Development of empirical and analytical fragility functions using kernel smoothing methods

Hae Young Noh1,*,†, David Lallemant2 and Anne S. Kiremidjian2

1Department of Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA, 15213, U.S.A.
2John A. Blume Earthquake Engineering Center, Department of Civil and Environmental Engineering, Stanford University, Stanford, CA, 94305, U.S.A.

SUMMARY

Fragility functions that define the probabilistic relationship between structural damage and ground motion intensity are an integral part of performance-based earthquake engineering or seismic risk analysis. This paper introduces three approaches based on kernel smoothing methods for developing analytical and empirical fragility functions. A kernel assigns a weight to each data that is inversely related to the distance between the data value and the input of the fragility function of interest. The kernel smoothing methods are, therefore, non-parametric forms of data interpolation. These methods enable the implicit treatment of uncertainty in either or both of ground motion intensity and structural damage without making any assumption about the shape of the resulting fragility functions. They are particularly beneficial for sparse, noisy, or non-homogeneous data sets. For illustration purposes, two types of data are considered. The first is a set of numerically simulated responses for a four-story steel moment-resisting frame, and the second is a set of field observations collected after the 2010 Haiti earthquake. The results demonstrate that these methods can develop continuous representations of fragility functions without specifying their functional forms and treat sparse data sets more efficiently than conventional data binning and parametric curve fitting methods. Moreover, various uncertainty analyses are conducted to address the issues of over-fitting, bias, and confidence intervals. Copyright © 2014 John Wiley & Sons, Ltd.

Received 12 October 2013; Revised 29 September 2014; Accepted 2 October 2014

KEY WORDS: fragility function; performance-based earthquake engineering; non-parametric analysis; kernel smoothing; uncertainty in damage; uncertainty in ground motion

1. INTRODUCTION

Performance-based earthquake engineering (PBEE) has received increasing attention among structural engineering researchers and practitioners [1, 2] in order to predict the performance of a structure subjected to earthquakes in a probabilistic manner and to design accordingly to achieve selected performance objectives [1, 3, 4]. In the conventional PBEE framework, four parameters are used to compute annual loss rate of a structure due to earthquakes. The first parameter is intensity measure (IM), which quantifies the intensity of an earthquake ground motion. The second is engineering demand parameter, which represents the structural response to the earthquake. The third is damage measure (DM), which describes the discrete physical damage state (DS) of the structure. The last is decision variable, which relates to the actual loss, such as casualties, downtime, and monetary loss [5]. A fragility function represents the conditional probability of a structure being or exceeding a specific DS, such as slight, moderate, or severe damage given the intensity of an earthquake (IM).
This relationship between IM and DM involves many uncertainties due to inherent randomness in ground motion and structural properties as well as modeling uncertainties and measurement errors.

Recent developments in analytical models of structures, numerical simulations, and various instrumentations, and numerous efforts to conduct post-earthquake assessments and to collect field data have provided considerable amount of information on structural damages that can be correlated to observed or inferred ground motion intensity. Given this benefit, this paper addresses issues related to the development of analytical and empirical fragility functions. The most common forms in which data are collected following an earthquake are identified, and the techniques to manipulate these data to a format that is useful for fragility function development are discussed. Three new approaches for fitting fragility functions to the data are introduced that enable the treatment of uncertainty in IM and/or DM. These new techniques are based on Gaussian kernel smoothing (GKS) methods, which estimate a nonparametric functional relationship between two variables. The fitting is achieved using weighted average of nearby data, and this weighting function, referred to as a kernel, defines the contribution of each data for fitting based on the distance from the target point. Noh and Kiremidjian [6] and Noh et al. [7] introduced a kernel smoothing based framework for developing fragility functions using features developed in the field of structural health monitoring, and the method is further extended and applied in this paper for seismic risk analysis purposes. Note that although they discuss kernel-based fragility functions, the previous publications [6, 7] introduced them for structural health monitoring purposes. Thus, they focus on developing accurate fragility functions as a damage detection tool that can utilize sensor data from structures to provide probability distribution of DSs and then validating their damage detection performance. On the other hand, this paper addresses the issues specific to risk analysis. Therefore, it explores different data types that are often used in the analysis and then introduces three different kernel-based methods depending on available data types. Moreover, this paper further expands the previous development of kernel-based fragility functions and conduct detailed uncertainty analysis.

The conventional methods to estimate fragility functions from analytical and empirical data include a data binning and a parametric curve fitting using maximum likelihood estimation or Bayesian methods. The data binning methods divides IM values into several bins of a width, Δim, and then computes the probability distribution of damage at each bin [4]. One of the main issues with data binning for fitting a fragility curve is that the resulting curve is very sensitive to the subjective choice of bin size and bin boundaries. It also produces discontinuous fragility functions. On the other hand, the parametric curve fitting method often fits a cumulative distribution function (CDF) to the data, such as a lognormal distribution, using maximum likelihood and least-squares estimation [4, 8–11]. Fitting a CDF provides a smooth fragility function, but specifying a functional form can restrict the choice of relationship and may not reflect the true structure of the data. Instead, a kernel assigns a weight to each data sample that is inversely related to its distance from the target value of interest. Then, kernel smoothing methods estimate a fragility function as a weighted average of data. Kernel smoothing methods are, therefore, a special form of data interpolation and allow us to estimate continuous fragility functions without limiting them to a specific functional form. In addition, they can reduce the bias due to discretization (e.g., data binning) and provide more detailed information about the probabilistic relationship between DM and IM. Using the kernel smoothing methods is particularly beneficial when data is non-homogeneous or sparse. They provide statistically more stable results when the data are sparse because it utilizes all the observations to estimate the fragility functions unlike the data binning method that limits the data by the value of the IM. Moreover, the selection of kernel shape and bandwidth allows us to incorporate uncertainties in the data, such as measurement error and data credibility.

