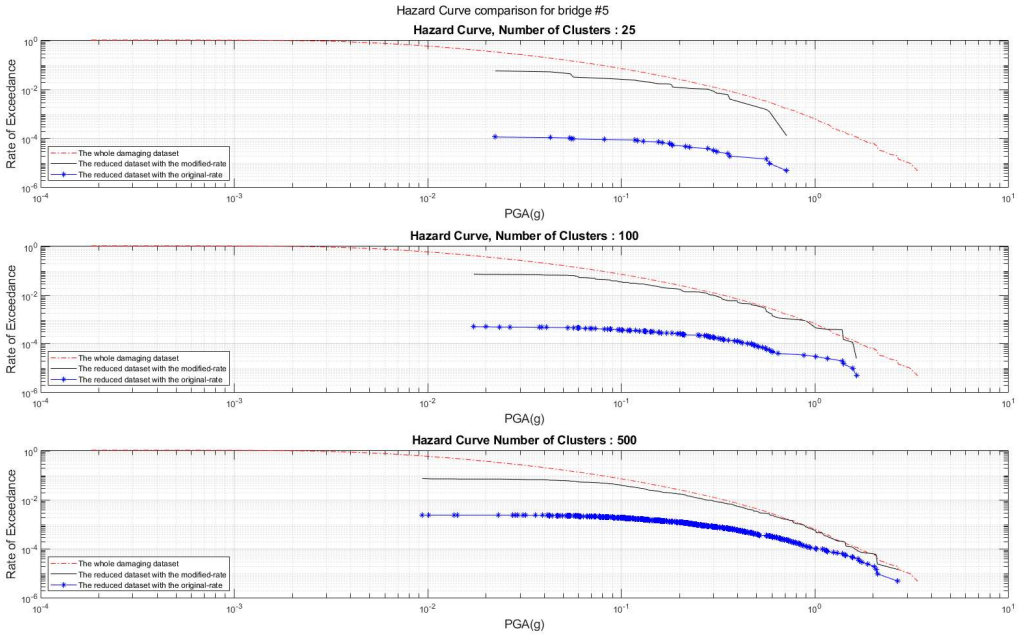


Figure 4-127 Hazard Curve Comparison for Bridge number 5 (for Case DWCL and Unweighted Risk-Hazard), (unit of y axis is 1/year).

The hazard curves above, show a slight improvement in the resulted hazard curve with reduced datasets in this approach in comparison with the rescued datasets with 2 previous approaches for DWCL.

4.2.5.4 Hazard-Consistent approach with clustering number like 3 previous approaches with DWCL

In this approach, the only features to be considered in calculation of distance metric is hazard in this study it is included by using 5 PGA values in bridge locations. Again, only damaging events are considered to be clustered, and it means that the accurate hazard curve is not expected as in this damaging event set many events that causes lower PGAs in the networks has been eliminated and consequently these values of PGAs will not be included in the resulted hazard curves.

Figure 4-128 shows the comparison of the resulted loss exceedance curve with the reduced dataset in this approach in 3 different clustering number of 25, 100, and 500 that are identical to cluster numbers in the other 3 approaches applied for DWCL. Discrepancies can be observed in fewer number of clusters like 25 and 100 with the reference risk curve (red line in the plot). However, by increasing the number of clusters, like 500 here, the satisfactory loss exceedance curves that are compatible with the reference curve can be obtained.

Figure 4-129 to Figure 4-133 show the resulted hazard curve using the reduced dataset in this Hazard-Consistent approach in 3 different sizes of dataset identical to the dataset sizes in previous 3 approaches for DWCL which are 25, 100, and 500. As mentioned before, using only damaging events cause causes inaccurate hazard curves even for larger number of clusters like 500. Therefore, this approach is not a purely Hazard-Consistent one as it includes the effect of damage in the eliminations, however in comparison with the other 3 approaches of this research, this approach is called Hazard-Consistent.

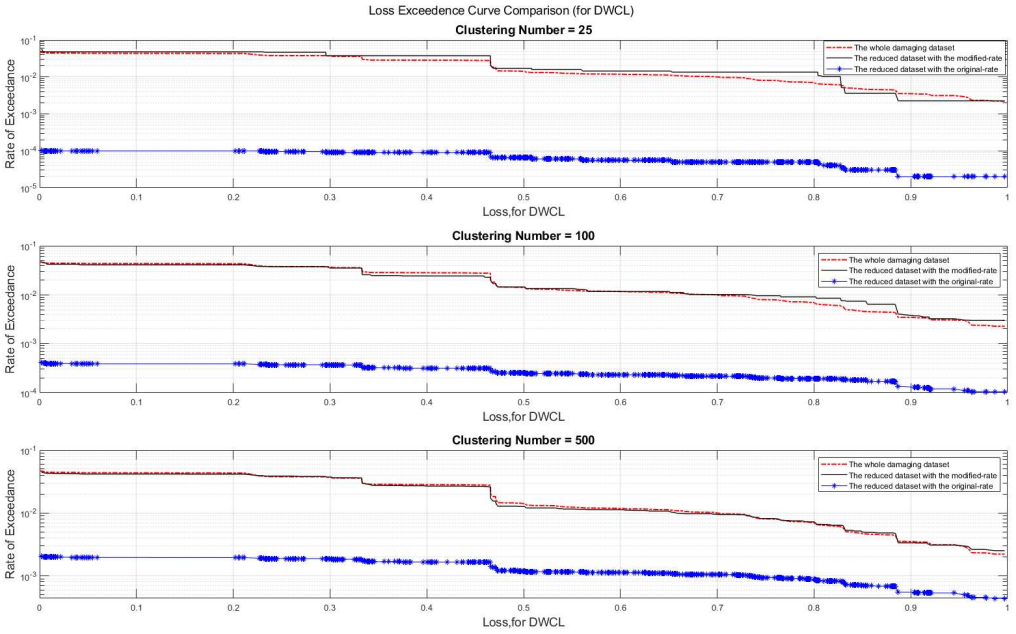


Figure 4-128 Comparison of LECs for reduced dataset with 3 different clustering number (for Case DWCL and Hazard-Consistent), (unit of y axis is 1/year).

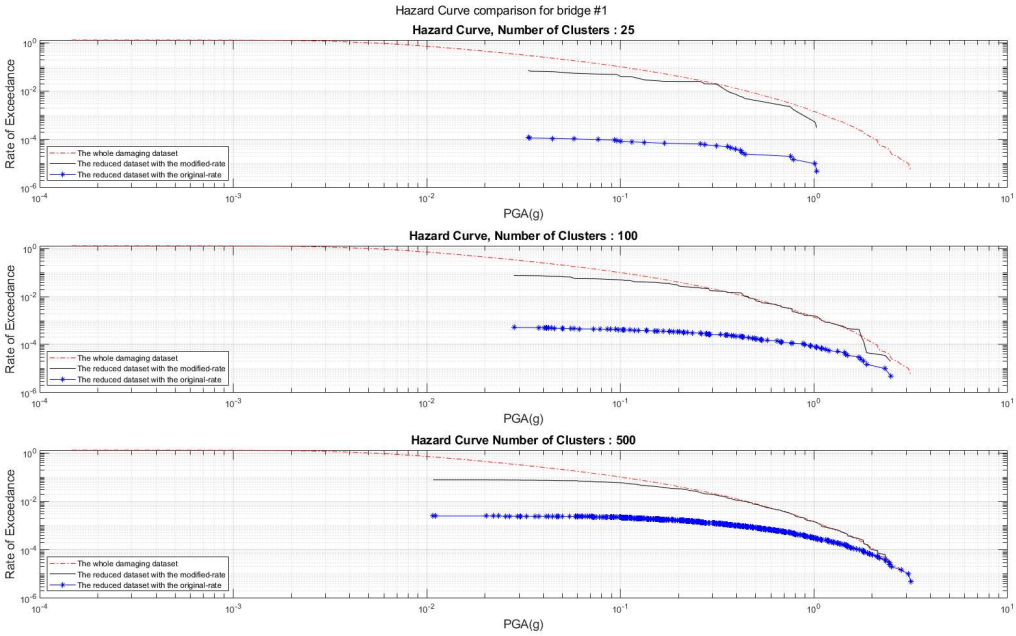


Figure 4-129 Hazard Curve Comparison for Bridge number 1 (to be compared with Case DWCL and Hazard-Consistent), (unit of y axis is 1/year).

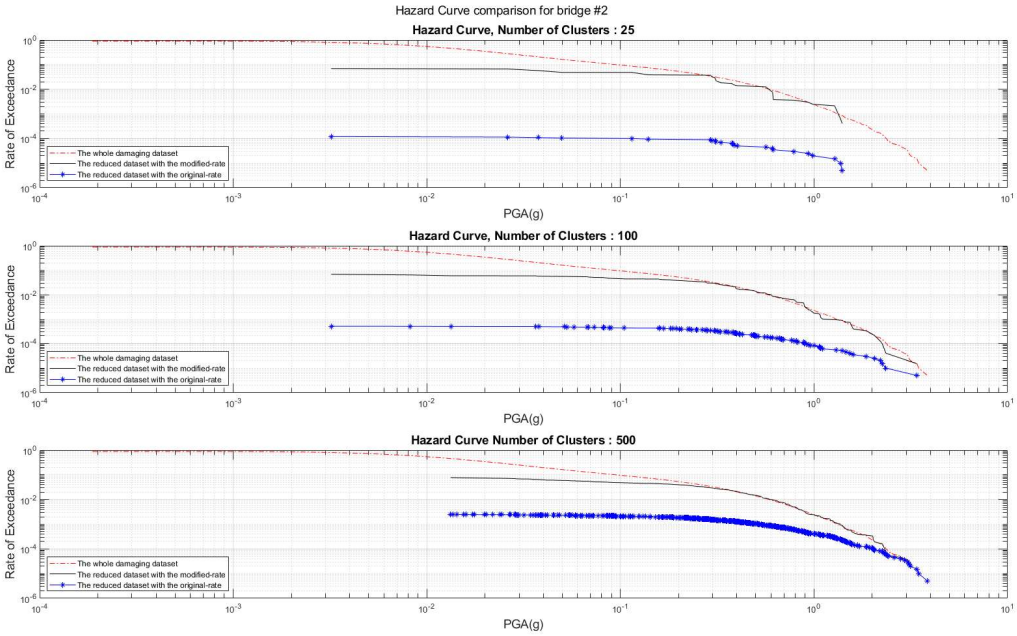


Figure 4-130 Hazard Curve Comparison for Bridge number 2 (to be compared with Case DWCL and Hazard-Consistent), (unit of y axis is 1/year).

