Cal Poly

Caltech

SCEC
AN NSF+USGS CENTER

UC Irvine

UCLA

UC Santa Barbara

USC

Hazard-Consistent Scenario-Based Correlated Ground Motions for California Gas Pipeline Infrastructure

Pengfei Wang, PhD (CEE, Old Dominion University, VA)

Linda Al Atik, PhD (Linda Alatik Consulting, San Francisco, CA)

Nick Gregor, PhD (Gregor Consulting, Oakland, CA)

Nicolas Kuehn, PhD (CEE & NHR3 Center, UCLA, CA)

Melanie A. Walling, PhD (GeoEngineers, Inc., WA)

Albert R. Kottke, PhD (PG&E, Oakland, CA)

Zehan Liu (CEE, UCLA, CA)

Paolo Zimmaro, PhD (University of Calabria, Italy)

Yousef Bozorgnia, PhD (Samueli Engineering, UCLA, CA)

Scott J. Brandenberg, PhD (Samueli Engineering, UCLA, CA)

Jonathan P. Stewart, PhD (Samueli Engineering, UCLA, CA)

A report on research supported by the California Energy Commission

Report GIRS-2023-02

DOI: 10.34948/N3001C

University of California, Los Angeles (headquarters)

Natural Hazards Risk & Resiliency Research Center

B. John Institute for the Risk Sciences

Hazard-Consistent Scenario-Based Correlated Ground Motions for California Gas Pipeline Infrastructure

Pengfei Wang, PhD (CEE, Old Dominion University, VA)

Linda Al Atik, PhD (Linda Alatik Consulting, San Francisco, CA)

Nick Gregor, PhD (Gregor Consulting, Oakland, CA)

Nicolas Kuehn, PhD (CEE & NHR3 Center, UCLA, CA)

Melanie A. Walling, PhD (GeoEngineers, Inc., WA)

Albert R. Kottke, PhD (PG&E, Oakland, CA)

Zehan Liu (CEE, UCLA, CA)

Paolo Zimmaro, PhD (University of Calabria, Italy)

Yousef Bozorgnia, PhD (Samueli Engineering, UCLA, CA)

Scott J. Brandenberg, PhD (Samueli Engineering, UCLA, CA)

Jonathan P. Stewart, PhD (Samueli Engineering, UCLA, CA)

A report on research conducted with support from the California Energy Commission

Report GIRS-2023-02
DOI: 10.34948/N3001C

Natural Hazards Risk and Resiliency Research Center
B. John Garrick Institute for the Risk Sciences
University of California, Los Angeles (Headquarters)

February 2023

Abstract

Spatial representations of the results of conventional probabilistic seismic hazard analyses (PSHA) typically take form of hazard maps, which represent results of independent analyses at many locations within the map. Hazard maps are not suitable for assessing risk to spatially distributed infrastructure (such as natural gas pipelines) because they significantly overpredict shaking intensities that the system could experience and thus the number and spatial extent of potential failures. This occurs because (1) hazard maps represent the aggregated effects of many events, and those events are likely to be different in different parts of a broad infrastructure system and (2) hazard maps provide relatively consistent levels of ground motions in space, whereas real earthquakes have more complex patterns that are reflected in models of the spatial correlation of ground motions. To overcome these problems, we present in this report an alternate scenario-based method, which utilizes spatially correlated hazard-consistent ground motions.

We describe the basis for the methodology and introduce procedures for developing spatially correlated hazard-consistent ground motions for seismic hazard risk analysis. There are three major steps: 1) conduct conventional point-based PSHA to obtain hazard curves and disaggregations as input (presented in Al Atik et al. 2022); 2) select hazard-consistent scenario earthquake events; and 3) generate spatially correlated ground motion realizations for each selected scenario event and select a manageable subset of hazard-consistent ground motion realizations. This report presents the methodology for Steps 2 and 3.

The aim of the scenario event selection is a manageable event subset that, in aggregate, approximately matches the hazard for single or multiple ground motion intensity measures across the spatially-distributed system while preserving contributions of different magnitudes and distances to the PSHA. We present a flexible and efficient regression-based method that meets these requirements using point-based PSHA results as inputs. The ground motion selection methodology is formulated similarly, but instead of selecting a subset of events among many candidate events, it selects realizations of ground motion from all selected events among many such possible realizations.

The procedure was applied to derive correlated hazard-consistent ground motion realizations from scenarios events for application to risk analyses for California natural gas pipeline infrastructure. We selected 1,220 gridded sites that are within 1 km of gas pipelines as the target hazard sites. We applied the regression-based method to select 599 hazard-consistent scenario events to preserve the hazard curves for Peak Ground Acceleration (PGA) and Peak Ground Velocity (PGV) from return periods of 200 years to 2,475 years and their magnitude distributions from disaggregation at the 1,220 target sites. We subsequently applied the regression-based method a second time to select 25 hazard-consistent correlated ground motion distributions for both PGA and PGV from scenario events. Lastly, we implement co-Kriging to interpolate the selected correlated maps to a 100 m square resolution.

Acknowledgments

Financial support for this research project was provided by the California Energy Commission, Pacific Gas & Electric Company, and California Department of Transportation. The support of these organizations is greatly appreciated. This research was also supported by the Research Computing clusters at Old Dominion University for computation.

The opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the study sponsors, the Natural Hazards Risk and Resiliency Research Center, or the Regents of the University of California.

Contents

List of Figures

List of Tables

1 Introduction

Conventional probabilistic seismic hazard analysis (PSHA) is performed at a specific site to quantify the annual probability of exceedance of a ground motion intensity measure based on the statistical distributions of ground motions that each of a large number of earthquake events might produce at the site. Magnitude (**M**) recurrence relations for earthquake sources are discretized into scenario events within magnitude bins, with each event having a rate of occurrence, a magnitude, and a location. Ground motion at the site is then characterized as a probabilistic function of event magnitude, source-to-site distance, and site conditions. Rates of exceedance of ground shaking are then summed over all of the considered events to form a hazard curve (McGuire, 2004). Hazard maps can then be generated by repeating the PSHA at many sites and then assembling ground motions from the hazard curves at a uniform exceedance rate across sites.

Uniform hazard maps are useful for characterizing demand for discrete infrastructure components, such as buildings. However, they are not suitable for assessing risk to spatially distributed infrastructure. First, no single event could produce the same ground shakings as the hazard map across a spatially distributed system. Second, the spatial correlation of ground motions within an event is a key component for spatially distributed infrastructure risk analysis but it is missing in the hazard map (Chang et al., 2000; Campbell and Seligson, 2003). Alternatively, a robust but computationally expensive approach is to analyze the spatially distributed infrastructure system for all possible spatial correlated ground motions produced by every event considered in the PSHA. However, PSHA often involves hundreds of thousands of events, each of which can produce infinite numbers of spatially correlated ground motion realizations. Therefore, this approach is generally not practical. Han and Davidson (2012) proposed a framework for regional probabilistic seismic risk analysis by using a finite number of hazard consistent spatially correlated ground motion maps.

We adopt their approach but propose a new framework, as shown in Figure 1.1. This framework contains five steps: target hazard calculation, event scenarios pre-selection, event scenarios final selection, ground motion map scenarios generation, and ground motion map scenarios selection. The first step, target hazard calculation, refers to the conventional point-based PSHA calculations for all sites in the study region. The results of hazard curves and disaggregations and the events considered in the PSHA are the inputs for the following steps. This step has been completed separately from the present effort and is described by Al Atik et al. (2022). The second and third steps operate together to select hazard-consistent scenario events. The second step uses the disaggregation results to pre-select a subset of events with the largest hazards contributions and the third step takes the pre-selected events and implements a newly proposed Least Absolute Shrinkage and Selection Operator (LASSO) regression-based method to select a final set of events with adjusted annual occurrence rates that make them hazard consistent with the full event set from PSHA. The fourth step generates a significant number of ground motion realizations using one or multiple intensity measures in which correlation structures are applied both spatially and between intensity measure types for each of the selected events. The fifth step applies the LASSO regression again to select the final set of ground motion maps and adjust their annual occurrence rates for hazard consistency.