This paper is organized as follows. Section 2 describes different types of data and explains how to process them for kernel smoothing methods. Section 3 then introduces three GKS-based approaches for developing fragility functions from these data. In Section 4, the methods are applied to a set of data collected from analytical model of a four-story steel moment-resisting frame and a set of field observations collected after the 2010 Haiti earthquake. The results are presented and compared with other conventional methods, along with discussions about issues regarding applying the kernel smoothing method and how to overcome them. Finally, Section 5 provides the summary and conclusions.
2. DATA TYPES

Data for ground motion intensity and corresponding structural damage can vary in two main ways. In the first case (referred to hereafter as individual data), individual structures are analyzed or surveyed. For analytical fragility functions, simulations usually involve an analytical model of a single structure subjected to a set of ground motions with varying intensities. For each simulation, the ground motion intensity and the corresponding structural damage are collected. For empirical fragility functions, the DS for individual structure is recorded, and if a ground motion instrument is present in the vicinity of the structure, that intensity is also recorded. It is, however, unlikely that there will be a strong motion instrument in the vicinity of the structure, and thus the ground motion is often inferred from a ground motion map, such as the Shake Maps produced by the US Geological Survey (USGS) [12], or estimated subsequently using an appropriate ground motion attenuation law. Individual data may be used in the development of fragility functions for individual structures that may be in a region. These functions are independent of the total number of structures that are affected by the earthquake; however, they can also be used to aggregate the number of structures in a particular DS in a region. This dataset, denoted as $X_A$, can be represented as a collection of pairs of DM and corresponding ground motion intensity as

$$X_A = \{(dm_i, im_i); i = 1, 2, \cdots n\}$$

where $dm_i$ and $im_i$ correspond to the DM and IM for the $i^{th}$ structure in the database, and $n$ is the size of the dataset. The data set $X_A$ may include a specific structural type or represent an aggregate for all or several structural types together. In order to compute the probability of exceeding a DS for a given ground motion intensity, we need to bin the data, fit a parametric curve, or apply kernel smoothing methods. Then, if desired, a conventional model, such as a lognormal CDF, can be fitted.

A second form of data often collected for empirical data (referred to hereafter as grouped data) is a count obtained for the total number of buildings at an estimated ground shaking level and of the number of buildings in each DS providing a damage ratio for the particular DS. The interpretation of these two types of data is not the same and should not be used in the same way. Grouped data are typically used by actuaries to estimate the exposure of building within a portfolio. The damage ratio that is obtained from these data represents the fraction of the total number of buildings that is likely to be in a particular DS from a population of buildings. Emergency response personnel also may use these damage ratios as they provide rapid information on the probable fraction of structures in each DS of the totality of structures that have been affected by an earthquake. This type of dataset, denoted as $X_B$, can be represented as a set given by

$$X_B = \{(n_{ij}, Ni, im_i); i = 1, 2, \cdots n, j = 0, 1, \cdots, n_{ds}\}$$

where $n_{ij}$ represents the number of buildings in DS $j$, surveyed from area $i$. The dataset $X_B$ contains $n_{ds} + 1$ DSs including the undamaged state (note $j = 0$ for undamaged state), and $n$ is the total number of surveyed areas. It should be noted that the total number of buildings in area $i$, denoted as $N_i$, in the data set is

$$N_i = \sum_{j=0}^{n_{ds}} n_{ij}$$

and the total number of all the buildings surveyed is

$$N = \sum_{i=1}^{n} N_i$$

In this formulation, note that buildings with no damage is included in the total building count $N$. Because we can directly compute the probability of exceeding each DS $j$ from the data as
\[ \sum_{k=j}^{n_{ij}} n_{ik}/N_i, \text{ a scatter plot of IM and corresponding probability of exceeding a DS can be obtained from such type of data. We can then apply a kernel smoothing method or fit a conventional model, such as a lognormal CDF, to this scatter plot. Applying kernel smoothing methods to these conditional probabilities is equivalent to converting these data to individual data type and then applying the methods. Grouped data can be converted to the individual data type by generating } n_{ij} \text{ pairs of } \{ im_i, dm_j \} \text{ for each DS } j \text{ and surveyed area } i. \text{ In this way, kernel smoothing methods can consider uncertainties on } im_i \text{ and } n_{ij} \text{ for developing fragility functions.}

We used both types of data to develop fragility functions to validate our kernel smoothing based methods. In the subsequent sections, we address some of the issues related to each type of data and the methods used to develop the respective functions.