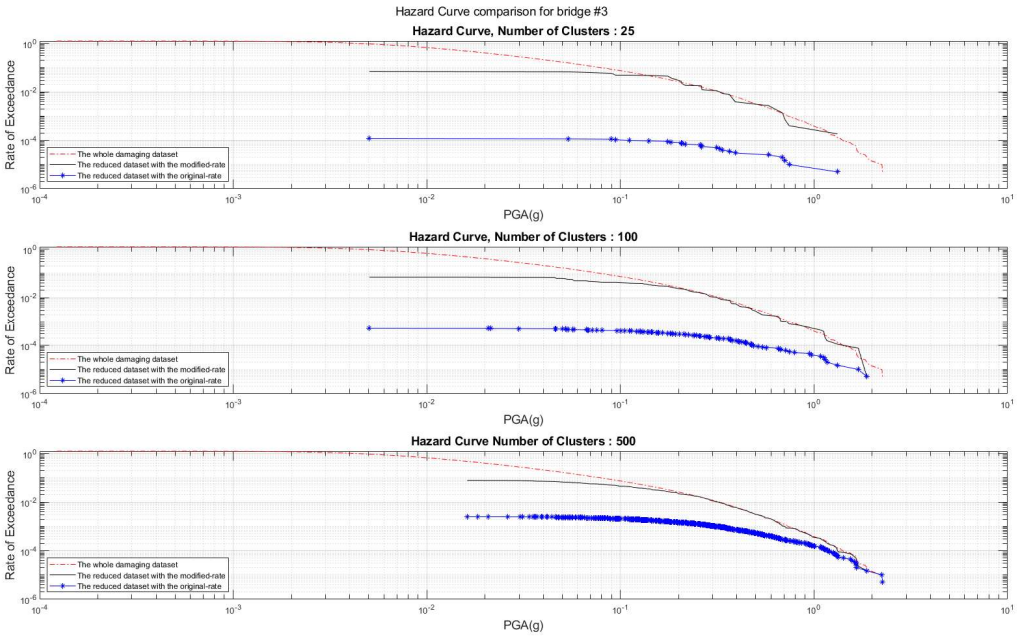


Figure 4-131 Hazard Curve Comparison for Bridge number 3 (to be compared with Case DWCL and Hazard-Consistent), (unit of y axis is 1/year).

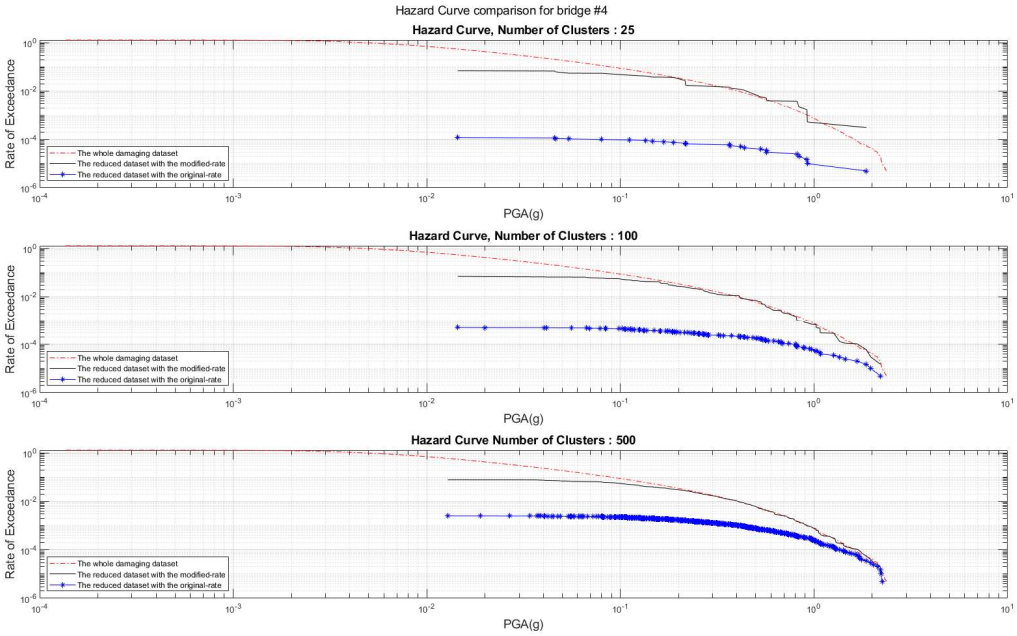


Figure 4-132 Hazard Curve Comparison for Bridge number 4 (to be compared with Case DWCL and Hazard-Consistent), (unit of y axis is 1/year).

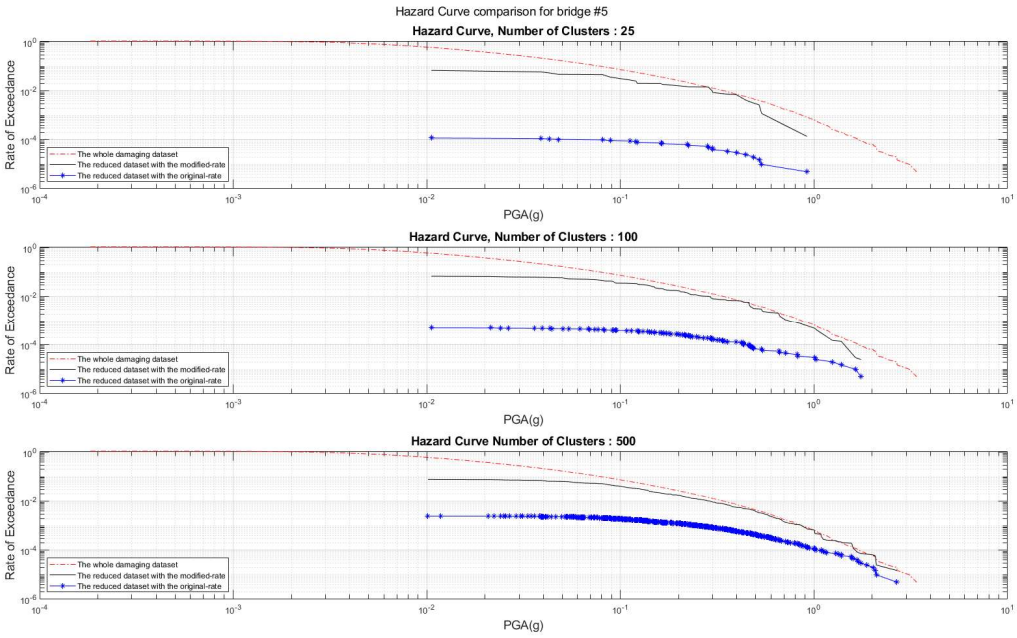


Figure 4-133 Hazard Curve Comparison for Bridge number 5 (to be compared with Case DWCL and Hazard-Consistent), (unit of y axis is 1/year).

4.2.5.5 Discussion on the results of DWCL using K-means clustering

The effects of using 4 different approaches on clustering with K-mean algorithm has been studied for DWCL as well and again like other applied loss metrics in this study, a similar pattern of decreasing

accuracy of resulted risk curves by decreasing the dominancy of the risk in clustering has been observed. Moreover, increasing the effect of the hazard in clustering improved the resulted hazard curves. However, to get both better risk and hazard curves, larger sizes of reduced datasets are required.

4.3 Using Density-Based Clustering Algorithm

As mentioned before in chapter of modeling and methodology, density-based clustering can spot outliers and do not force all the events to be grouped, and in contrast to K-means clustering that the shape of the clusters all where spherical here produced clusters can have arbitrary shapes. To study the effect of these features of density-based clustering in the reduction method, density-based clustering was applied for the Risk-Consistent approach, and it is tested for two different loss metrics which are the total length of damaged bridges (LDB) and the Distance-Based Weighted Connectivity Loss (DWCL). In following, the resulted reduced dataset by this clustering method is evaluated.

4.3.1 Density-based clustering in Risk-Consistent approach for LDB loss metric

In this section the density-based clustering algorithm is applied in Risk-Consistent approach and for the loss metric of this study that is related to the direct damage to the component of the network, which is total length of damaged bridges, as mentioned before, this metric can be converted to the economic loss.

Figure 4-134 shows the comparison of loss exceedance curve for the resulted reduced data set in 3 different sizes using the density-based clustering algorithm. The resulted curves have a good compatibility with the less frequent loss values unlike the more frequent loss values. In comparison of the resulted risk curve for the k-means clustering in Risk-Consistent approach for LDB, Figure 4-5, these curves show better accuracy for less frequent loss values even with smaller size of reduced datasets.