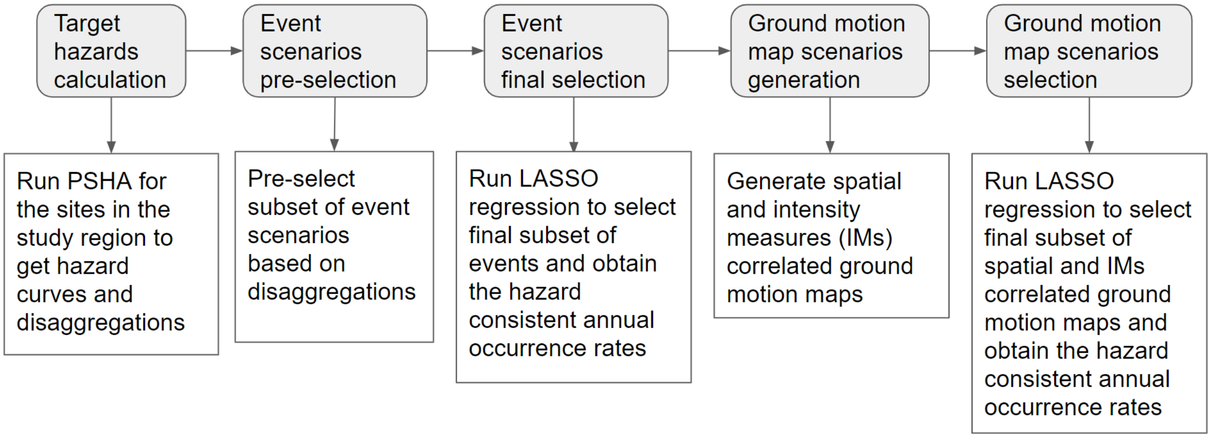

Target hazards calculation
Event scenarios pre-selection
Event scenarios final selection
Ground motion map scenarios generation
Ground motion map scenarios selection
Run PSHA for the sites in the study region to get hazard curves and disaggregations
Pre-select subset of event scenarios based on disaggregations
Run LASSO regression to select final subset of events and obtain the hazard consistent annual occurrence rates
Generate spatial and intensity measures (IMs) correlated ground motion maps
Run LASSO regression to select final subset of spatial and IMs correlated ground motion maps and obtain the hazard consistent annual occurrence rates

Figure 1.1. Framework for spatial correlated hazard consistent ground motion map scenarios.

In this report, we first describe the mathematical formulation for region-specific hazard calculations in Chapter 2, which are the bases of our methodology for event and ground motion map scenarios selection. In Chapter 3, we introduce LASSO regression and a series of required data manipulations (tensor rank reduction transformation, matrix representation of distributions) to be able to implement LASSO for scenarios selection. In Chapter 4, we describe how to generate ground motion maps that preserve spatial and cross-intensity measure correlations and how to re-implement LASSO for ground motion map scenario selection. Chapter 4 also describes co-Kriging interpolation to densify selected correlated ground motion maps. We present in Chapter 5 the steps for selecting the final set of correlated ground motions and show some representative example results. Finally, we summarize our conclusions, limitations of the study, and future work opportunities in Chapter 6.

2 Region-Specific Hazard Calculation

Ground motion hazard curves express the annual exceedance rate of ground motion at the site as a function of a specified ground motion intensity measure (*IM*) level x, as described in Eq. (2.1) based on McGuire (2004),

$$\lambda(IM > x) = \sum_{i=1}^{N_E}(v_i P(IM > x \mid M_i, R_i)) = \sum_{i=1}^{N_E} \Lambda_i(IM > x) \tag{2.1}$$

where IM is a log-normally distributed random variable whose mean and standard deviation are specified by ground motion models (GMMs) (e.g., Boore et al., 2014), x is a specific value of *IM*, $\lambda(IM > x)$ is the annual rate at which *IM* exceeds *x* due to all considered events (or total hazards), N_E is the total number of considered events, v_i is the annual occurrence rate of the ith event, and $P(IM > x \mid M_i, R_i)$ is the probability that *IM* exceeds x [shaded area in Figure 2.1(a)] given the event with magnitude M_i and site-to-source distance R_i occurs. The product of annual occurrence rate and exceedance probability is herein defined as $\Lambda_i(IM > x)$, which represents the annual ground motion exceedance rate or the hazard produced by event i, which can also be conceptualized as an event-specific hazard curve. Given the predicted log-normal distribution by GMMs, $f_{\mathrm{d}}(IM|M_i, R_i)$ (probability density function), the ground motion exceedance probability in the shaded area in Figure 2.1(a) can be calculated by

$$P(IM > x \mid M_i, R_i) = 1 - \Phi\left(\frac{ln(x) - \mu_{GMM}(M_i,R_i)}{\sigma_{GMM}(M_i,R_i)}\right) \tag{2.2}$$

where $\mu_{GMM}(M_i, R_i)$ is the predicted logarithmic mean ground motion, $\sigma_{GMM}(M_i, R_i)$ is the associated standard deviation, $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution, and $\frac{ln(x) - \mu_{GMM}(M_i,R_i)}{\sigma_{GMM}(M_i,R_i)}$ is the standardized z value.

For a site-specific PSHA in Eq. (2.1), the calculation is more commonly expressed as a summation of integrations over different magnitudes and distances for each considered event rather than a discrete summation over all events. The infinite set of possible events that might occur on a fault (e.g., fault rupture location and magnitude) is discretized into a discrete set of events with individual annual rates to set a rate of moment release that is compatible with the rate of moment build-up on a fault, which is strongly correlated with the slip rate on the fault. The rate for each discrete event is therefore a function of the fault slip rate and the discretization strategy, and selecting a larger number of events will result in lower annual rates for each event and viceversa. Often, hundreds of thousands of events are considered [e.g., UCERF3 (Field et al. 2015) utilizes over 400,000 fault rupture events for California]. Hazard integrals are then evaluated numerically in discrete form. We represent them from the outset in the discrete form here, which is better suited for our scenario selection methodology.

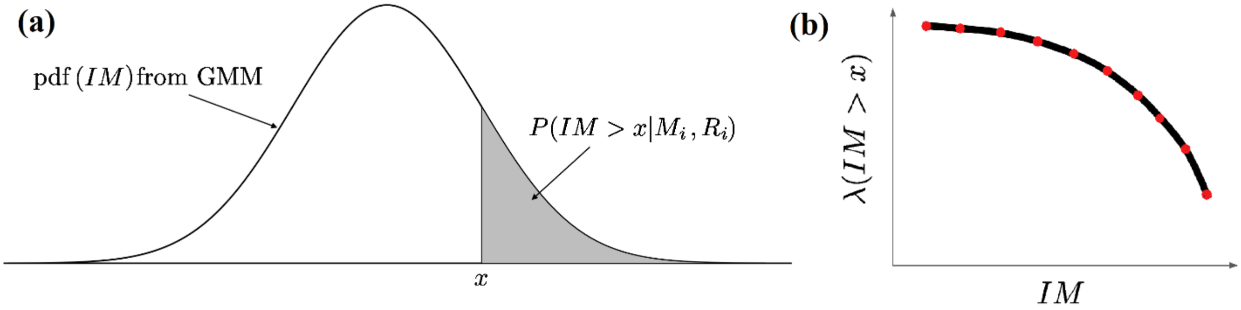


Figure 2.1. (a) Schematic plot of ground motion exceedance probability and (b) hazard curve.

For the problem at hand, we wish to select a manageable subset of events that, in a least-squares sense, preserves the hazard curves at multiple sites for multiple intensity measures. Reducing the number of events requires an increase in the annual rate of each event to preserve hazard. Furthermore, we optionally wish to match, in a least squares sense, the magnitude and distance distributions from the PSHA disaggregation at each site and intensity measure level (or return period). Variables utilized in the derivation that follows are defined in Table 2.1.

Table 2.1. Variables considered in regional PSHA.