3. KERNEL SMOOTHING METHODS

A kernel is a weighting function that assigns a weight to each noisy observation, and the weight is inversely related to the distance between the observed value and the value that we want to estimate [13]. In other words, if we are interested in computing the fragility function for DM, \( dm^* \), at ground motion intensity \( im^* \), then the weight for a data point \( (dm_i, im_i) \) is a function of the distance between \( im^* \) and \( im_i \) for one-dimensional kernel smoothing and that between \( (dm^*, im^*) \) and \( (dm_i, im_i) \) for two dimensional. General description of kernel smoothing is first given for one-dimensional case, and the two-dimensional expansion is described in Section 3.3. A kernel assigns smaller weights to data points with larger distances than others with shorter distances. This implies that the data points farther away from the point of interest contribute less to the computation of the fragility function than those that are closer. In this paper, we use the Gaussian probability density function (PDF) as a kernel, without loss of generality. The method using the Gaussian PDF is referred to as the GKS. The variance of the Gaussian PDF represents the uncertainty of the data, and, if DM is continuous, by using the multi-variate Gaussian PDF, we can consider the uncertainties of both DM and IM and their correlation. Note that we may apply other forms of kernels in a similar way.

This kernel smoothing approach is particularly beneficial when the data are sparse because it efficiently utilizes the entire data set to compute the fragility function at every IM value, instead of confining the data into different bins. In addition, it avoids the discretization error unlike the data binning method. For example, a small measurement error may assign a sample to a different bin in data binning method, which may result in a large difference in regression. Moreover, all the data within a bin is lumped at a central IM value, no matter what their actual individual IM values are. On the other hand, the kernel smoothing method uses a continuous kernel to assign a weight to each sample, thus the effect of a small measurement error on the regression result is mitigated and their actual IM values contribute to computation of their weights. Another benefit of kernel smoothing is that we do not need to assume a particular form of fragility functions unlike parametric curve fitting methods. In other words, kernel smoothing provides a very flexible model that adjusts its profile according to data.

Similarly, a kernel can represent the uncertainty in the observed value. We can estimate a PDF of a random variable \( X \), denoted as \( f_X(x) \), using the kernel as

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K \left( \frac{x - x_i}{h} \right)
\]

where \( x_i \)s are \( n \) realizations of the random variable \( X \), \( K \) is a kernel, and \( h \) is a smoothing parameter or the bandwidth of the kernel \( K \). For Gaussian kernel, \( K(x) \) is a standard Gaussian probability density function given as

\[
f_X(x) \approx \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K \left( \frac{x - x_i}{h} \right)
\]
and the bandwidth corresponds to the standard deviation.

Figure 1(a) shows an illustrative example of the Gaussian kernel density estimation based on the five observations, \( x_1, x_2, \ldots, x_5 \). The dots indicate observed values. The dotted lines correspond to a kernel for each observation representing the uncertainty of the observation and/or the contribution of the observation for computing the PDF of a random variable \( X \). Therefore, the bandwidth of the kernel (i.e., the standard deviation of the Gaussian kernel) can be selected on the basis of the measurement error, if available, or engineering judgment. If the observations are independent identically distributed samples of a Gaussian distribution, we can use Silverman’s optimum bandwidth \( h \) for the Gaussian kernel [14]. It is given as

\[
h = 1.06 \hat{\sigma} n^{\frac{1}{5}}
\]

where \( \hat{\sigma} \) is the sample standard deviation, and \( n \) is the number of samples. Finally, the blue dashed line is the estimated PDF using the kernels. Note that the resulting PDF does not have an explicit parametric form. Instead, it is represented as a sum of kernels the number of which equals the size of the observations. For this reason, the kernel methods are often used as an intermediate step to obtain a parametric model or for model selection.

It should be noted that the kernel can be interpreted in two ways: from the perspective of function estimation and of data uncertainty representation. First, when we want to approximate a function output \( y^* \) (e.g., the probability of exceeding a DS for a fragility function or any other generic function) for the input value of \( x^* \), which we may or may not have observed, one way to estimate \( y^* \) is to compute the weighted average of the nearby observations. Here, the kernel smoothing is a systematic way of assigning the weights on the basis of how near the observations are to \( x^* \). The shape and the bandwidth, \( h \), of the kernel determine how fast the weight decays with the distance between the observation and \( x^* \). Therefore, in the case of the Gaussian kernel, we can visualize the kernel as the Gaussian PDF function centered at \( x^* \) with standard deviation \( h \) as shown in Figure 1(b). The weight for each observation \( x_i \) is determined by this kernel evaluated at \( x_i \), and the standard deviation depends on the specific application of interest.

The second approach interprets the kernel as the measure of uncertainty for the observation, equivalent to the degree of belief for or the assumed distribution of the true value. Therefore, the kernel can be represented as a Gaussian PDF centered at each observation with a specific standard deviation \( h \) as shown in Figure 1(a). This standard deviation can be determined from the uncertainty of the data, and it does not need to be identical for each observation. These two approaches are equivalent because the kernel is a function of the absolute difference between \( x^* \) and \( x_i \).