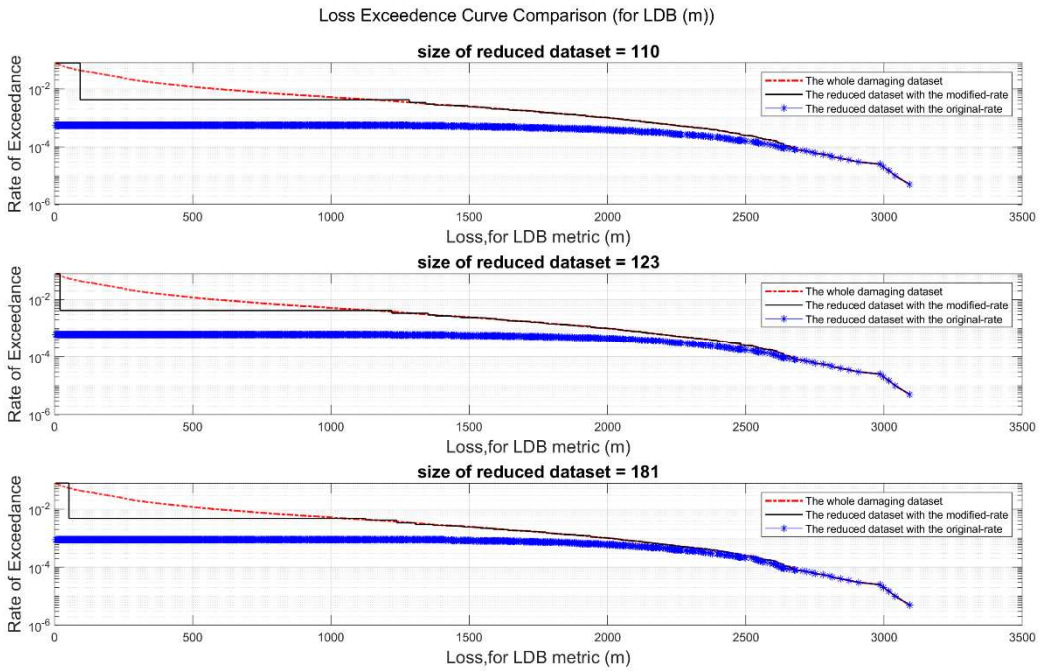


Figure 4-134 Comparison of LECs for different sizes of reduced dataset, using Density-Based clustering (for Case LDB and Risk-Consistent), (unit of y axis is 1/year).

Figure 4-135 to Figure 4-140 shows the resulted hazard curves with the reduced datasets using density-based clustering algorithm in different sizes of reduced dataset. Because it is in the Risk-Consistent approach, errors are seen in the resulted hazard curves as was expected. Like what have been observed in the results of K-means clustering, a large range of PGAs are missing in Risk-Consistent approach. But unlike what has been mostly observed in results of k-means clustering, increasing the size of reduced dataset does not necessarily improve the resulted hazard curves with the density-based clustering algorithm.

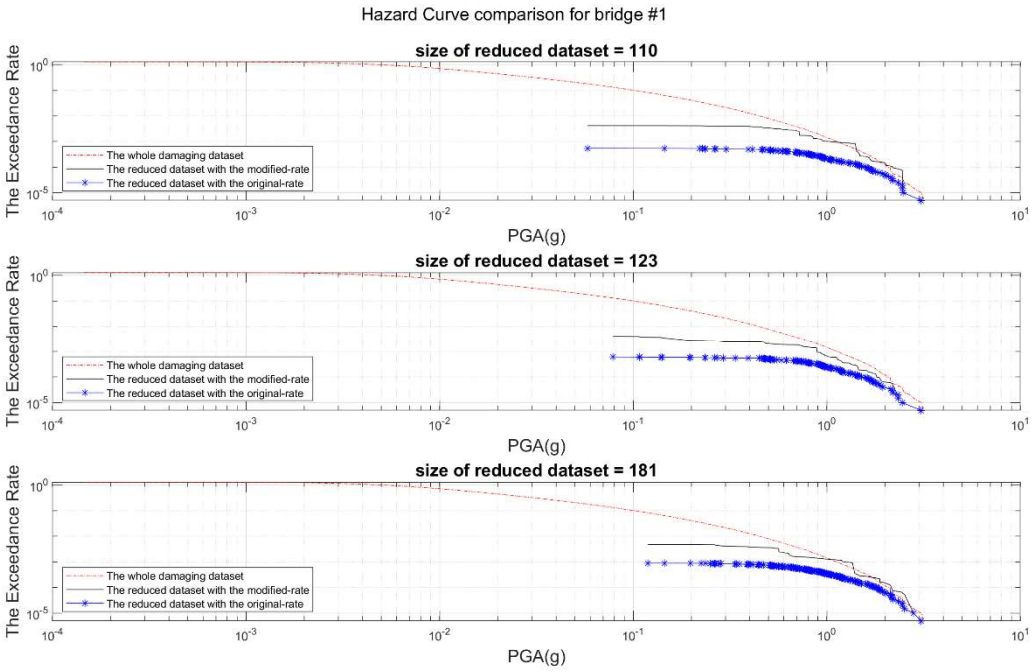


Figure 4-135 Hazard Curve Comparison for Bridge number 1 using Density-Based clustering (for Case LDB and Risk-Consistent), (unit of y axis is 1/year).

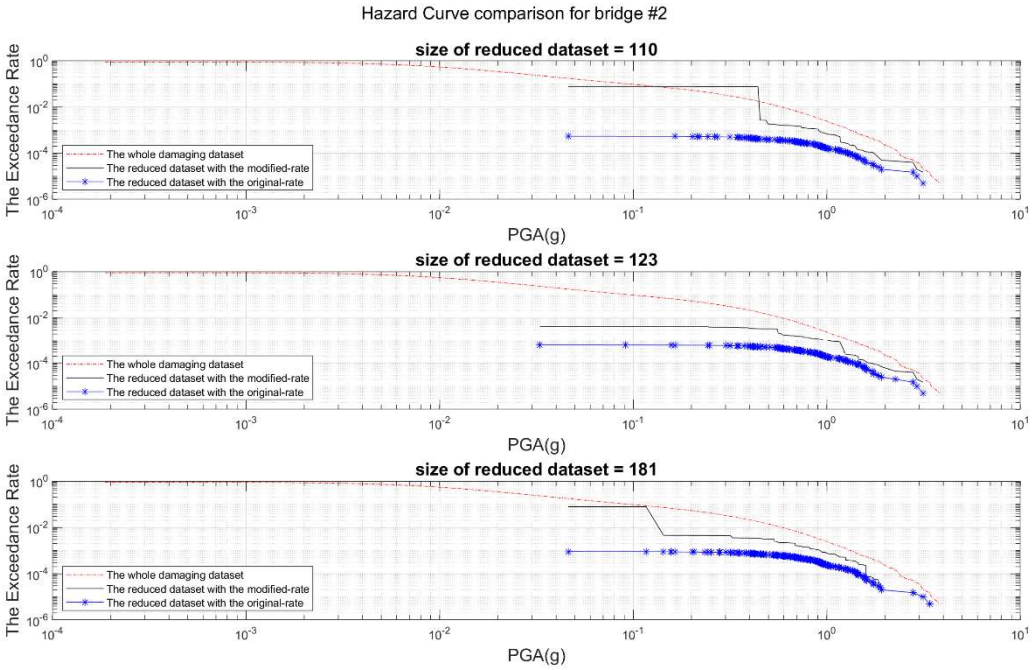


Figure 4-136 Hazard Curve Comparison for Bridge number 2 using Density-Based clustering (for Case LDB and Risk-Consistent), (unit of y axis is 1/year).

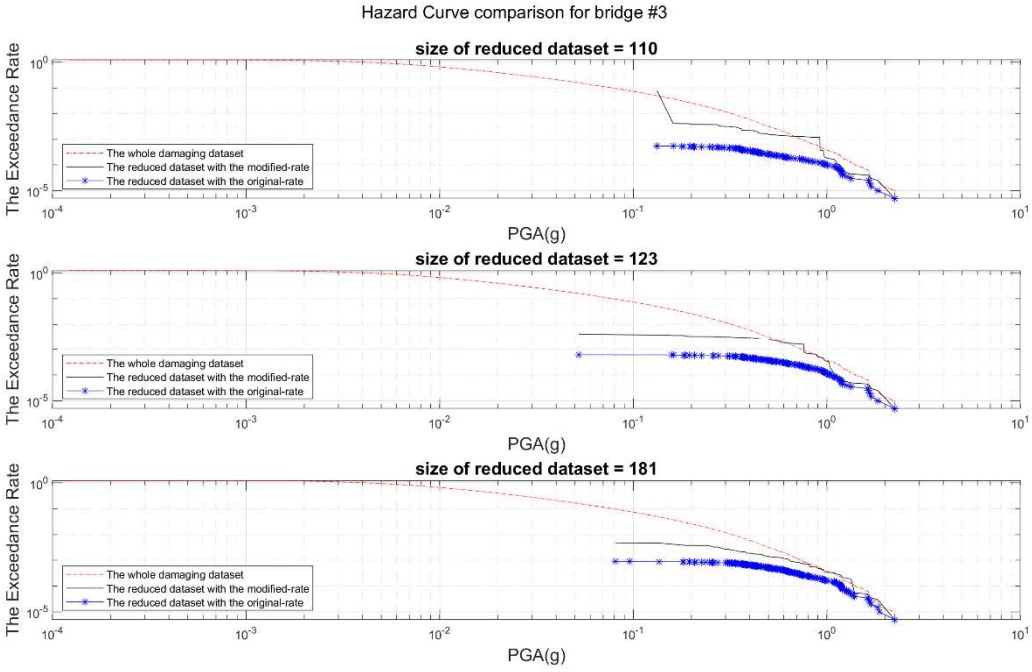


Figure 4-137 Hazard Curve Comparison for Bridge number 3 using Density-Based clustering (for Case LDB and Risk-Consistent), (unit of y axis is 1/year).