Meaning	Index	Number of elements
Significant events	i	N_E
Site	j	N_S
Intensity measure type	k	N_T
Intensity measure level	l	N_X
Magnitude bins	b	N_M
Distance bins	d	N_R

Assuming the hazard curve [e.g., Figure 2.1(b)] at a site is represented in discrete form by N_X number of separate x values (intensity measure levels), we can define rank-1 and rank-2 tensors, ${}^1\lambda_l$ and ${}^2\Lambda_{l,i}$, to represent the total hazard curve and the hazard curve produced by each event i, respectively. The left superscripts 1 and 2 of ${}^1\lambda_l$ and ${}^2\Lambda_{l,i}$ indicate the ranks of the tensors, which are equal to the number of indices (i.e., the conditions that may be varied in a particular analysis, as reflected by the different rows in Table 2.1). Therefore, Eq. (2.1) can be represented for one intensity measure type at one site by Eq. (2.3).

$$ {}^{1}\lambda_{l} = \sum_{i=1}^{N_E} {}^{2}\Lambda_{l,i} \tag{2.3}$$

Equivalently, we can also discretize the hazard curve horizontally by different hazard levels λ or return periods $RP = 1/\lambda$.

For regional PSHA, ${}^{1}\lambda_{l}$ and ${}^{2}\Lambda_{l,i}$ must be computed at N_S different sites (from $j = 1$ to $j = N_S$), which can be represented by adding an index j ($j \in \{1, \cdots, N_S\}$), thereby increasing the tensor ranks, as defined by Eq. (2.4).

$$ {}^{2}\lambda_{l,j} = \sum_{i=1}^{N_E} {}^{3}\Lambda_{l,j,i} \tag{2.4}$$

Similarly, we can introduce different types of intensity measures (e.g., PGA, PGV, and pseudo-spectral acceleration, PSA at different oscillator periods) by introducing an index k ($k \in \{1, \cdots, N_T\}$), thereby further increasing the tensor ranks as defined by Eq. (2.5).

$$ {}^{3}\lambda_{l,j,k} = \sum_{i=1}^{N_E} {}^{4}\Lambda_{l,j,k\ i} \tag{2.5}$$

Having now established equations for a single site single *IM* PSHA [Eq. (2.3)], a single *IM* regional PSHA [Eq. (2.4)], and a multi-*IM* regional PHSA [Eq. (2.5)], we will next describe our proposed regression-based method for event selection. Further expansion of tensor ranks to consider magnitude and distance distributions is discussed after defining the regression methodology.

3 Methodology for Scenario Earthquake Selection

In this chapter, we will first describe LASSO regression as a tool for event selection. The method requires a tensor rank reduction transformation that is explained. We also present a new representation of magnitude and distance distributions from disaggregation such that they can be incorporated into the event selection framework.

3.1 LASSO

The Least Absolute Shrinkage and Selection Operator (LASSO) (also known as regression with L1 regularization) proposed by Tibshirani (1996) is a regularized regression method capable of performing variable selection. Unlike linear regression, which aims to find the best coefficients to minimize the sum of squared residuals, LASSO minimizes the sum of squared residuals and simultaneously reduces the number of variables. The LASSO regression formula and its objective function can be expressed by Eq. (3.1) and (3.2),

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{3.1}$$

$$\arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X\beta})^T (\boldsymbol{y} - \boldsymbol{X\beta}) + \gamma ||\boldsymbol{\beta}||_1 \tag{3.2}$$

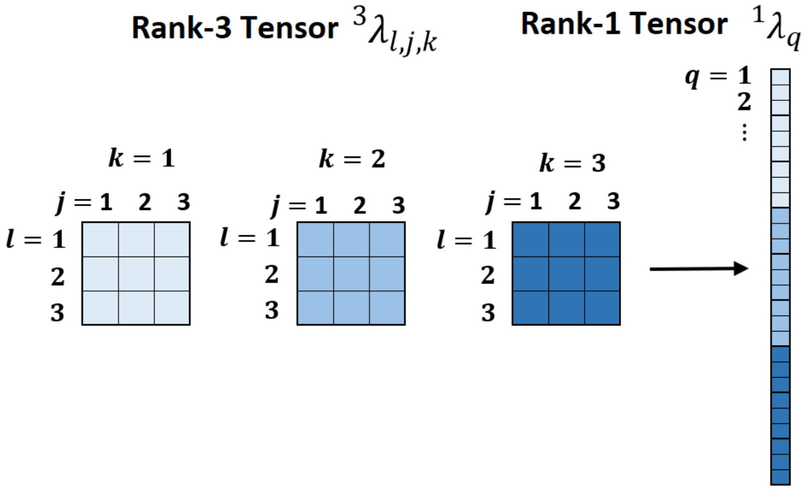
where $\boldsymbol{y}$ is the target response vector (i.e., a column of values), $\boldsymbol{X}$ is the predictor matrix, $\boldsymbol{\beta}$ is the coefficient vector, $\boldsymbol{\varepsilon}$ is the error variable (follows multivariate normal distribution), $|| \cdot ||_1$ is L1 norm (sum of absolute values), $(\boldsymbol{y} - \boldsymbol{X\beta})^T (\boldsymbol{y} - \boldsymbol{X\beta})$ is the error value, which is calculated as the inner product of $(\boldsymbol{y} - \boldsymbol{X\beta})$ (or sum of squared errors), and γ is a parameter to tune the model. Increasing γ will reduce the sum of squared errors (as this term has a larger contribution to the total objective function), which results in more zeros in $\boldsymbol{\beta}$, where zeros are interpreted as unselected variables (i.e., those associated columns in $\boldsymbol{X}$ may be removed). On the other hand, decreasing γ leads to fewer zeros in $\boldsymbol{\beta}$ but a better fit for a smaller sum of squared residuals. If $\gamma = 0$, all of the variables are retained, and the LASSO regression is a multiple linear regression. We use bold notations to indicate vectors or matrices, which will be applied subsequently. In the context of multi-*IM* regional PSHA, if we substitute for $\boldsymbol{y}$ with ${}^{3}\lambda_{l,j,k}$ and $\boldsymbol{X}$ with ${}^{4}\Lambda_{l,j,k\;i}$, then the LASSO regression can be used to simultaneously select event subsets and coefficients (i.e., adjusted event rates). However, ${}^{3}\lambda_{l,j,k}$ and ${}^{4}\Lambda_{l,j,k\;i}$ are rank-3 and rank-4 tensors, respectively, whereas $\boldsymbol{y}$ and $\boldsymbol{X}$ in LASSO must be a vector (rank-1 tensor) and a matrix (rank-2 tensor), respectively. Thus, a rank reduction transformation is required before conducting LASSO regression for event selection.

3.2 Tensor Rank Reduction Transformation

We take ${}^3\lambda_{l,j,k}$ and ${}^4\Lambda_{l,j,k,i}$ as examples to introduce a transformation method for tensor rank reduction. The transformation method is also effective for higher-order rank reductions that are required when magnitude and distance distributions from disaggregation are also considered during event selection.

In Figure 3.1, we use an example of a target hazard with three *IM* types ($N_T = 3$) for three *IM* levels ($N_X = 3$) at three sites ($N_S = 3$) to illustrate the method of reducing a rank-3 tensor ${}^3\lambda_{l,j,k}$ to a rank-1 tensor ${}^1\lambda_q$. Each 3 by 3 block represents the regional hazard curves for each of the three *IM* types (i.e., $k = \{1, 2, 3\}$). The three columns in each block are the hazard curves at three sites (i.e., $j = \{1, 2, 3\}$). For each column, the entries in the three rows correspond to the annual exceedance rates at three different *IM* levels (i.e., $l = \{1, 2, 3\}$), which are the outputs of PSHA. The long column on the right expresses the transformed rank-1 tensor, which is transformed relative to the previous rank-3 tensors, and is obtained by re-arranging the 3 by 3 grids according to Eq. (3.3),

$$q = l + (j - 1) \times N_X + (k - 1) \times N_X \times N_S \tag{3.3}$$



Rank-3 Tensor ${}^3\lambda_{l,j,k}$
Rank-1 Tensor ${}^1\lambda_q$
$q = 1$
2
⋮
$k = 1$
$k = 2$
$k = 3$
$j = 1$ 2 3
$l = 1$
2
3

Figure 3.1. An example of rank reduction transformation from a rank-3 tensor (with indices l, j, k) to a rank-1 tensor (with index q) for a target hazard from all considered seismic sources.