One additional consideration when using the Gaussian kernel is the domain of the probability distribution function. If the random variable \( X \) is bounded, for example, \( 0 \leq X \leq 1 \), its PDF is also bounded on the interval zero to one while the standard Gaussian kernel estimator is unbounded. This results in non-zero probability of \( X \) beyond the possible interval. It also means that the integral of the probability distribution over the interval zero to one is less than one. A simple method to
solve this problem is to transform the variable. One such transformation is the inverse normal cumulative distribution function. Given a monotonically transformed variable, we can estimate its distribution $g_T(T(x))$ through the kernel estimator. Using the relationship between the cumulative distributions of $X$ and $T(X)$ and the chain rule, the original probability distribution $f_X(x)$ of the untransformed variable is then obtained as

$$f_X(x) = g_T(T(x)) \left| \frac{dT(x)}{dx} \right|$$  (8)

There are three methods using the GKS approach for different types of DM and consideration of uncertainties: one dimensional Gaussian kernel for discrete DSs, one dimensional Gaussian kernel for continuous DMs, and two dimensional Gaussian kernel for continuous damage and IMs. Because the kernel smoothing is a non-parametric method, the resulting fragility function does not have a simple parametric description. Therefore, a conventional function can be fitted to the non-parametric fragility function using the method of moments, maximum likelihood, or Bayesian methods. The lognormal CDF is used in conventional fragility functions, but other functions, such as beta and normal CDF, can also be used depending on the characteristics of the data. Note that the GKS method estimates a function based on the pattern of the data, which allows a great flexibility in the resulting functional shape. However, this flexibility may prevent the method from incorporating necessary constrains, such as enforcing monotonically increasing fragility functions in our application. Ideally, the data should contain this characteristic, which in turn gets reflected in the resulting fragility functions through the GKS method. For noisy data, we can increase the kernel bandwidth to smooth out the noise and highlight the underlying structure of monotonic increase. In case the noise level is too high to remove, we can impose the non-decreasing constraint by fitting conventional CDFs to the GKS-based functions. The three GKS methods are summarized in Table I and are described in detail in the following subsections.

3.1. One-dimensional Gaussian kernel smoothing for discrete DSs, $DS_i$

Damage states are discrete variables most often defined as ‘no damage,’ ‘slight damage,’ ‘moderate damage,’ and ‘severe damage.’ Alternatively, each DS ($DS_0$) can be defined by a range of continuous DM. Thus, a set of threshold values of DM is specified for each DS $DS_i$ as follows:

Table I. Summary of three probabilistic mapping methods between intensity measure and damage measure (or damage state).

<table>
<thead>
<tr>
<th>Methods</th>
<th>Outcome</th>
<th>Advantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>1D GKS for discrete damage state</td>
<td>Fragility function $(\text{Prob}(DS \geq DS_i</td>
<td>IM = im))$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Considers uncertainty in IM measurement.</td>
</tr>
<tr>
<td>1D GKS for continuous damage measure</td>
<td>Conditional mean and standard deviation $(\mu_{DM</td>
<td>IM}, \sigma_{DM</td>
</tr>
<tr>
<td></td>
<td>Conditional probability of DM given IM $(\text{Prob}(DM = dmi</td>
<td>IM = im))$</td>
</tr>
<tr>
<td>2D GKS for continuous damage measure</td>
<td>Conditional probability of DM given IM $(\text{Prob}(DM = dmi</td>
<td>IM = im))$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Considers uncertainty in IM and DM measurement.</td>
</tr>
</tbody>
</table>

GKS, Gaussian kernel smoothing; DS, damage state; DM, damage measure; IM, intensity measure.
\[
\text{Damage State} = \begin{cases} 
DS_0 & \text{if } DM_0 \leq DM < DM_1 \\
DS_1 & \text{if } DM_1 \leq DM < DM_2 \\
\vdots \\
DS_n & \text{if } DM_n \leq DM < DM_{n+1} 
\end{cases}
\]

(9)

where \(DM_i\)s are monotonically increasing threshold values for increasing \(i\)'s, and \(n\) is the number of DS. Similarly, FEMA 356 uses four different DSs, namely, very light, light, moderate, and severe, which correspond to the performance levels of operational, immediate occupancy, life safety, and collapse prevention, respectively [15]. Using this definition of DSs and of the conditional probability, the fragility function, \(G_i(im) = \text{Prob}\{DS \geq DS_i | IM = im\}\) can be rewritten as

\[
G_i(im) = \frac{\text{Prob}\{DS \geq DS_i, IM = im\}}{\text{Prob}\{IM = im\}}
\]

(10)

Using the kernel density estimation in Equation (5), the empirical fragility function for DS \(i\) given the ground motion intensity value of \(im\) is defined as

\[
\hat{G}_i(im) = \frac{1}{n} \sum_{p=1}^{n} I(ds_p \geq DS_i) \times I(im_p = im) \cdot \frac{1}{n} \sum_{q=1}^{n} I(im_q = im) \\
= \frac{1}{n} \sum_{p=1}^{n} \frac{1}{h} K\left(\frac{im - im_p}{h}\right) \cdot \frac{1}{n} \sum_{q=1}^{n} \frac{1}{h} K\left(\frac{im - im_q}{h}\right) \\
= \frac{1}{n} \sum_{p=1}^{n} K\left(\frac{im - im_p}{h}\right) \\
= \frac{1}{n} \sum_{q=1}^{n} K\left(\frac{im - im_q}{h}\right)
\]

(11)

where \(I(x)\) is an indicator function that is 1 if \(x\) is true and 0 otherwise. Note that this method considers the uncertainty in the IM measurements but not in the DS measurements.