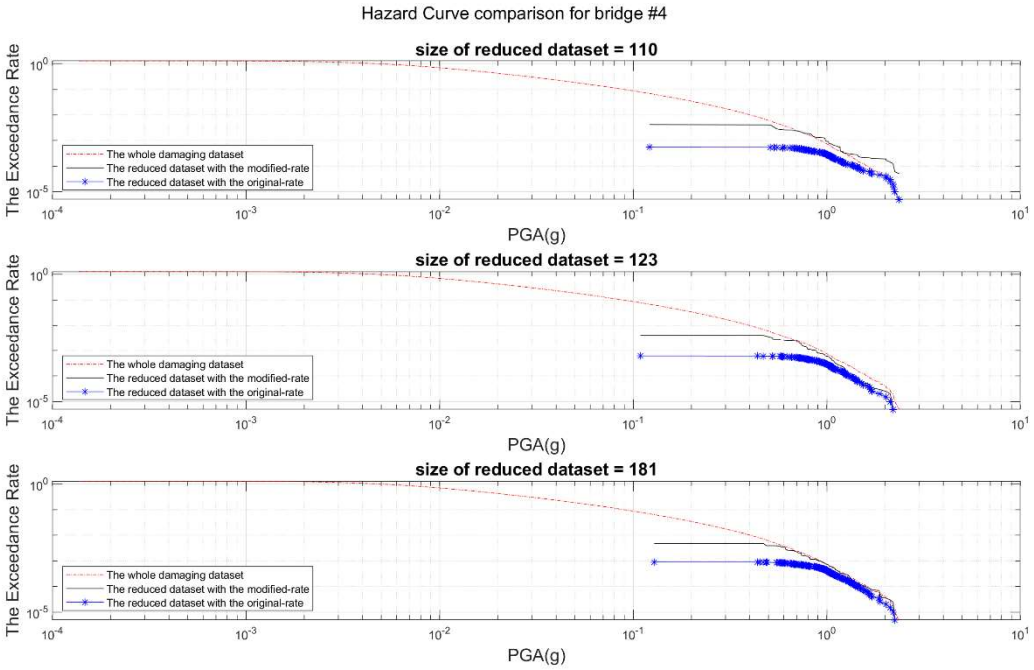


Figure 4-138 Hazard Curve Comparison for Bridge number 4 using Density-Based clustering (for Case LDB and Risk-Consistent), (unit of y axis is 1/year).

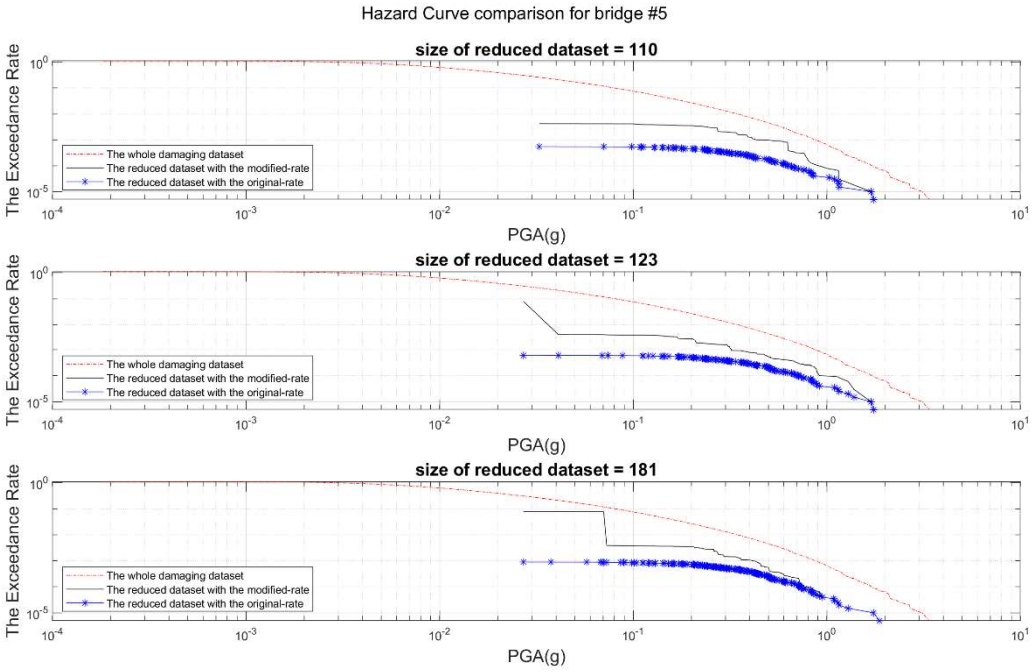


Figure 4-139 Hazard Curve Comparison for Bridge number 5 using Density-Based clustering (for Case LDB and Risk-Consistent), (unit of y axis is 1/year).

4.3.2 Density-based clustering in Risk-Consistent approach for DWCL loss metric

In this section risk and hazard curves of reduced dataset using density-based clustering algorithm in Risk-consistent approach for distance based weighted connectivity loss (DWCL) have been generated. Figure 4-140 shows the comparison of loss exceedance curves resulted by 3 different sizes of reduced datasets in Risk-Consistent approach. theses curves show for DWCL in 90 in comparison with the case of using K-means clustering for DWCL I Risk-Consistent in clustering number of 100 in Figure 4-110, a better compatibility is observed with the true risk curve with the whole damaging dataset.

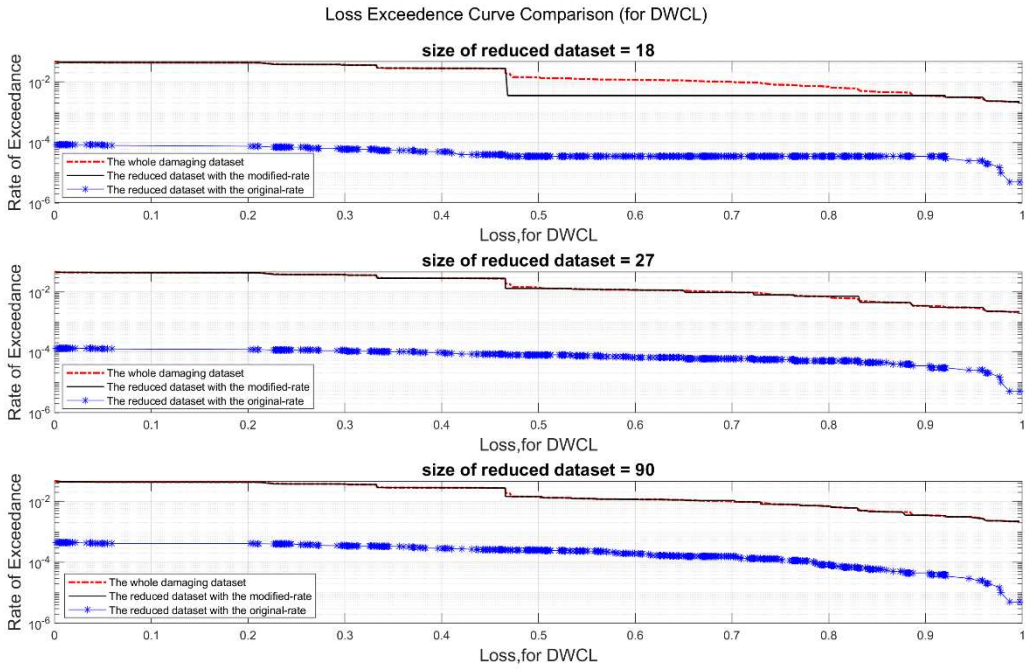


Figure 4-140 Comparison of LECs for different sizes of reduced dataset, using Density-Based clustering (for Case DWCL and Risk-Consistent), (unit of y axis is 1/year).

Figure 4-141 to Figure 4-145 show the resulted hazard curves with 3 different sized of reduced dataset in Risk-Consistent approach using density-based clustering. here again the Risk-Consistent approach does not result an accurate hazard result and like what was observed in case of LDB, in density-based clustering increasing the number of dataset does not necessarily improve the resulted hazard curve because of having partial clustered events and partial outliers, the trend in resulted curves is less predictable than K-means clustering.