Figure 3.2 illustrates the rank reduction of the rank-4 tensor, which includes index i for event (i.e., ${}^4\Lambda_{l,j,k,i}$), to a rank-2 tensor (${}^2\Lambda_{q,i}$). Each element in the rank-4 tensor represents the hazard produced by a single event i, not the total hazard from PSHA (right side of Eq. 2.5). We use the same 3 by 3 block structure to represent the regional hazard curves produced by each event. For the illustration in Figure 3.2, we consider three events (i.e., $i = \{1, 2, 3\}$), thus there are three 3 by 3 blocks for each *IM* type. The three long columns on the right then express the transformed rank-2 tensor, where each column corresponds to the hazard produced by a single event.

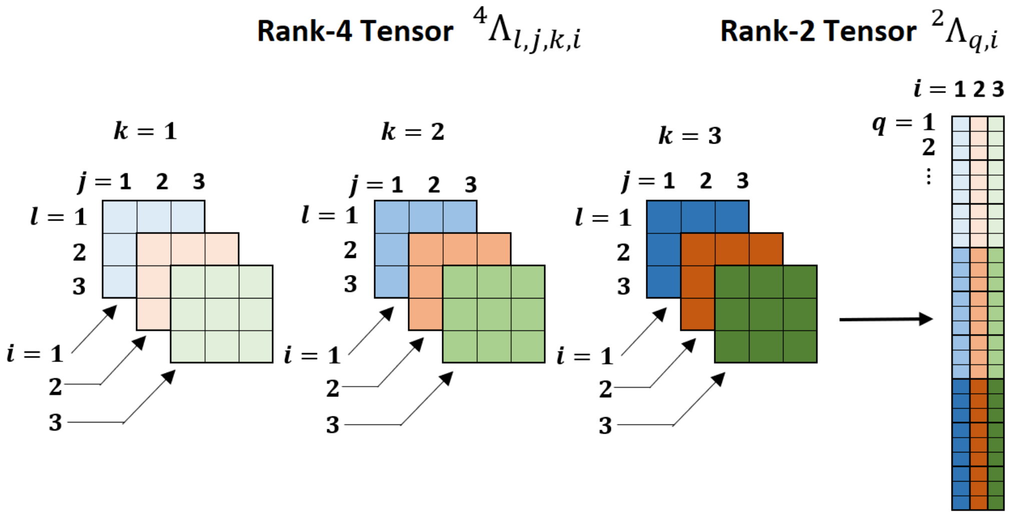

Rank-4 Tensor ${}^{4}\Lambda_{l,j,k,i}$
Rank-2 Tensor ${}^{2}\Lambda_{q,i}$
$k = 1$
$k = 2$
$k = 3$
$j = 1$ 2 3
$l = 1$
2
3
$i = 1$
2
3
$i = 1$ 2 3
$q = 1$
2
⋮

Figure 3.2. An example of rank reduction transformation from a rank-4 tensor (with indices l, j, k, i) to a rank-2 tensor (with indices q, i) for hazard produced by three events.

The transformation implied by Eq. (3.3) provides one re-ordering rule for tensor rank reduction in which we first stack the hazard curves for each site, then by each intensity measure type. However, other re-ordering operations could also be utilized if each row q in the total hazard ${}^{1}\lambda_q$ vector is consistent with the corresponding row in the event hazard matrix ${}^{2}\Lambda_{q,i}$ for the same *IM* level at the same site for the same *IM* type.

3.3 Event Selection

Following tensor rank reduction, we re-define the event selection by LASSO regression as,

$$\boldsymbol{\lambda} = \boldsymbol{\Lambda\beta} + \boldsymbol{\varepsilon} \tag{3.4}$$

$$\arg\min_{\boldsymbol{\beta}} (\boldsymbol{\lambda} - \boldsymbol{\Lambda\beta})^T \boldsymbol{W} (\boldsymbol{\lambda} - \boldsymbol{\Lambda\beta}) + \gamma ||\boldsymbol{\beta}||_1 \text{ and subject to } \boldsymbol{\beta} \geq \boldsymbol{0} \tag{3.5}$$

where $\boldsymbol{\lambda}$ is the rank-1 total hazard ${}^{1}\lambda_q$ that replaces response vector $\boldsymbol{y}$ from Eq. (3.1), and $\boldsymbol{\Lambda}$ is the corresponding rank-2 hazard matrix $\Lambda_{q,i}$ that replaces predictor matrix $\boldsymbol{X}$ from Eq. 3.1. The column vector $\boldsymbol{\beta}$ contains the associated rate adjustments for each event (or column) in $\boldsymbol{\Lambda}$, and $\boldsymbol{\varepsilon}$ represents the hazard misfits. For event i with an original annual rate of occurrence v_i, the adjusted rate after selection is $\beta_i \times v_i$ (β_i is the i-th element in $\boldsymbol{\beta}$). We constrain all elements of $\boldsymbol{\beta}$ to be non-negative ($\beta_i \geq 0$) to ensure that adjusted rates are physically meaningful. We apply a weighted LASSO regression in which $\boldsymbol{W}$ is a weighting diagonal matrix with diagonal elements equal to $\mathbf{1}/\boldsymbol{\lambda}$. We adopt this weighting scheme because hazard values are generally plotted on a log scale rather than a linear scale, and as a result, we apply equal weights to the logarithm of each data point in the regression. If this weighting was not applied, values at short return periods would have significantly more weight than those at long return periods (e.g. the weight for a return period of 50 years would be 1/50, which is roughly 50 times that for a return period of 2475 years, which would be 1/2475).

Equations 3.4 and 3.5 are fully general and can be applied for full PSHA without reducing the number of events by using all N_E events to develop $\boldsymbol{\Lambda}$, in which case $\boldsymbol{\beta} = \mathbf{1}$ (all elements in $\boldsymbol{\beta}$ are 1) and $\boldsymbol{\varepsilon} = \mathbf{0}$ (all elements in $\boldsymbol{\varepsilon}$ are 0). For event selection, we seek a subset of n events (n columns in $\boldsymbol{\Lambda}$) from the complete set of N_E events ($n < N_E$) and the corresponding rate adjustments in $\boldsymbol{\beta}$ that, in aggregate, are consistent with the total hazard $\boldsymbol{\lambda}$ at all sites, for all *IM* types, and all *IM* levels within certain error bounds represented by $\boldsymbol{\varepsilon}$. The regressed values of $\boldsymbol{\beta}$ are generally higher than 1.0 because rates of events in the reduced event set must be higher than those in the full set to overcome the omission of the unselected events. The number of selected events equals the number of positive elements in $\boldsymbol{\beta}$, which can be tuned by adjusting γ.

For a large region or a region where the seismicity is complex, the number of considered events (N_E) for PSHA is large (e.g., over 400,000 fault rupture events and over 2,500 grid point sources for background seismicity from two branches modeled by UCERF3 in California). The number of columns in $\boldsymbol{\Lambda}$ is equal to the number of events and inverting $\boldsymbol{\Lambda}$ to solve for $\boldsymbol{\beta}$ can therefore become computationally expensive. However, many of the events are unlikely to significantly influence seismic hazard at sites of interest and with a modest degree of approximation they can be excluded from the $\boldsymbol{\Lambda}$ matrix before performing LASSO regression. The pre-selection of events can be conducted based on seismic hazard disaggregation results (Bazzurro and Cornell, 1999), in which only events that contribute more than a certain amount (e.g., 5%) to the hazard for any intensity measure at any site are included, as demonstrated subsequently.