3.2. One-dimensional Gaussian kernel smoothing for continuous damage measure

This method estimates the conditional mean and variance of the DM given the IM, denoted as \(\hat{\mu}_{DM|IM}\) and \(\hat{\sigma}^2_{DM|IM}\), respectively, and then fits a conventional function to this conditional distribution, such as the lognormal and beta distribution functions, using the method of moments. In other words, we can obtain the conditional probability distribution of the DM given the IM. This method is particularly useful when the DM is a continuous value and/or when the conditional density function of the DM needs to be convoluted with other conditional density function for further risk analysis, and thus the full conditional distribution of the DM given IM is necessary. The estimates of the conditional mean, \(\hat{\mu}_{DM|IM}\), and the conditional variance, \(\hat{\sigma}^2_{DM|IM}\), for the IM value of \(im\) can be computed using a kernel as follows:

\[
\hat{\mu}_{DM|IM=im} = \frac{\sum m d_m \times K\left(\frac{im - im_m}{h}\right)}{\sum n K\left(\frac{im - im_n}{h}\right)}
\]

(12)
\[ \sigma_{\text{DM}|\text{IM}=im}^2 = \sum_m \left( dm_m - \mu_{\text{DM}|\text{IM}=im_m} \right)^2 \times K \left( \frac{im - im_m}{h} \right) \frac{1}{\sum_n K \left( \frac{im_m - im_m}{h} \right)} \] (13)

Once the mean and the variance are computed, a conventional distribution function can be fitted to the conditional distribution of the DM given the IM by the method of moments. For example, the obtained mean and variance are used as the mean and the variance of the lognormal or beta distribution function, and the corresponding parameters are obtained.

### 3.3. Two-dimensional Gaussian kernel smoothing for continuous damage measure

Alternatively, we can estimate the conditional probability of the DM given the IM using a two-dimensional kernel as follows:

\[
\text{Prob}(\text{DM} = dm | \text{IM} = im) = \frac{1}{n} \sum_{p=1}^{n} \frac{1}{h_1 h_2} K \left( \frac{im - im_p}{h_1}, \frac{dm - dm_p}{h_2} \right) \frac{1}{n} \sum_{q=1}^{n} K \left( \frac{im - im_q}{h_1} \right) \frac{1}{h_2} \sum_{p=1}^{n} K \left( \frac{dm - dm_p}{h_2} \right) \] (14)

where \( K(x, y) \) is a two-dimensional kernel at \((x, y)\). This equation follows directly from the definition of the conditional probability and the kernel density estimation in Equations (5) and (10). For Gaussian kernel, \( K(x, y) \) is a bivariate standard Gaussian probability density function. In this formulation, it is assumed that the measurement errors for DM and IM are uncorrelated. If the two-dimensional kernel \( K(im, dm) \) can be factorized into \( K(im) \) and \( K(dm) \), then the previous equation can be rewritten as

\[
\text{Prob}(\text{DM} = dm | \text{IM} = im) = \frac{1}{n} \sum_{p=1}^{n} \frac{1}{h_1} K \left( \frac{im - im_p}{h_1} \right) \times \frac{1}{h_2} K \left( \frac{dm - dm_p}{h_2} \right) \] (15)

For convenience, we can fit a conventional PDF to the resulting nonparametric conditional probability distribution by minimizing the fitting error, such as a root-mean-square error (RMSE). The lognormal distribution is appropriate for this conditional probability at each IM value because the IM values are bounded by zero on the lower side. It is noted that the outcome of the two-dimensional kernel method, the conditional distribution of DM given IM, is same as that of the one-dimensional kernel introduced in Section 3.2. One of the main advantages of using the two-dimensional kernel is that we can directly compute the conditional probability of the DM given the IM without using the second moment approximation as before. In addition, this method considers the uncertainties in both the IM and the DM measurements unlike the previous method that considers the uncertainty of only the IM by using the one-dimensional kernel. The conditional distributions of DM given IM are bounded on the interval zero to one (no damage to collapse), thus the inverse normal cumulative distribution function is used to transform the DM. For an inverse normal cumulative distribution transformation, Equations (8) and (15) can be combined as
4. APPLICATIONS

The three kernel smoothing techniques introduced in Section 3 are applied to two sets of data. The two data sets are from field observations obtained after the 2010 Haiti earthquake and numerical simulations of a four-story steel moment resisting frame model, respectively.

4.1. Empirical fragility function using field observations from the 2010 Haiti Earthquake

4.1.1. Description of data. The data consists of field-based damage assessments conducted in 2010-2011 following the 2010 Haiti earthquake [16]. Led by the Haitian Ministry of Public Works, the evaluation was conducted by civil engineers and architects having been trained in the ATC-20 damage assessment methodology [17]. Damage was categorized by assigning a DS as per ATC-13 [18]: none, slight, light, moderate, heavy, extreme, and destroyed. The modified Mercalli intensity (MMI) scale values obtained from the USGS shakemap [12] were used as the ground motion IM. These are not observed MMI values, but rather ‘instrumental intensity’ values, estimated through a combined regression of peak acceleration and velocity amplitudes, as well as USGS’ ‘did-you-feel it?’ data.
4.1.2. Results. The scatter plot of the data is shown in Figure 2(a). The samples at each MMI level and the corresponding ratio of buildings exceeding the DS of interest are shown as a circle whose size is proportional to the number of buildings at that MMI level. The data from over 250,000 buildings are included in this figure.