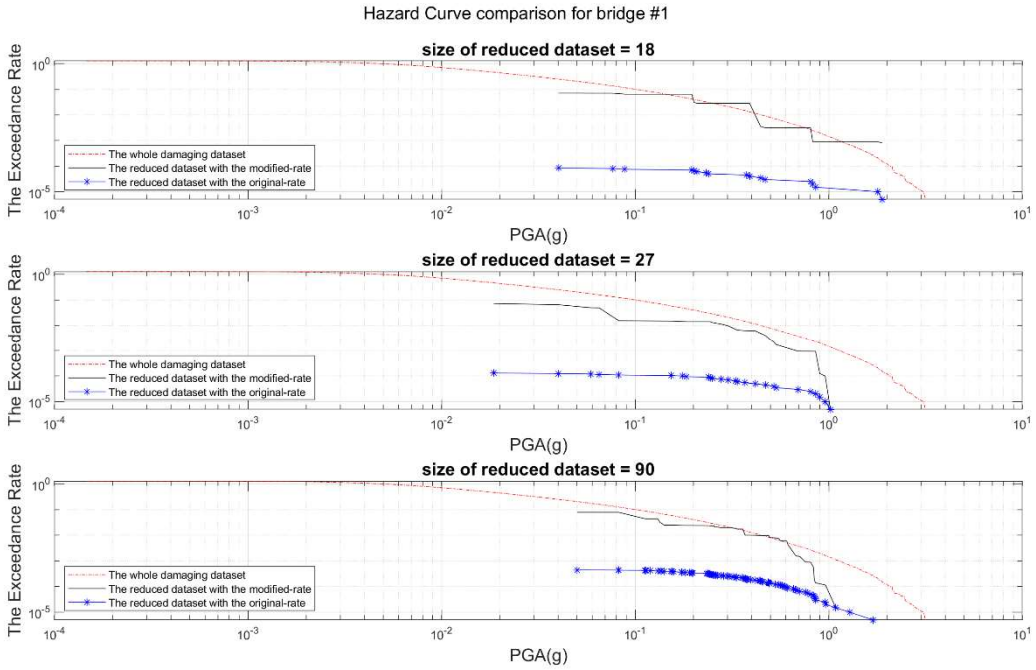


Figure 4-141 Hazard Curve Comparison for Bridge number 1 using Density-Based clustering (for Case DWCL and Risk-Consistent), (unit of y axis is 1/year).

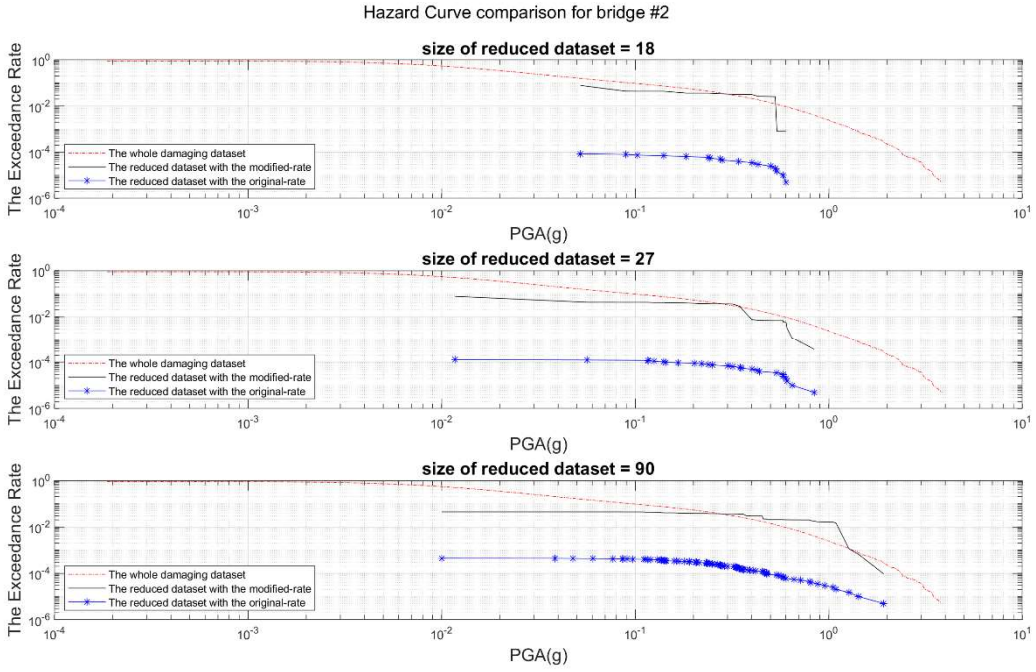


Figure 4-142 Hazard Curve Comparison for Bridge number 2 using Density-Based clustering (for Case DWCL and Risk-Consistent), (unit of y axis is 1/year).

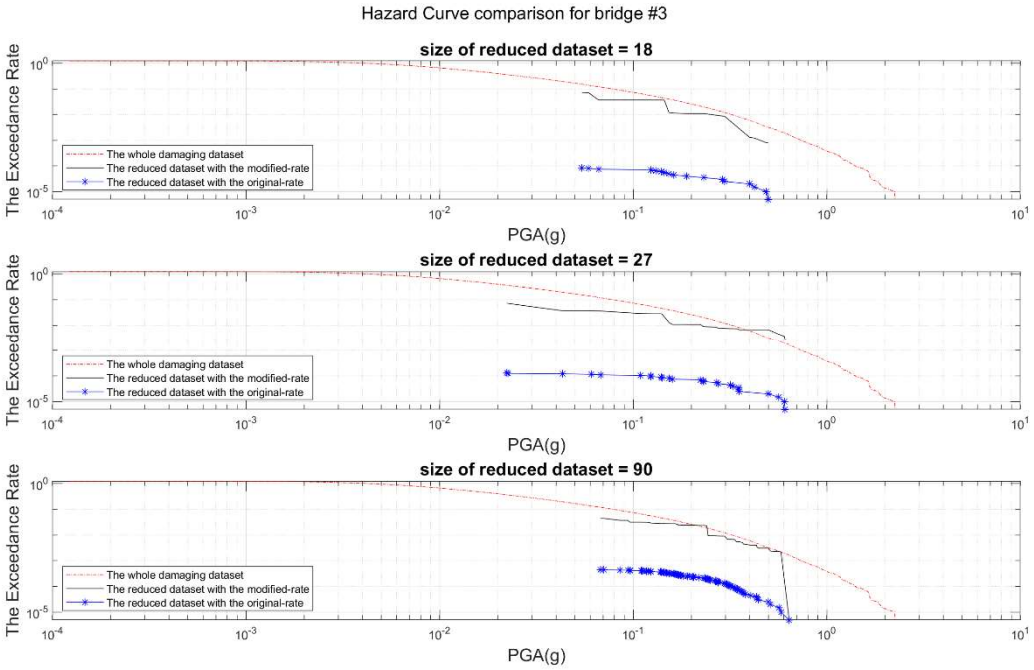


Figure 4-143 Hazard Curve Comparison for Bridge number 3 using Density-Based clustering (for Case DWCL and Risk-Consistent), (unit of y axis is 1/year).

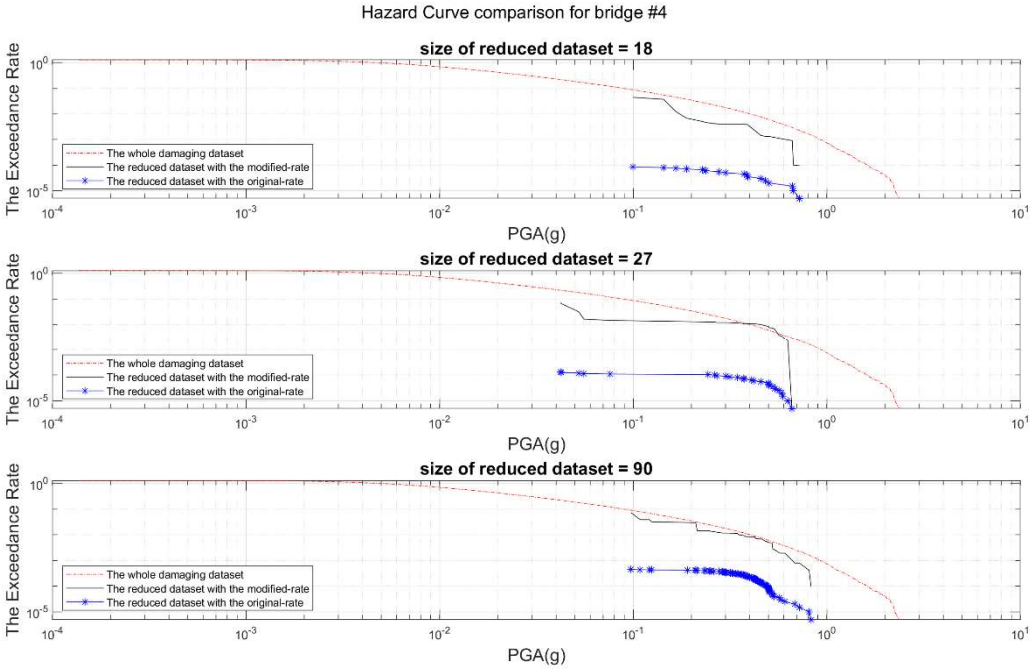


Figure 4-144 Hazard Curve Comparison for Bridge number 4 using Density-Based clustering (for Case DWCL and Risk-Consistent), (unit of y axis is 1/year).