Since the output $\boldsymbol{\beta}$ in LASSO are regularized coefficients (minimizing the sum of weighted squared errors and penalty term $\gamma||\boldsymbol{\beta}||_1$), the rate adjustments for the selected events are not optimal values to minimize hazard misfits. Therefore, we propose an additional refit (linear regression without penalty term) to re-calculate the optimal $\boldsymbol{\beta}$ after subset events are selected. The complete event selection process can now be summarized as follows:

i. Run a traditional PSHA to calculate the total hazard $\boldsymbol{\lambda}$ at each site for each intensity measure;

ii. Use a pre-selected set of events that contribute significantly to the hazard based on disaggregation results to develop $\boldsymbol{\Lambda}$;

iii. Run constrained ($\boldsymbol{\beta} \geq \mathbf{0}$) weighted LASSO regression specified by Eq. (3.5) for a specified γ to obtain a regularized coefficient vector, which is denoted $\widehat{\boldsymbol{\beta}_{\boldsymbol{L}}}$;

iv. After selecting the event subset in *iii*, re-optimize Eq. (3.5) without $\gamma||\boldsymbol{\beta}||_1$ term (i.e., weighted least square) to obtain the updated regressed coefficient vector, $\widehat{\boldsymbol{\beta}_{\boldsymbol{R}}}$. This regression is subject to the constraint that each element of $\widehat{\boldsymbol{\beta}_{\boldsymbol{R}}}$ is positive;

v. Repeat steps *iii* – *iv* for different γ values until the number of selected events equals a desired target.

3.4 Matrix Representation of Magnitude and Distance Hazard

The events selected using the LASSO regression procedure presented above can match hazard curves for multiple sites and multiple intensity measures, but the relative contributions of events with different magnitudes and source-to-site distances to the hazard are not likely to be preserved. Preserving magnitude and distance distributions may be necessary for some applications. For example, the triggering of soil liquefaction depends not only on shaking intensity but also on magnitude because longer-duration ground motions at a given shaking intensity are more likely to induce liquefaction. To incorporate magnitude and distance distributions into the LASSO framework, we must formulate their contributions to total hazard as a vector and contributions from considered events as a matrix that can then be included in the regression equations. In the case of the events matrix, since a given event has a particular magnitude and a particular source-to-site distance, the event contributions will be calculated differently from total hazard contributions (which is explained below). For notational simplicity, we will express them as tensors first and then apply the rank reduction method to transform them into a vector and a matrix.

The target distributions of magnitude and distance are derived from disaggregation results, as illustrated by Figure 3.3(a). The heights of the blue bars indicate the relative contributions to a specified hazard level (i.e., intensity measure type and exceedance rate or return period) for binned values of magnitude and distance. The joint distribution of magnitude and distance for a given site, intensity measure, and return period is a rank-2 tensor ${}^{2}P_{b,d}$, where b and d are the magnitude and distance bin indices, respectively. Its dimension is N_M by N_R, where N_M and N_R represent the number of considered magnitude and distance bins, respectively.

Alternatively, we can use marginal distributions as the target distributions, as illustrated in Figure 3.3(b) for magnitude. The benefit of using marginal distributions is that the target distribution size is reduced from $(N_M \times N_R)$ to $(N_M + N_R)$, thereby reducing computational demand. For example, suppose we wish to select a subset of events from $N_E = 1000$ events, while preserving hazard at 200 sites for 7 intensity measure levels and 7 intensity measure types. Suppose the joint distribution with 7 magnitude bins and 7 distance bins is also to be preserved. In that case, our experience suggests that we will require at least 30 GB of memory to load matrices for LASSO regression (exceeding the capacity of a typical personal computer). However, if the marginal distribution is used, we only require about 8 GB of memory, and the computational time would be reduced by approximately a factor of ten. The drawback of using marginal distributions is that the joint distribution of magnitude and distance may not be preserved accurately.

Marginal magnitude and distance distributions are calculated by summing the joint magnitude and distance disaggregation bars, shown in yellow and red columns in Figure 3.3(a). The marginal magnitude distribution is taken as an example and replotted in Figure 3.3 (b), in which the magnitude bin is denoted as m_b and bar heights are denoted as $P(m_b)$, which represents the relative contribution to hazard from magnitude bin, m_b. We can also calculate the cumulative sum, $F(m_b) = \sum_{z=1}^{b} P(m_z \geq m_b)$ where z is the running index and plot $F(m_b)$ as in Figure 3.3(c). The cumulative sum $F(m_b)$ provides a form that is consistent with the hazard curves, which are also cumulative distribution functions. Similarly, the marginal distance distribution $P(r_d)$ (where r_d is the d -th distance bin) and cumulative sum of marginal distance distribution $F(r_d) = \sum_{z=1}^{d} P(r_z \geq r_d)$ can also be calculated. We adopt the notation ${}^{1}P_b$ and ${}^{1}P_d$ (rank-1 tensors) to represent the marginal magnitude and distance distributions and ${}^{1}F_b$ and ${}^{1}F_d$ (rank-1 tensors) to represent the marginal cumulative magnitude and distance distributions.

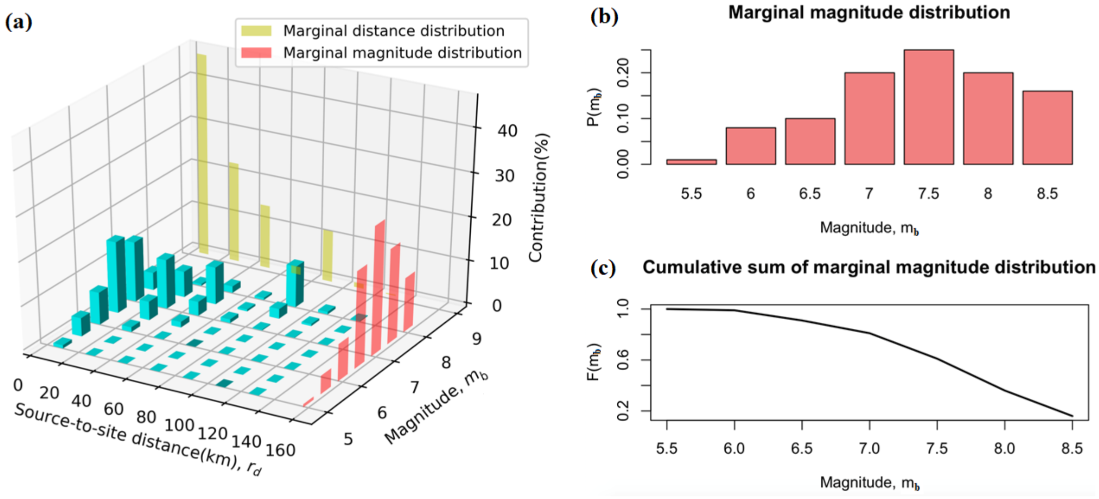


Figure 3.3. A schematic plot of (a) disaggregation of the seismic hazard by magnitude and distance, (b) marginal magnitude distribution, and (c) the corresponding cumulative sum of marginal magnitude distribution.

The joint distribution ${}^{2}P_{b,d}$ and marginal distributions ${}^{1}P_b$ and ${}^{1}P_d$ (or ${}^{1}F_b$ and ${}^{1}F_d$) are derived from disaggregation results for a particular site, *IM* type, and *IM* level (equivalently, return period). If a regional multi-*IM*s hazard is analyzed, the calculation must be conducted repeatedly for all sites, *IM* types, and *IM* levels. A rank-5 tensor ${}^{5}P_{b,d,l,j,k}$ is used to represent the relative hazard contribution at *IM* level l for *IM* type k at site j from the magnitude b and distance d bin. We can use two rank-4 tensors ${}^{4}P_{b,l,j,k}$ and ${}^{4}P_{d,l,j,k}$ (or ${}^{4}F_{b,l,j,k}$ and ${}^{4}F_{d,l,j,k}$) to represent the marginal magnitude and distance distributions (or marginal cumulative sum distributions) for the magnitude b bin and the distance d bin, respectively, at *IM* level l for *IM* type k at site j. These distributions then must be multiplied by the corresponding hazard $\lambda_{l,j,k}$ (annual exceedance rate) to obtain the absolute hazard contribution distributions before incorporating them into LASSO regression. The reason for the multiplication is that we need to assign equal weights to magnitude and distance distributions as well as the ground motion hazards when optimizing the LASSO objective function. Taking the marginal magnitude and distance hazard distributions as an example, the calculations are,

$$ {}^{4}\lambda_{b,l,j,k} = {}^{4}P_{b,l,j,k} \circ \lambda_{l,j,k} \tag{3.6}$$

$$ {}^{4}\lambda_{d,l,j,k} = {}^{4}P_{d,l,j,k} \circ \lambda_{l,j,k} \tag{3.7}$$

where ∘ represents Hadamard product (also known as the element-wise product). To incorporate Eqs. (3.6) and (3.7) into the LASSO framework, their ranks must be reduced to a rank-1 tensor or a vector. We use $\boldsymbol{\lambda_M}$ and $\boldsymbol{\lambda_R}$ to denote the transformed marginal magnitude and distance hazard vectors. The same procedure of rank reduction transformation illustrated in Figure 3.1 can be applied here. The index q in Eq. (3.3) can now be updated for the marginal magnitude hazard distribution ${}^{4}\lambda_{b,l,j,k}$ transformation as,

$$q = b + (l-1) \times N_M + (j-1) \times N_M \times N_X + (k-1) \times N_M \times N_X \times N_S \tag{3.8}$$

The same calculation can be applied to the index q for the marginal distance hazard distribution.