Data binning is often used to identify trends or even to develop non-parametric regression curves. The data are divided into bins of ground motion intensity and the damage averaged over that bin. Binning, however, introduces additional error to the data. One of the main issues with data binning for fitting a fragility curve is that the resulting curve is very sensitive to the subjective choice of binning. The curves in Figure 2(b) and (c) have the same bin size, but the bins are shifted, resulting in different curves. Reducing the bin size also results in a very different fitted curve, as shown in Figure 2(d). The figure also demonstrates two other common disadvantages of using binned average to fit fragility curves. When binning, all IM values within a bin are considered the same, even if they can vary quite broadly within a single bin. Furthermore, DMs are computed only at each bin, resulting in sparse representation of fragility curves. For instance, Figure 2(d) has a gap around MMI level 6, for which no data exist. For these reasons, GKS is recommended as a mean to develop fragility curves that are smooth and maximize the use of all data points, as explained in Section 3.

Figure 3(a) shows fragility curves for the extreme DS, obtained by weighting every data point by the Gaussian kernel centered at each IM. As can be seen from Figure 3(a) and (b), the choice of kernel range, or the variance of the Gaussian function, influences the resulting curve, yet the influence is primarily on the smoothness of the curve rather than its shape. The kernel method can also be used to obtain the standard deviation of the fragility function following Equation (13), as seen in Figure 3(c). Figure 3(d) demonstrates that a continuous curve can still be created using the kernel smoothing method even when there is a large data gap, which was not possible using binning as shown in Figure 3(d). Finally, we note that kernel smoothing results in significantly smoother curves than...
those obtained from binning. Table II summarizes the advantages of using the kernel method. Figure 3 also provides a comparison between GKS and a parametric fragility curve (dashed line) whose functional form is that of a lognormal cumulative density function and fit by maximum likelihood estimation. Indeed, the lognormal CDF has often been used to model earthquake damage fragility [4, 5, 7, 8, 10, 11]. Such parametric functions have the advantage of being very simple (summarized with only two parameters). As is seen in Figure 3, however, they cannot reflect the complexity of the data, such as local non-linearity, due to their parametric constraints.

Figure 4 shows the mean damage factor, as well as three full distributions of damages conditioned on IM = {7, 8.5 and 10}. Although DSs are ordinal data, it is common to analyze such data by assuming that they have an underlying continuous distribution [19]. The expected (mean) damage factor reflects the assumption that the ordinal DS data represent an underlying latent variable of damage. Scores are assigned to each DS corresponding to a central damage factor as defined in ATC-13 [18]. Equation (12) can then be used to develop a mean damage factor curve. Equation (13) is then used to obtain the complete damage distribution conditioned on IM. Assuming a lognormal conditional distribution ensures that damage is bounded at zero damage. Other distributions may also be used, such as the beta distributions, the moments of which can be obtained similarly by the method of moments.

The two-dimensional GKS provides further insight into the behavior and response of structures to earthquake ground motion. This is carried out by developing a complete joint-probability plane, giving the probability of every DM-IM pair. Assuming that the measurement error in DM and IM are independent, Equation (15) can be used to obtain the conditional probability of damage given IM. Figure 5 shows that for higher IM levels, buildings do not simply shift to higher DSs, but there is a disproportionate increase in extreme damage. This is consistent with the expected behavior of very brittle buildings, such as those commonly found in Haiti.

### 4.2. Analytical fragility function using a four-story steel moment resisting frame model

#### 4.2.1. Description of data

A set of numerically simulated data is obtained from a four-story two-bay steel special moment-resisting frame model. This frame is a perimeter lateral resisting system of an office building designed in Los Angeles based on current seismic provisions such as the IBC [20].

Table II. Advantages of the Gaussian kernel density estimation in comparison to the data binning method.

<table>
<thead>
<tr>
<th>Kernel density estimation</th>
<th>Data binning method</th>
</tr>
</thead>
<tbody>
<tr>
<td>All the available ({im_i, dm_i}) pairs are used with different weights to estimate the fragility function for damage state (j), (\hat{G}_j(IM)). Thus, this method reduces problems caused by lack of ({im_i, dm_i}) data or biased sampling.</td>
<td>Lack of data or biased data can result in some bins with no data—(\hat{G}_i(IM)) values cannot be defined for these bins (Figure 3(d)).</td>
</tr>
<tr>
<td>(\hat{G}_i(IM)) values can be computed at all the values of IM, thus resulting in a dense representation of (\hat{G}_i(IM)). Also, all IM values are considered as they are measured.</td>
<td>(\hat{G}_i(IM)) values can be computed only at each bin, resulting in a sparse representation of (\hat{G}_i(IM)) (Figure 3(b) and (c)). All the IM values within a bin are considered the same (i.e., they are treated as the center value of the bin). This can introduce a bias in the fit (Figure 3(a)-(c)).</td>
</tr>
<tr>
<td>Discretization error for IM is avoided.</td>
<td>A small measurement error may assign a sample to a different bin, which may affect the regression results significantly.</td>
</tr>
<tr>
<td>The resulting fragility function is a smooth curve and the degree of smoothness can be varied to reflect the uncertainty in the data.</td>
<td>The resulting fragility function is not always smooth (Figure 3(b) and (c)).</td>
</tr>
<tr>
<td>Uncertainty in the IM and DM values in the data can be accounted for by the Kernel’s bandwidth</td>
<td>It is difficult to include the uncertainty in IM and DM measurements.</td>
</tr>
</tbody>
</table>