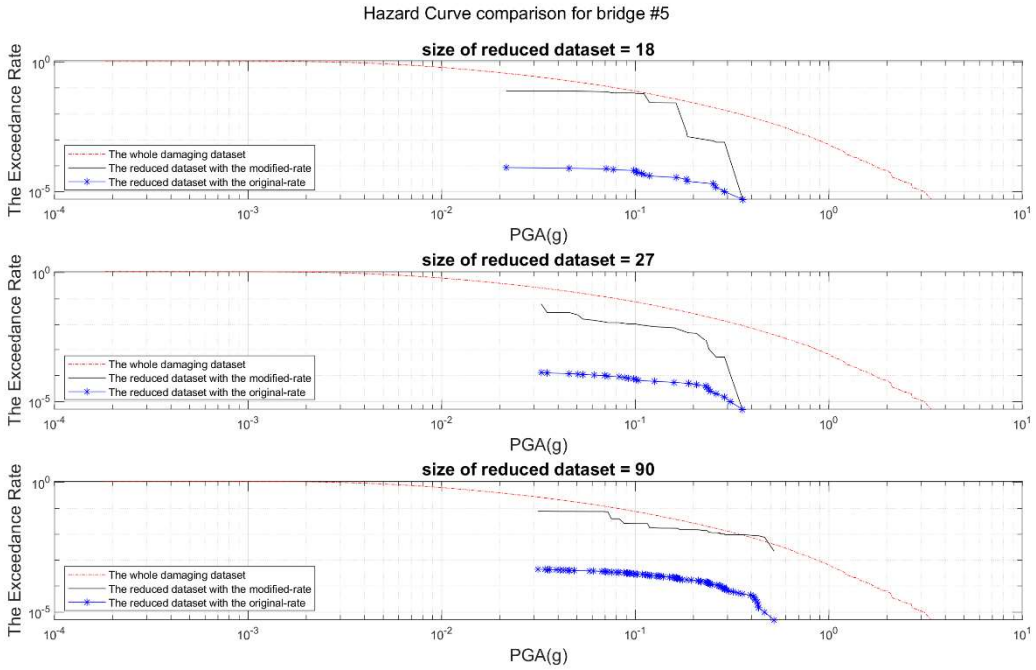


Figure 4-145 Hazard Curve Comparison for Bridge number 5 using Density-Based clustering (for Case DWCL and Risk-Consistent), (unit of y axis is 1/year).

4.4 Exploring the application of the Reduced dataset. Example: results of Risk-Consistent approach

In this section, the applicability of the resulted reduced dataset in Risk-Consistent approach is explored in testing improvement plans of the networks. Means that, if we want to improve our network for example retrofit scenarios and use this dataset in generating risk curves to the see the effect of retrofit scenario, is it compatible with the risk curve resulted by using the whole dataset or not. For this reason, here a retrofit scenario is conserved as the performance of 15 bridges (these bridges have been selected randomly just to test the procedure) is improved in a way that they do not experience any damage at all. The loss modeling implemented for the improved network once with the whole event set and another time just with 3 different sizes of the reduced dataset that were produced with Risk-Consistent approach of this research by using K-means clustering and the resulted risk curves were compared. It should be mentioned that the modification of the current rate of these events in the reduced dataset is based on the weakest performance of the network which is in the pre-retrofitted condition. this retrofitted case study was implemented for the reduced data set for two different loss metrics LDB and DWCL. Figure 4-146 and Figure 4-147 are the resulted risk curves for LDB and DWCL respectively. As it can be observed in figure 5-145 related to LDB, the predicted loss exceedance curve with the error based optimum number of the k-means clustering mostly conservative as it overestimates the risk. But the larger number of clustering like 500 shows a good accuracy in the resulted risk for the improved network. In Figure 4-147 and case of DWCL, the results are not mostly overestimating yet the errors are not significant.

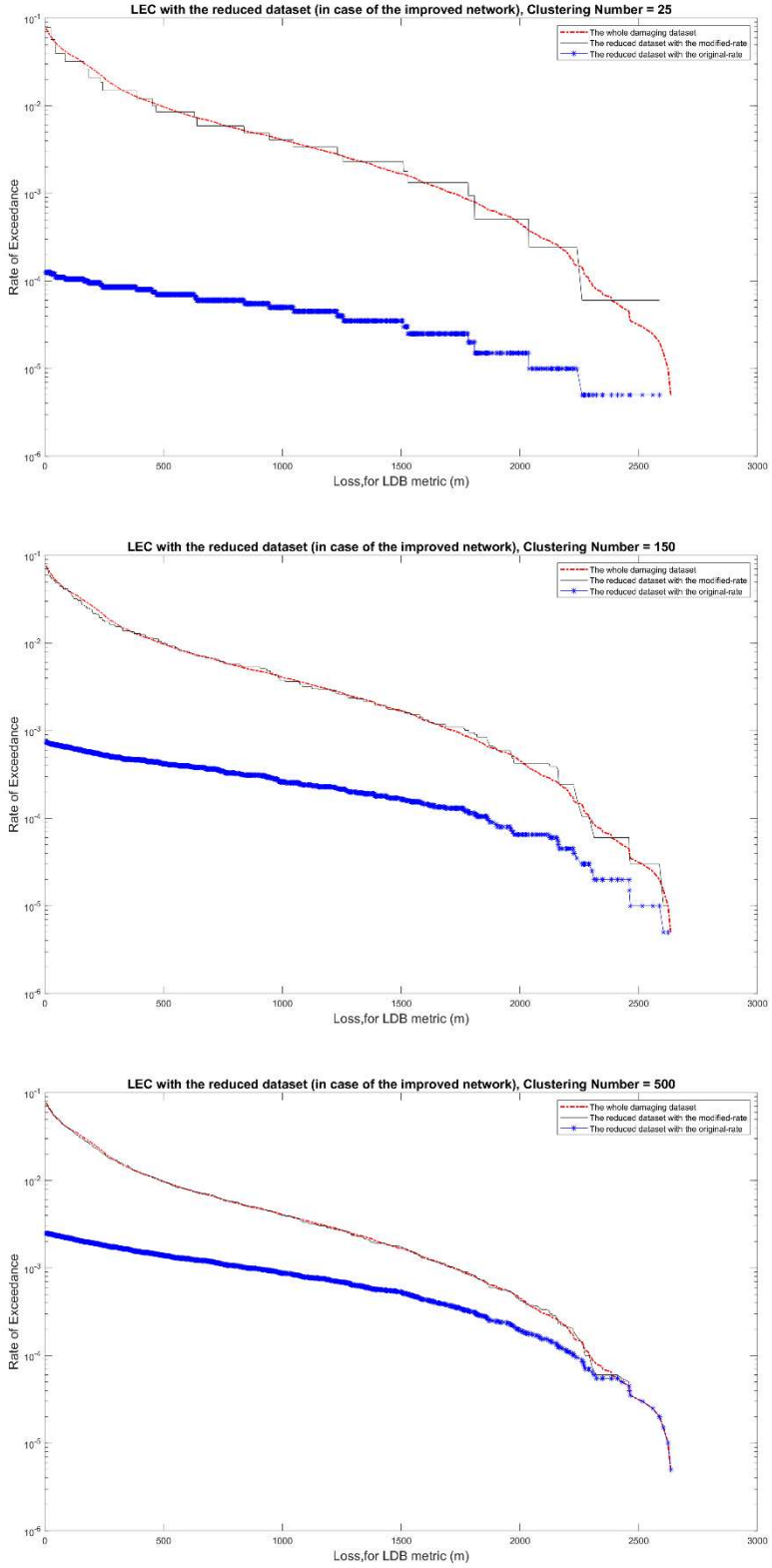


Figure 4-146 comparison of risk assessment of the improved network with LDB based risk consisted reduced dataset and the whole dataset, size of Reduced dataset: 25, 150 and 500 respectively, (unit of y axis is 1/year).