Similar to the event hazard matrix $\boldsymbol{\Lambda}$ defined in Eq. 3.4 (composed of elements ${}^{4}\Lambda_{l,j,k\,i}$ in Eq. 2.5), we also need to develop the hazard distribution matrix for each event. A particular event i has a specified magnitude, thus, the hazard produced by the event only contributes to the magnitude bin that includes the event magnitude and the hazard contribution for other magnitude bins should be zero. Equivalently, for a particular event i and site j, the source-to-site distance is fixed and the hazard contribution from the event occurs only in the distance bin that includes that distance. Consequently, the following equations provide the marginal magnitude and distance distributions for event i,

$$ {}^{5}\Lambda_{b,l,j,k,i} = \begin{cases} \Lambda_{l,j,k,i}, & m_i \in (m_b - \frac{\Delta m}{2}, m_b + \frac{\Delta m}{2}) \\ 0, & \text{otherwise} \end{cases} \tag{3.9}$$

$$ {}^{5}\Lambda_{d,l,j,k,i} = \begin{cases} \Lambda_{l,j,k,i}, & r_{i,j} \in (r_d - \frac{\Delta r}{2}, r_d + \frac{\Delta r}{2}) \\ 0, & \text{otherwise} \end{cases} \tag{3.10}$$

where m_i is the magnitude of event i, $r_{i,j}$ is source-to-site distance for event i and site j, Δm and Δr are the bin widths for magnitude and distance. These hazard distributions then need to be transformed into a rank-2 tensor by the procedure illustrated in Figure 3.2. The corresponding index q is calculated by Eq. (3.8). We denote the transformed event hazard distribution matrix for magnitude and distance as $\boldsymbol{\Lambda_M}$ and $\boldsymbol{\Lambda_R}$.

To preserve the magnitude and distance hazard contribution in the event selection for regional multi-*IM* hazard analysis, we can expand $\boldsymbol{\lambda}$ and $\boldsymbol{\Lambda}$ by including the magnitude and distance hazard distributions as,

$$\boldsymbol{\lambda}' = \begin{bmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\lambda_M} \\ \boldsymbol{\lambda_R} \end{bmatrix},\ \boldsymbol{\Lambda}' = \begin{bmatrix} \boldsymbol{\Lambda} \\ \boldsymbol{\Lambda_M} \\ \boldsymbol{\Lambda_R} \end{bmatrix} \tag{3.11}$$

If these expanded $\boldsymbol{\lambda}'$ and $\boldsymbol{\Lambda}'$ are substituted for $\boldsymbol{\lambda}$ and $\boldsymbol{\Lambda}$ in equations 3.4 and 3.5, the selected events will, in a least squares sense, match the hazard curves and simultaneously preserve magnitude and distance hazard distributions.

4 Methodology for Scenario Ground Motion Generation and Selection

This chapter first describes the generation of correlated scenario ground motions using multivariate normal distribution randomization. Then it describes the tensor and matrix representation of hazard produced by each ground motion realization and implements LASSO regression for selection of a subset of hazard-consistent ground motion realizations.

4.1 Correlated Scenario Ground Motion Generation

Once the hazard-consistent scenario event subset is selected (Chapter 3), we will need to generate correlated ground motion realizations for the region for each event. We use a rank-3 tensor, ${}^{3}Z^{GM}_{j,i,g}$, to represent the simulated ground motion at site j from event i in realization g for a single intensity measure (e.g., PGA). A vector representation of a single realization of ground motions from a particular scenario event (i.e., given i and g) is provided with one entry per site, which is expressed as follows (note the i and g indices are dropped for simplicity),

$$\tilde{\boldsymbol{z}} = \widehat{\boldsymbol{\mu}} + \widetilde{\boldsymbol{\eta}} + \widetilde{\boldsymbol{\delta W}} \tag{4.1}$$

where

$$\tilde{\boldsymbol{z}} = \begin{bmatrix} \ln\,(\tilde{z}_1) \\ \ln\,(\tilde{z}_2) \\ \vdots \\ \ln\,(\tilde{z}_J) \end{bmatrix} \tag{4.2}$$

is the vector of generated ground motions in natural log units for J sites (the tilde symbols indicate the values are simulated),

$$\widehat{\boldsymbol{\mu}} = \begin{bmatrix} \ln(\hat{\mu}_1) \\ \ln(\hat{\mu}_2) \\ \vdots \\ \ln(\hat{\mu}_J) \end{bmatrix} \tag{4.3}$$

is the vector of predicted natural log median ground motions from a GMM (e.g., Boore et al., 2014) (the hat symbols indicate the values are estimated),

$$\widetilde{\boldsymbol{\eta}} = \tilde{\eta} \times \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \tag{4.4}$$

is the vector of simulated event terms or event bias (note that there is a single event bias for one ground motion realization), and

$$\widetilde{\boldsymbol{\delta W}} = \begin{bmatrix} \widetilde{\delta W}_1 \\ \widetilde{\delta W}_2 \\ \vdots \\ \widetilde{\delta W}_J \end{bmatrix} \tag{4.5}$$

is the vector of simulated within event residuals, which are different for each site. Typically, we assume that event term follows a univariate normal distribution,

$$\tilde{\eta} \sim N(0, \hat{\tau}^2) \tag{4.6}$$

where $\hat{\tau}$ is the between-event standard deviation, which is provided in GMMs. Within-event residuals are spatially correlated (Jayaram and Baker, 2009) and follow a multivariate normal distribution,

$$\widetilde{\boldsymbol{\delta W}} \sim N(\mathbf{0}, \boldsymbol{\Sigma}) = N\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \hat{\phi}_1{}^2 & \hat{\phi}_1\hat{\phi}_2\rho_{1,2} & \cdots & \hat{\phi}_1\hat{\phi}_J\rho_{1,J} \\ \hat{\phi}_2\hat{\phi}_1\rho_{2,1} & \hat{\phi}_2{}^2 & \cdots & \hat{\phi}_2\hat{\phi}_J\rho_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\phi}_J\hat{\phi}_1\rho_{J,1} & \hat{\phi}_J\hat{\phi}_2\rho_{J,2} & \cdots & \hat{\phi}_J{}^2 \end{bmatrix} \right) \tag{4.7}$$

where $\hat{\phi}_j$ is the standard deviation of within event residuals for site j (also provided by a GMM) and $\rho_{j,j'}$ is the correlation of within event residuals between sites j and j'. The standard deviations of event term and within event residual are usually modeled in a GMM, while the spatial correlation $\rho_{j,j'}$ is not. Models for spatial correlation have been developed by applying geostatistical tools such as semivariograms to relatively small subsets of the ground motions considered in GMM development (e.g., Jayaram and Baker, 2009; Loth and Baker, 2013).

Once the standard deviations and correlations are estimated by a GMM and a correlation model, we can establish the complete distributions for $\tilde{\eta}$ and $\widetilde{\boldsymbol{\delta W}}$ and simulate realizations for event terms and within event residuals by multivariate normal distribution randomization. By performing the summation in Eq. (4.1), we generate the logarithmic ground motion realizations. We then repeat this routine for different events i and different realizations g to develop ${}^{3}Z^{GM}_{j,i,g}$.