GKS, Gaussian kernel smoothing; DS, damage state; DM, damage measure; IM, intensity measure.
and the AISC [21]. This model is subjected to 40 ground motions, each of which is scaled to various intensities. The spectral acceleration at the first mode period ($S_a(T_1, 2\%)$), measured in g, is used as an IM of the ground motion, and the story drift ratio (SDR) is used to quantify a DM. More details about the model, data, and the analysis procedure can be found in Lignos and Krawinkler [22] and Noh et al. [7].

4.2.2. Results. Figure 6(a) and (b) show the scatter plot of IM and DM and the fragility functions obtained using the data binning methods, whereas Figure 6(c) shows the fragility functions obtained from the one-dimensional GKS method with a bandwidth of 0.4. The five discrete DSs, DS0, DS1, ..., DS4, are defined as no damage (i.e., within the elastic limit) ($0\% \leq SDR < 1\%$), slight damage ($1\% \leq SDR < 2\%$), moderate damage ($2\% \leq SDR < 3\%$), severe damage ($3\% \leq SDR < 6\%$), and collapse ($6\% \leq SDR$), respectively. These threshold values for SDR are selected as representative values to describe different DSs based on current practice (FEMA 356 and FEMA 440) [15, 23].

We observe that the GKS method provides a smooth and continuous fragility functions unlike the data binning method as explained in Table II. Using the continuous DM, we can obtain a more detailed relationship between IM and DM. Figure 6(d) shows the conditional mean and standard deviation of DM given IM for various IM values. Figure 7 shows the scatterplot of the data and the conditional mean with one standard deviation above and below it. The right panel of Figure 7 Lognormal distribution is fitted using the method of moments based on the conditional mean and standard deviation for this figure. We can observe that as the IM values increases, both the conditional mean and the variance of the DM increases.
Figure 8 shows the conditional distribution of DM given the IM values of 0.7, 1.3, and 1.9 g using two-dimensional GKS method, with the bandwidths of 0.1 and 0.01 for IM and DM, respectively. The conditional distributions on the right panel are obtained from the two-dimensional GKS without fitting the lognormal distribution. We can observe further details of structural damage in response to different ground motion intensities. In particular, the solid line (IM value of 1.3 g) shows a very irregular shape of probability distribution and has high probability for large damage. This phenomenon is observed because most of collapses occurred around this IM value. This information could have been lost if conventional distribution fitting methods were used due to parametric constraints.

Figure 6. Relationship between intensity measure (IM) and damage measure (DM) for the numerical simulation: (a) scatter plot of IM versus DM; (b) fragility functions using data binning; (c) fragility functions using Gaussian kernel smoothing with bandwidth of 0.4; and (d) conditional mean and standard deviations.

Figure 7. One-dimensional kernel application to the numerically simulated data with a bandwidth of 0.4.
4.3. Discussions

4.3.1. Over-fitting. It is observed that, because the fragility functions based on GKS are no longer constrained to a particular functional shape, it tends to more closely fit the data, with the possible danger of over-fitting. This is entirely dependent on the characteristics of the data used and the kernel bandwidth. The width of the kernel is a critical parameter in the smoothing method and its choice is based on a bias-variance trade-off:

- A narrow bandwidth will result in higher variance of the resulting fragility functions, as fewer contributing points are averaged within the kernel bandwidth. On the other hand, the bias will be small, as the results fit the data closely.
- A wide bandwidth will have small variance because more values are being averaged, but will have higher bias, because points further away are incorporated.

The effect of different bandwidths is illustrated in Figure 9. In practice, various bandwidths can be tested using engineering judgement and prior information about data uncertainties until an adequate one is found. If Gaussian assumption is appropriate, the Silverman’s optimum bandwidth can be used as shown in Equation (7). Finally, it is noted that the bandwidth does not have to be constant across the domain. This is particularly the case when data are not equally spaced.

4.3.2. Bias. One common issue with kernel smoothing is that the resulting functions exhibit bias at or near the boundaries of the domain, as shown in Figure 10, because data exist only on one side of the kernel. It is analogous to applying asymmetric kernels near the boundary. A similar issue can arise within the domain in a less severe way when data are not equally spaced [24]. The issue of localized bias at the boundary results in non-zero probabilities of damage at negligible IM values, which clearly does not reflect underlying physical constraints. Typically, the bias is proportional to

Figure 8. Two-dimensional kernel application to the numerically simulated data with bandwidths of 0.1 and 0.01 for intensity measure and damage measure, respectively.