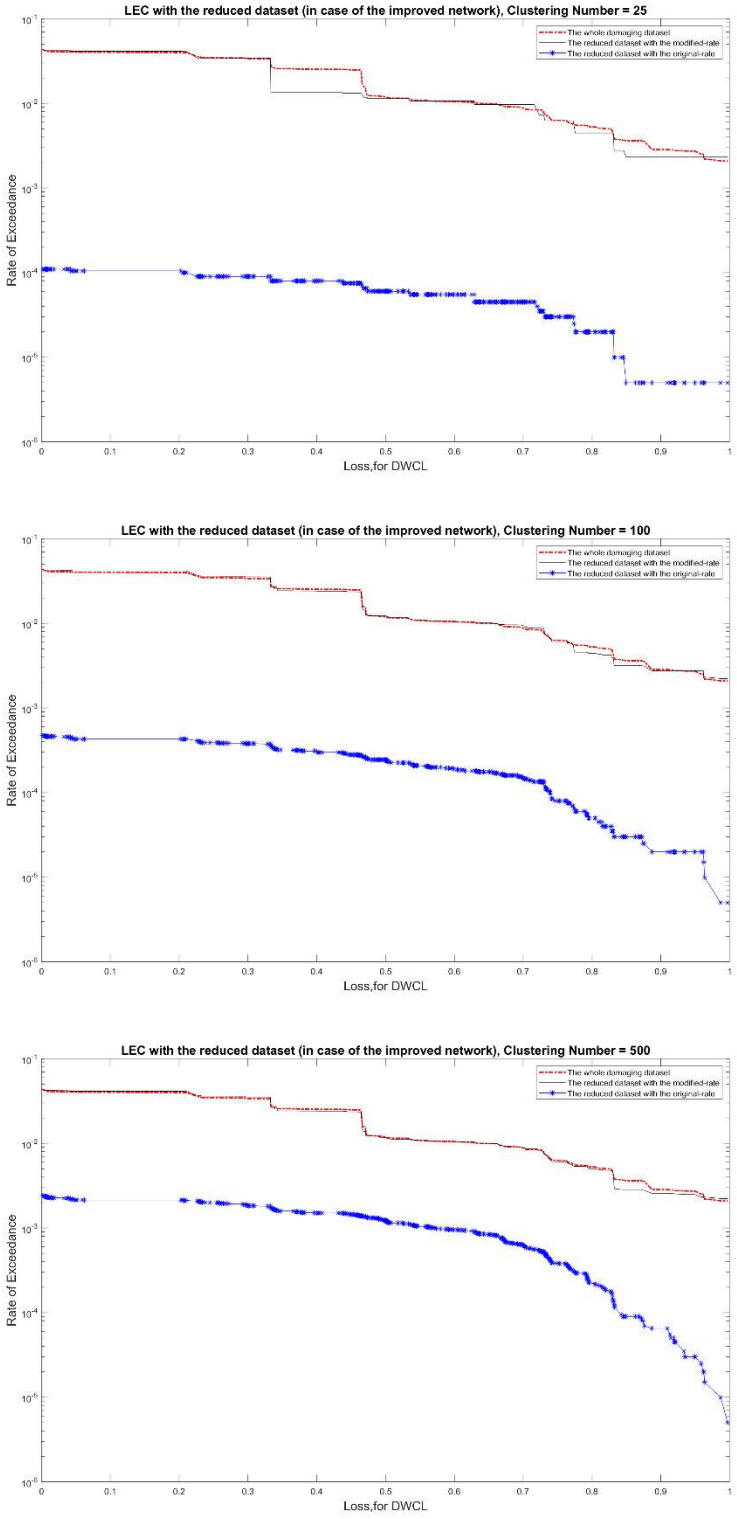


Figure 4-147 comparison of risk assessment of the improved network with DWCL based risk consisted reduced dataset and the whole dataset, size of Reduced dataset: 25, 100 and 500 respectively, (unit of y axis is 1/year).

4.5 Conclusions

The goal of this research was to find a methodology to reduce the required large number of events that are used in stochastic seismic risk assessment. Because this method of risk assessment is computationally intensive especially in the case of infrastructures due to the complexity of loss modeling of the networks this issue is exacerbated. For this reason, some clustering methods including simple spatial griding method, K-means clustering algorithm, and density-based clustering algorithm have been applied to classify similar events and group them. 4 different approaches were taken in defining the similarity of the events 1) Risk-Consistent approach: when the only resulted loss values differentiate events 2) Weighted Risk-Hazard approach: when the criterion in finding similar events is a weighted combination of loss value (50%) and intensity measures in 5 important sites (10% each). 3) Unweighted Risk-Hazard approach: similar combination with the weighted method yet without any weight allocations. 4) Hazard-consistent approach: intensity measures in 5 important sites are the criterion for evaluating the similarity between events. Moreover, to explore the potential of the reduced dataset in expediating the loss modeling and the risk estimation for improved networks, one hypothetical example of the improved network was evaluated, and the results were compared with the reference risk curve.

In the conclusion of this chapter, based on all the explained results, 3 aspects will be discussed here. First is the effect of considering different types of clustering methods and their pros and cons. The second is about the results of the applied different approaches of this research and their comparisons. And finally, the results of an exploration of the potential of the reduced dataset in expediating further risk assessments are mentioned.

Pros and cons of the applied clustering methods:

The first grouping method that has been applied in this thesis was the simple spatial griding method which uses a grid over the area to classify the events based on their locations and in the next step divides the range of loss values into small segments and groups the events that are in the same spatial grid and the same loss segment. Pros are this method is easy and fast and eventually produces a small subset of events that to some extent can imitate the spatial distribution of the whole package. The cons of this method are the size and number of clusters should be defined in advance, the shapes of the produced clusters are similar, and all the events are forced to be grouped in a cluster. This means that events that are in the same group might not be similar enough. Therefore, this method might result in a biased collection of the events.

The second applied clustering method in this study was the K-means algorithm, which is an iterative method to find the most similar events based on a predefined distance metric and a predefined number of clusters. Pros are this method is easy and straightforward and the methods like elbow point or the risk error method that are applied in this study are useful to decide about the appropriate number of clustering. the cons are shapes of the produced clusters are all spherical and all the events are finally joining a group. which means that this method like the previous one force all events to be clustered regardless of insufficient similarities with the events in the cluster.

The third clustering method that has been applied here was density-based clustering (DBSCAN). In this clustering method, predefined conditions like the limitation for the distance and a minimum number of members in a cluster are required to be determined in advance. Pros: It is known as a robust clustering method for finding outliers which means that, unlike the previous clustering method, this one does not

force the events to be clustered. Also, the resulted clusters can be produced in any shape. These features lead to a less biased selection of the events. Cons: finding a reasonable distance limit and a minimum number of members as the predefined parameters make it less straightforward in comparison with the K-means clustering.

Comparison of the results of different applied approaches

As mentioned above four different approaches have been applied in this research and the only clustering method that was used in studying all these approaches was the k-means clustering. The *Risk-Consistent approach* resulted in a reduced dataset that mostly generated risk curves that included a complete range of loss values, and even in smaller sizes of clustering an acceptable accuracy in the resulted risk curves was observed. However, the resulted hazard curves with the reduced datasets in the Risk-Consistent approach were not consistent with the reference hazard curve of the site. The second approach that was applied here was the *Weighted Risk-Hazard approach* a decreasing trend in the accuracy of the resulted risk curves and an improvement in the resulted hazard curves with the reduced dataset were observed (improvement in comparison with the results of risk consistent approach). However, by increasing the number of events in the reduced dataset, the accuracy of both risk curves and hazard curves seemed improvable. The *Unweighted Risk-Hazard approach* resulted in less accurate risk curves in comparison with the two previous approaches. However, resulted in slight improvements in the hazard curves. Lastly, the applied *Hazard-Consistent approach* which intentionally was not purely Hazard-Consistent, and the effect of the events on damaging bridges were included by using only damaging events in the clustering process. Therefore, accurate hazard results were not obtained. This was expected; because many events that produced lower values of intensity measures in the location of the 5 selected bridges in the network were eliminated in the first step of the reduction process which affected the accuracy of the resulted hazard curve. For the less frequent hazard, in cases of using larger sizes of the reduced datasets, satisfactory compatibilities with the reference hazard curve were observed. Also, less accurate risk curves resulted from the Hazard-Consistent approach that was improved with larger sizes of the reduced dataset.

Exploring about the potential of reduced dataset in expediating retrofit prioritizations

A hypothetical retrofitted scenario was considered as 15 bridges out of 87 bridges in the road network of this research, were randomly selected and improved in a way that they never will be damaged due to any of the events. The loss modeling has been performed once with the whole dataset and the next time with the reduced datasets, results of this research from the Risk-Consistent approach for two cases of loss metrics LDB and DWCL. The resulted risk estimations of the retrofitted road network with the reduced dataset and the whole datasets were compared. Satisfactory risk results with the reduced dataset were observed for the case of LDB, especially when larger sizes of reduced event sets were applied, like 500. For the case of DWCL, however, the results were not as good as LDB but could be improved by using larger sizes of the reduced dataset. In other words, the results of this exploration example showed that the reduced dataset with the Risk-Consistent approach generated a good risk estimation for the retrofitted scenario when the size of the applied reduced dataset is large enough. However, concluding about the applicability of the reduced event set (that has been generated with the Risk-Consistent approach) in expediting the retrofit prioritization, requires more study, and many examples should be tested, and it is out of the scope of this thesis.

Recommended Future research recommended research for future is a study about the applicability of the reduced dataset in expediting the retrofit prioritization. the reduction method can be in the Risk-Consistent or Weighted Risk-Hazard Consistent approach.

References

Chiou, B.-J., & ,Youngs. , R. (2008). An NGA model for the average horizontal component of peak ground motion and response spectra. *Earthquake spectra 24.1 (2008): 173-215., 24.1*, 173-215. doi:https://doi.org/10.1193/1.2894832

Aghabozorgi, S. (2020). Machine Learning With Python: A Practical Introduction. EDX. Retrieved from https://www.edx.org/course/machine-learning-with-python-a-practical-introduct

Akkar, S., & Çağnan, Z. (2010). A local ground-motion predictive model for Turkey, and its comparison with other regional and global ground-motion models. *Bulletin of the Seismological Society of America, 100.6*, 2978-2995. doi:https://doi.org/10.1785/0120090367

Akkar, S., Sandıkkaya, M., & Bommer, J. (2014). Empirical ground-motion models for point-and extended-source crustal earthquake scenarios in Europe and the Middle East. *Bulletin of earthquake engineering, 12.1*, 359-387. doi:https://doi.org/10.1007/s10518-013-9461-4

Ambraseys, N. (2009). *Earthquakes in the Mediterranean and Middle East: a multidisciplinary study of seismicity up to 1900.* (C. U. Press, Ed.)

Anagnos, T., & Kiremidjian, A. (1988). A review of earthquake occurrence models for seismic hazard analysis. *Probabilistic Engineering Mechanics , 3.1*, 3-11. doi:https://doi.org/10.1016/0266-8920(88)90002-1

Askan, A., Karimzadeh, S., Asten, M., Kilic, N., ŞİŞMAN, F. N., & Erkmen, C. (n.d.). Assessment of seismic hazard in the Erzincan (Turkey) region: construction of local. *Turkish Journal of Earth Sciences, 24.6*, 529-565. doi:10.3906/yer-1503-8

Atkinson, G., & Adams, J. (2013). Ground motion prediction equations for application to the 2015 Canadian national seismic hazard maps. *Canadian Journal of Civil Engineering, 40*, 988-998. doi:https://doi.org/10.1139/cjce-2012-0544

Avşar, Ö., yakut, a., & Caner, A. (2012). Development of analytical seismic fragility curves for ordinary highway bridges in Turkey. *The 15th World Conference on Earthquake Engineering (WCEE),*. Lisboa.