Equations (4.1)-(4.7) describes ground motion generation when a single *IM* is considered, whereas in many applications we need to generate ground motion realizations in which multiple *IM*s are considered and their correlations (both spatially for each *IM* and between-*IM*s) are modelled. For example, in this project, we need to provide correlated ground motion realizations of PGA and PGV for subsequent seismic geo-hazard analyses. A rank-4 tensor, ${}^{4}Z^{GM}_{k,j,i,g}$, is used to represent the simulated ground motion at site j, *IM* k, event i, and ground motion realization g. Due to the correlation between different *IM*s, the multi-*IM*s simulations are not independent. A correlated realization of multi-*IM*s for a given event i and for realization g is expressed as a rank-2 tensor ${}^{2}Z^{GM}_{k,j}$. We can apply the tensor transformation approach described in Ch. 3 to convert the rank-2 tensor to a long vector (stacking first by *IM* and then by site) to simplify the description. For example, if we consider two *IM*s, k_1 and k_2, Eq. (4.1) is updated as,

$$\begin{bmatrix} \tilde{\boldsymbol{z}}_{\boldsymbol{k_1}} \\ \tilde{\boldsymbol{z}}_{\boldsymbol{k_2}} \end{bmatrix} = \begin{bmatrix} \widehat{\boldsymbol{\mu}}_{\boldsymbol{k_1}} \\ \widehat{\boldsymbol{\mu}}_{\boldsymbol{2}} \end{bmatrix} + \begin{bmatrix} \widetilde{\boldsymbol{\eta}}_{\boldsymbol{k_1}} \\ \widetilde{\boldsymbol{\eta}}_{\boldsymbol{k_2}} \end{bmatrix} + \begin{bmatrix} \widetilde{\boldsymbol{\delta W}}_{\boldsymbol{k_1}} \\ \widetilde{\boldsymbol{\delta W}}_{\boldsymbol{k_2}} \end{bmatrix} \tag{4.8}$$

where

$$\begin{bmatrix} \tilde{\boldsymbol{z}}_{\boldsymbol{k_1}} \\ \tilde{\boldsymbol{z}}_{\boldsymbol{k_2}} \end{bmatrix} = \begin{bmatrix} \ln\,(\tilde{z}_{k_1,1}) \\ \ln\,(\tilde{z}_{k_1,2}) \\ \vdots \\ \ln\,(\tilde{z}_{k_1,J}) \\ \ln\,(\tilde{z}_{k_2,1}) \\ \ln\,(\tilde{z}_{k_2,2}) \\ \vdots \\ \ln\,(\tilde{z}_{k_2,J}) \end{bmatrix} \tag{4.9}$$

is the vector of simulated logarithmic ground motions for the region with J sites for *IM* k_1 (the first J rows in the vector) and *IM* k_2 (the second J rows in the vector),

$$\begin{bmatrix} \widehat{\boldsymbol{\mu}}_{\boldsymbol{k_1}} \\ \widehat{\boldsymbol{\mu}}_{\boldsymbol{k_2}} \end{bmatrix} = \begin{bmatrix} \ln(\hat{\mu}_{k_1,1}) \\ \ln(\hat{\mu}_{k_1,2}) \\ \vdots \\ \ln(\hat{\mu}_{k_1,J}) \\ \ln(\hat{\mu}_{k_2,1}) \\ \ln(\hat{\mu}_{k_2,2}) \\ \vdots \\ \ln(\hat{\mu}_{k_2,J}) \end{bmatrix} \tag{4.10}$$

is the vector of predicted median logarithmic ground motions from GMMs for *IM*s k_1 and k_2,

$$\begin{bmatrix} \widetilde{\boldsymbol{\eta}}_{\boldsymbol{k_1}} \\ \widetilde{\boldsymbol{\eta}}_{\boldsymbol{k_2}} \end{bmatrix} = \begin{bmatrix} \tilde{\eta}_{k_1} \\ \tilde{\eta}_{k_1} \\ \vdots \\ \tilde{\eta}_{k_1} \\ \tilde{\eta}_{k_2} \\ \tilde{\eta}_{k_2} \\ \vdots \\ \tilde{\eta}_{k_2} \end{bmatrix} \tag{4.11}$$

is the vector of simulated event terms for *IM*s k_1 and k_2, and

$$\begin{bmatrix} \widetilde{\boldsymbol{\delta W}}_{\boldsymbol{k_1}} \\ \widetilde{\boldsymbol{\delta W}}_{\boldsymbol{k_2}} \end{bmatrix} = \begin{bmatrix} \widetilde{\delta W}_{k_1,1} \\ \widetilde{\delta W}_{k_1,2} \\ \vdots \\ \widetilde{\delta W}_{k_1,J} \\ \widetilde{\delta W}_{k_2,1} \\ \widetilde{\delta W}_{k_2,2} \\ \vdots \\ \widetilde{\delta W}_{k_2,J} \end{bmatrix} \tag{4.12}$$

is the vector of simulated within event residuals for *IM*s k_1 and k_2. The event terms for *IM*s k_1 and k_2 are described by a bivariate normal distribution,

$$\begin{bmatrix} \tilde{\eta}_{k_1} \\ \tilde{\eta}_{k_2} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \hat{\tau}_{k_1}{}^2 & 0 \\ 0 & \hat{\tau}_{k_2}{}^2 \end{bmatrix} \right) \tag{4.13}$$

where the diagonal elements are the corresponding variances of event terms from GMMs for *IM*s k_1 and k_2, and the off-diagonal element is the covariance of event terms, which is usually assumed to be zero. For within event residuals of J sites for k_1 and k_2, the multivariate normal distribution is

$$\begin{bmatrix}\widetilde{\boldsymbol{\delta W}}_{\boldsymbol{k_1}}\\ \widetilde{\boldsymbol{\delta W}}_{\boldsymbol{k_2}}\end{bmatrix} \sim N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{k_1}} & \boldsymbol{\Sigma}_{\boldsymbol{k_1,k_2}}\\ \boldsymbol{\Sigma}_{\boldsymbol{k_2,k_1}} & \boldsymbol{\Sigma}_{\boldsymbol{k_2}}\end{bmatrix}\right) \tag{4.14}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{k_1}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{k_2}}$ are the covariance matrices among J sites for *IM*s k_1 and k_2, respectively, in the same manner as the $\boldsymbol{\Sigma}$ in Eq. (4.7). The off-diagonal matrix $\boldsymbol{\Sigma}_{\boldsymbol{k_1,k_2}}$ (or $\boldsymbol{\Sigma}_{\boldsymbol{k_2,k_1}}$) is the covariance matrix considering spatial correlation and cross-*IM*s correlation, which is expressed as,

$$\boldsymbol{\Sigma}_{\boldsymbol{k_1,k_2}} = \begin{bmatrix} \widehat{\phi}_{k_1,1}\widehat{\phi}_{k_2,1}\rho_{k_1-1,k_2-1} & \widehat{\phi}_{k_1,1}\widehat{\phi}_{k_2,2}\rho_{k_1-1,k_2-2} & \cdots & \widehat{\phi}_{k_1,1}\widehat{\phi}_{k_2,J}\rho_{k_1-1,k_2-J}\\ \widehat{\phi}_{k_1,2}\widehat{\phi}_{k_2,1}\rho_{k_1-2,k_2-1} & \widehat{\phi}_{k_1,2}\widehat{\phi}_{k_2,2}\rho_{k_1-2,k_2-2} & \cdots & \widehat{\phi}_{k_1,2}\widehat{\phi}_{k_2,J}\rho_{k_1-2,k_2-J}\\ \vdots & \vdots & \ddots & \vdots\\ \widehat{\phi}_{k_1,J}\widehat{\phi}_{k_2,1}\rho_{k_1-J,k_2-1} & \widehat{\phi}_{k_1,J}\widehat{\phi}_{k_2,2}\rho_{k_1-J,k_2-2} & \cdots & \widehat{\phi}_{k_1,J}\widehat{\phi}_{k_2,J}\rho_{k_1-J,k_2-J} \end{bmatrix} \tag{4.15}$$

where $\widehat{\phi}_{k_1,j}$ and $\widehat{\phi}_{k_2,j}$ are the standard deviations of within event residuals at site j for *IM*s k_1 and k_2, and $\rho_{k_1-j,k_2-j'}$ is the correlation between within event residuals at site j for *IM* k_1 and site j' for *IM* k_2. Loth and Baker (2013) provide a correlation model for this cross-*IM* spatial correlation. Once these distributions are established, we can implement multivariate normal distribution randomization to generate a correlated ground motion realization for PGA and PGV. By repeating the generation for different events i and different realizations g, we establish the complete rank-4 tensor of ground motion realizations ${}^{4}Z^{GM}_{k,j,i,g}$.