Figure 9. Effect of Gaussian kernel bandwidth on bias and variance.
the size of the kernel at the boundary. Therefore, a kernel with varying bandwidth can be used, with small bandwidth at the boundaries to reduce the bias, as shown in Figure 10(a). Note that varying bandwidth can also resolve the bias at the other boundary for large IM values. The kernel with varying bandwidth results in a steeper slope at large IM values, which reflects the data more accurately. The issue of non-zero probability of damage at negligible IM can also be addressed through zero-padding at the boundary. This process simply adds artificial values of zero at the lowest IM level and is displayed in Figure 10(b). Due to the kernel weighting of every data point, the effect of zero-padding is localized near the boundary and therefore has no effect at higher IM levels.

A more rigorous, although more involved, method to solve the issue of localized bias is through local linear regression. In this approach, straight lines rather than constants are fit through kernel weighted linear regression. This is carried out by solving a weighted least square problem at each target IM value of \( \text{im}_0 \):

\[
\min_{\alpha(\text{im}_0), \beta(\text{im}_0)} \sum_{i=1}^{N} K \left( \frac{\text{im}_0 - \text{im}_i}{h} \right) \left( y_i - \alpha(\text{im}_0) - \beta(\text{im}_0)\text{im}_i \right)^2
\]

where \( \alpha(\text{im}_0) \) and \( \beta(\text{im}_0) \) are coefficients that vary with \( \text{im}_0 \) value. This is solved for each \( \text{im}_0 \) value to obtain a local linear regression curve:

\[
\hat{G}(\text{im}_0) = \alpha(\text{im}_0) + \beta(\text{im}_0)\text{im}_0
\]
The closed form solution is described as

\[ \hat{G}(im_0) = (1, im_0)(B^T W(im_0)B)^{-1}B^T W(im_0)y \]  

where \( B \) is a \( N \) by 2 matrix with the \( i^{th} \) row equal to \( (1, im_i) \) and \( W(im_0) \) is a diagonal matrix with the \( i^{th} \) diagonal element equal to \( K(im_0, im_i) \). Figure 11 shows the results of fragility functions using GKS (solid line) and weighted local linear regression (dash line). We can observe that weighted local linear regression reduced the bias, particularly near the boundaries, resulting in a less flat fragility function.

4.3.3. Confidence intervals. The fragility functions obtained from the methods described earlier provide the expected probability of damage as a function of ground motion intensity. These predictions have some uncertainty, which can be characterized in terms of confidence intervals. It is noted that confidence interval refers to the uncertainty in the fragility functions itself, resulting from the uncertainty in the function estimates. One common method to obtain the confidence intervals on regression curves is through the non-parametric bootstrap method. In this approach, regression analysis is conducted consecutively for numerous random samples of the data. The process involves randomly sampling data sets with replacement from the original data set, each of which has equal sample size [24]. The fragility functions are computed for each bootstrap data set, from which a distribution of fragility functions is obtained. The point-wise \( \alpha \)% confidence interval can then be obtained by taking the \( 100\alpha/2 \)th and \( 100\alpha/2 \)th percentiles of all bootstrap functions at each IM level, which is referred to as the bootstrap percentile interval.

Figure 12 shows the 95% confidence interval obtained by the bootstrap method. The region between dashed lines indicates the 95% confidence interval, and the solid line is the mean fragility function. Note that the confidence intervals reflect the uncertainties from the data used in the analysis as well as the assumptions made in the fragility function estimation. For this example, a sample of 10000 data points from the Haiti dataset was used as demonstration, because the uncertainty on the entire set of 250000 is negligible (uncertainty due to limited data).

5. CONCLUSIONS

This paper introduces a new framework for computing fragility functions from empirical and analytical data using GKS methods. The kernel smoothing methods estimate a functional relationship between two variables by taking weighted average of nearby data, and this weighting function is referred to as a kernel. Three different methods are presented in this paper to provide different levels of information for discrete and continuous DMs.

The first method uses one-dimensional kernel to compute fragility functions for discrete DSs while the second and the third methods use one-dimensional and two-dimensional kernels, respectively, for

Figure 12. Confidence intervals developed from non-parametric bootstrap for Gaussian Kernel Smoothing with kernel bandwidth of 0.3.
continuous DMs. The latter two provide more detailed information about the relationship between IM and DM, because they allow us to take advantage of the continuous characteristics of DM instead of discretizing it as many conventional methods do.

For validation, the three kernel-based methods are applied to two sets of data collected from field observations after the 2010 Haiti earthquake and numerical simulations of a four-story steel special moment-resisting frame. The results demonstrate that conditional distributions of DM given IM can be estimated for various forms of DM. Moreover, the results show that the kernel-based methods can reduce the bias due to discretization and parameterization (e.g., data binning and curve fitting).

This paper also investigates over-fitting and bias issues regarding the fragility function estimation using the kernel smoothing method and presents possible solutions. In addition, it demonstrates how to obtain confidence intervals using bootstrap. Variable transformation method is introduced to resolve the issue of unbounded conditional distributions, which result from applying the GKS method for fragility function estimation. With careful selection of kernel bandwidth, boundary correction, and uncertainty quantification, the kernel smoothing method provides the means to estimate fragility functions from empirical and analytical data that are noisy, sparse, and/or non-homogeneously distributed.

ACKNOWLEDGEMENTS

The research presented in this paper is partially supported by the Global Earthquake Model, National Science Foundation CMMI Research Grants No. 0800932 and 106756, the John A. Blume Fellowship, and the Samsung Scholarship. This support is greatly appreciated. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the sponsors.

REFERENCES