Baker, J. (2013). An introduction to probabilistic seismic hazard analysis. *White paper version 2.1*, 79.

Booth, E. (2007). The estimation of peak ground-motion parameters from spectral ordinates. *Journal of Earthquake Engineering , 11.1*, 13-32. doi: https://doi.org/10.1080/13632460601123156

Chang, S., Shinozuka, M., & Moore, J. (2000). Probabilistic earthquake scenarios: Extending risk. *Earthquake Spectra, 16.3*, 557-572. doi:https://doi.org/10.1193/1.1586127

Cornell, C. (1968). Engineering seismic risk analysis. *Bulletin of the seismological society of America, 58.5*, 1583-1606. doi:https://doi.org/10.1785/BSSA0580051583

Crowley, H., & Bommer, J. (2006). Modelling seismic hazard in earthquake loss models with spatially distributed exposure. *Bulletin of Earthquake Engineering, 4.3*, 249-273. doi:https://doi.org/10.1007/s10518-006-9009-y

Danciu, L., Şeşetyan, K., Demircioglu, M., Erdik, M., & Giardini, D. (2016). Danciu, L., et al. "Input files for OpenQuake used to compute the seismic hazard of the Middle East region within the Earthquake Hazard Assessment of Middle East (EMME) Project. doi:doi:10.12686/a3

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische mathematik, 1.1*, 269-271.

Douglas, J. (2010). *Estimation of strong ground motion: Aleatory variability and epistemic uncertainties.* Doctoral dissertation, Université Joseph-Fourier-Grenoble I. Retrieved from https://tel.archives-ouvertes.fr/tel-00545546/

Ebel, J., & Kafka, L. (1999). A Monte Carlo Approach to Seismic Hazard Analysis. *Bulletin of the Seismological Society of America, 89.4* , 854-866. doi:https://doi.org/10.1785/BSSA0890040854

Ester, M. e. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *kdd. , Vol. 96. No. 34.* Retrieved from https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf?source=post_page

Franchin, P., & Cavalieri, F. (2013). Seismic vulnerability analysis of a complex interconnected civil infrastructure. In *Handbook of seismic risk analysis and management of civil infrastructure systems.* (pp. 465-514e.). Woodhead Publishing. doi:https://doi.org/10.1533/9780857098986.4.465

Gan, G., Chaoqun, M., & Jianhong, W. (2020). *Data clustering: theory, algorithms, and applications.* Society for Industrial and Applied Mathematics.

GEM. (2022). *The OpenQuake-engine User Manual. Global Earthquake Model (GEM) OpenQuake Manual for Engine version 3.13.0.* doi:doi: 10.13117/GEM.OPENQUAKE.MAN.ENGINE.3.13.0

Grossi, P., Kunreuther, H., & Windeler, D. (2005). An introduction to catastrophe models and insurance. In *Catastrophe modeling: A new approach to managing risk* (pp. 23-42). Boston, MA: Springer.

Gutenberg, B., & Richter, C. (1944). Frequency of earthquakes in California. *Bulletin of the Seismological society of America, 34.4*, 185-188. doi:10.1785/BSSA0340040185

Han, Y., & Davidson. , R. (2012). Probabilistic seismic hazard analysis for spatially distributed infrastructure. *Earthquake Engineering & Structural Dynamics, 41.15* , 2141-2158. doi:https://doi.org/10.1002/eqe.2179

HAZUS. (2020). *Earthquake loss Estimation Methodology.* Technical Manual, FEMA, National Institute of Building for the Federal Emergency Managment Agency, Washington (DC). doi:https://www.fema.gov/sites/default/files/2020-10/fema_hazus_earthquake_technical_manual_4-2.pdf

Jayaram , N., & Baker, J. (2010). Efficient sampling and data reduction techniques for probabilistic seismic lifeline risk assessment. *Earthquake Engineering & Structural Dynamics, 39.10*, 1109-1131. doi: https://doi.org/10.1002/eqe.988

Jayaram, N., & Baker, J. (2009). Correlation model for spatially distributed ground-motion intensities. *Earthquake Engineering & Structural Dynamics, 38.15*, 1687-1708. doi:https://doi.org/10.1002/eqe.922

Kiremidjian, A., Moore, J., Fan, Y., Yazlali, O., Basoz, N., & Williams, M. (2007). Seismic Risk Assessment of Transportation Network Systems. *Journal of Earthquake Engineering, 11*, 371-382 . doi:https://doi.org/10.1080/13632460701285277

Kiremidjian, A., Stergiou , E., & Lee, R. (2007). Issues in seismic risk assessment of transportation. *Earthquake geotechnical engineering*, 461-480. doi:https://doi.org/10.1007/978-1-4020-5893-6_19

Kramer, S. L. (1996). *Geotechnical Earthquake Engineering.* Pearson Education India.

Kyriazis, P., Franchin, P., Khazai, B., & Wenzel, H. (2014). *SYNER-G: systemic seismic vulnerability and risk assessment of complex urban, utility, lifeline systems and critical facilities: methodology and applications.* (Vol. Vol. 31). Springer.

Lee, Y., & Graf, W. (2008). Development of earthquake scenarios for use in earthquake risk analysis for lifeline. *Proceedings of the 14th World Conference on Earthquake Engineering.* Bejing (CN).

Miller, M., & Baker, J. (2015). Ground-motion intensity and damage map selection for probabilistic infrastructure network risk assessment using optimization. *Earthquake Engineering & Structural Dynamics , 44.7* , 1139-1156. doi: https://doi.org/10.1002/eqe.2506

Musson, R. M. (2000). The use of Monte Carlo simulations for seismic hazard assessment in the UK.

Nielson, B. G. (2005). *Analytical fragility curves for highway bridges in moderate seismic zones.* Georgia Institute of Technology.

Pagani, M., Garcia-Pelaez, J., Gee, R., Johnson, K., Poggi, V., Styron, R., . . . Monelli, D. (2018). *Global Earthquake Model (GEM) Seismic Hazard Map (version 2018.1 - December 2018).* doi:DOI: 10.13117/GEM-GLOBAL-SEISMIC-HAZARD-MAP-2018.1

Pagani, M., Monelli, D., Weatherill, G., & Garcia, J. (2014-08). *The OpenQuake-engine Book: Hazard. Global Earthquake Model (GEM) Technical Report.* doi:doi: 10.13117/-GEM.OPENQUAKE.TR2014.08

Poljanšek, K., Bono, F., & Gutiérrez, E. (2012). Seismic risk assessment of interdependent critical infrastructure systems: The case of European gas and electricity networks. *Earthquake Engineering & Structural Dynamics, 41.1*, 61-79. doi: https://doi.org/10.1002/eqe.1118

Shinozuka, M., Feng, M., & Lee, J. (2000). Statistical analysis of fragility curves. *Journal of engineering mechanics, 126.12*, 1224-1231. doi:https://doi.org/10.1061/(ASCE)0733-9399(2000)126:12(1224)

Shinozuka, M., Murachi, Y., Dong, X., & Zhou, Y. (2003). Seismic Performance of Highway Transportation Networks. *Proceedings of China-US Workshop on protection of urban infrastructure and public buildings against earthquakes and man-made disasters. .*

Shiraki, N., Shinozuka, M., Moore, J., Chang, S., Kameda, H., & Tanaka, S. (2007). System risk curves: Probabilistic performance scenarios for highway networks subject to earthquake damage. *Journal of Infrastructure Systems, 13.1*, 43-54. doi:https://doi.org/10.1061/(ASCE)1076-0342(2007)13:1(43)

Shirazian, S., Ghayamghamian, M., & Nouri, G. (2011). Developing of fragility curve for two-span simply supported concrete bridge in near-fault area. *51*, pp. 571-575. World Acad Sci Eng Technol.

Sokolov, V., & Wenzel, F. (2013). Spatial correlation of ground motions in estimating seismic hazards to civil infrastructure. In *Handbook of seismic risk analysis and management of civil infrastructure systems* (pp. 57-78). Woodhead Publishing. doi:https://doi.org/10.1533/9780857098986.1.57

U.S. Geological Survey, 2. (2022, March 9). *Introduction to the National Seismic Hazard Maps.* Retrieved from https://www.usgs.gov/programs/earthquake-hazards/science/introduction-national-seismic-hazard-maps

Vaziri, P. (2009). Earthquake risk mitigation: hazard identification and resource allocation. *Ph.D. Dissertation, School of Civil and Environmental Engineering, Cornell University*. Retrieved from https://hdl.handle.net/1813/11393

Vaziri, P., Davidson, R., Apivatanagul, P., & Nozick, L. (2012). Identification of optimization-based probabilistic earthquake scenarios for regional loss estimation. *Journal of Earthquake Engineering, 16.2*, 296-315. doi:https://doi.org/10.1080/13632469.2011.597486

Wang, Z. (2011). Seismic hazard assessment: issues and alternatives. *Pure and Applied Geophysics, 168.1*, 11-25.

Weatherill, G. A. (2014). *OpenQuake Ground Motion Toolkit - User Guide. Global Earthquake Model (GEM). .* Technical Report.

Zhao, J., Zhang, J., Asano, A., Ohno, Y., Oouchi, T., Takahashi, T., . . . Fukushima, Y. (2006). Attenuation relations of strong ground motion in Japan using site classification based on predominant period. *Bulletin of the Seismological Society of America, 96.3*, 898-913. doi:https://doi.org/10.1785/0120050122