4.2 Hazard-Consistent Ground Motion Maps Selection

The procedure described in Section 4.1 can be applied to generate as many correlated ground motion realizations as needed from each of the selected hazard-consistent events from Chapter 3. For seismic risk analysis of spatially distributed infrastructure systems, analysis run times may dictate that a reduced set of ground motion realizations need to be generated. To meet this need, we adapt the procedure of Han and Davidson (2012), which begins by generating a large number of correlated ground motion realizations from each selected event to ensure that extreme ground motion (at the tails of ground motion distributions) are captured. For the present analyses, we generated $n_G = 50$ ground motion realizations per event. Next we combine all ground motion realizations to develop a hazard matrix $\boldsymbol{\Lambda^{GM}}$ in the same manner as hazard matrix $\boldsymbol{\Lambda}$ for the considered events in Chapter 3 (Eq. 3.4). We then implement LASSO regression to select ground motion realizations and estimate their hazard-consistent annual occurrence rates. The development of $\boldsymbol{\Lambda^{GM}}$ includes two steps, converting ground motion maps ${}^{4}Z^{GM}_{k,j,i,g}$ to hazard curves ${}^{5}\Lambda^{GM}_{l,k,j,i,g}$ and conducting tensor reduction to transform ${}^{5}\Lambda^{GM}_{l,k,j,i,g}$ to $\boldsymbol{\Lambda^{GM}}$.

A hazard curve for *IM* type k at site j is a series of annual exceedance rates at different *IM* levels l, as shown by the data points along the hazard curve in Figure 4.1. For a particular event i with the corrected annual occurrence rate $v_i{}'$ ($v_i{}' = \beta_i \times v_i$ from Chapter 3), suppose n_g ground motion realizations are generated. The annual occurrence rate for each realization g is $\alpha_{i,g} = v_i{}' \div n_g$ (each map is generated with equal likelihood). To illustrate the conversion from a ground motion to a hazard curve, we take the annual occurrence rate of a simulated ground motion realization g generated for event i as $\alpha_{i,g} = 0.001$, and the simulated ground motion for *IM* type k at site j is $\exp(\ {}^{4}Z^{GM}_{k,j,i,g}) = 0.01$, then the associated exceedance rate of this ground motion is expressed as

$$ {}^{5}\Lambda^{GM}_{l,k,j,i,g} = \begin{cases} 0.001, & IM \text{ level } l < 0.01 \\ 0, & \text{otherwise} \end{cases} \tag{4.16} $$

which is shown by the red curve in Figure 4.1.

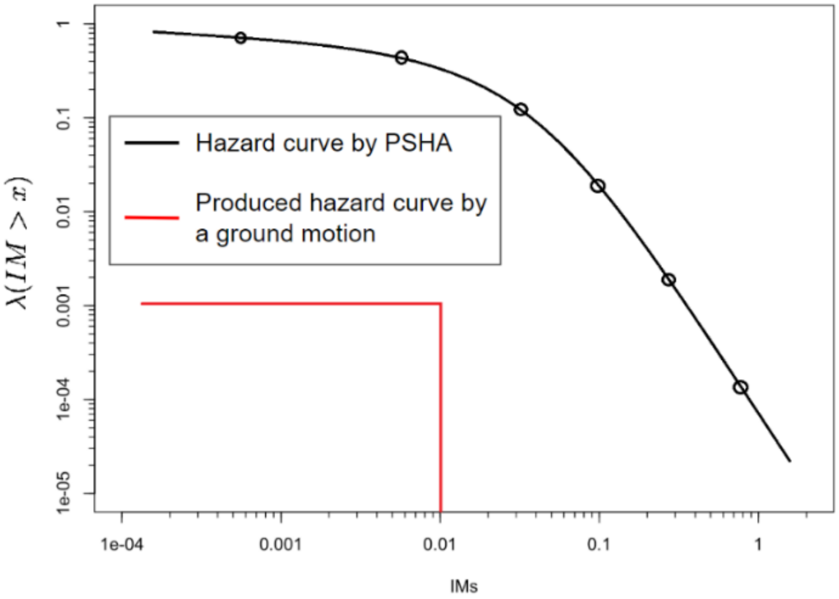


Figure 4.1. A schematic plot of the hazard produced a ground motion.

This conversion can be repeated for all *IM* type k at site j from event i for ground motion realization g to obtain the complete rank-5 tensor ${}^{5}\Lambda^{GM}_{l,k,j,i,g}$ for all realizations.

To incorporate ${}^{5}\Lambda^{GM}_{l,k,j,i,g}$ into LASSO to conduct the selection process, we transform the rank-5 tensor into a matrix $\boldsymbol{\Lambda^{GM}}$ and replace $\boldsymbol{\Lambda}$ with $\boldsymbol{\Lambda^{GM}}$ in Eqs. (3.4) and (3.5). In $\boldsymbol{\Lambda}$, we need to make sure the index of row q is consistent with the target hazard $\boldsymbol{\lambda}$ and each column is associated with each considered event. Then for $\boldsymbol{\Lambda^{GM}}$, we keep the same row order as defined by Eq. (3.3) and each column is associated with a ground motion realization. For row index q in $\boldsymbol{\Lambda}^{\boldsymbol{GM}}$, we apply Eq. (3.3). Column index p combines indices for event i and realization g, such that

$$p = g + (i - 1) \times {N_E}' \tag{4.17}$$

where ${N_E}'$ is the number of selected events.

5 Hazard-Consistent Scenario Results

Point-based PHSA was conducted by Al Atik et al. (2022) at 19,316 sites (with a grid spacing of 0.05 by 0.05 degrees in longitude and latitude) in California, locations of which are shown in Figure 5.1. For subsequent geo-hazard analyses (Stewart et al. 2023), we require hazard-consistent correlated PGA and PGV ground motion realizations. Incorporating all 19,316 sites into LASSO would require more than 1TB of memory, which is practically impossible given the currently available computation resources. Since this project aims to assess seismic risk for California natural gas pipelines (red lines in Figure 5.1), we only enforce hazard-consistency for sites close to the gas pipelines and allow hazard discrepancies for other sites. We drew a buffer with a 1 km width along the natural gas pipelines and chose the sites within the buffer as the target sites, which led to 1,220 sites shown in Figure 5.2. These sites were overall evenly distributed along all pipelines with a few small segments missing.

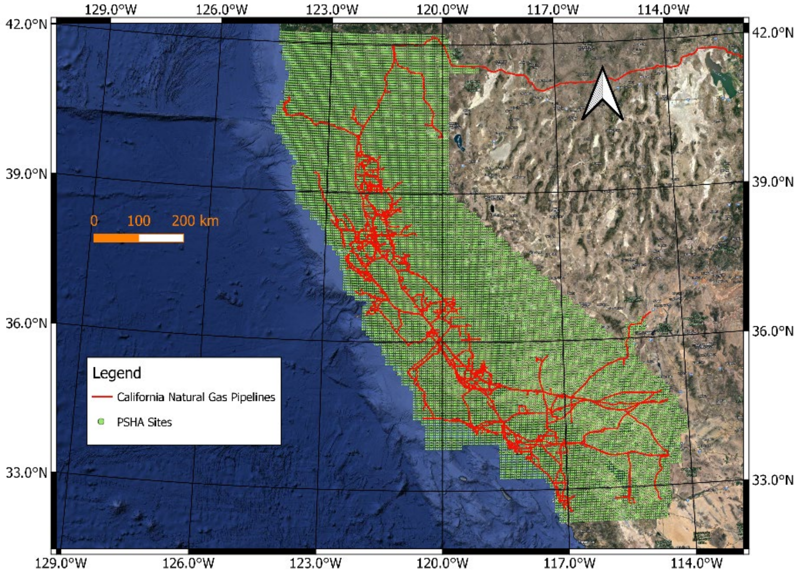


Figure 5.1. A map of all PSHA sites considered by Al Atik et al. (2023).

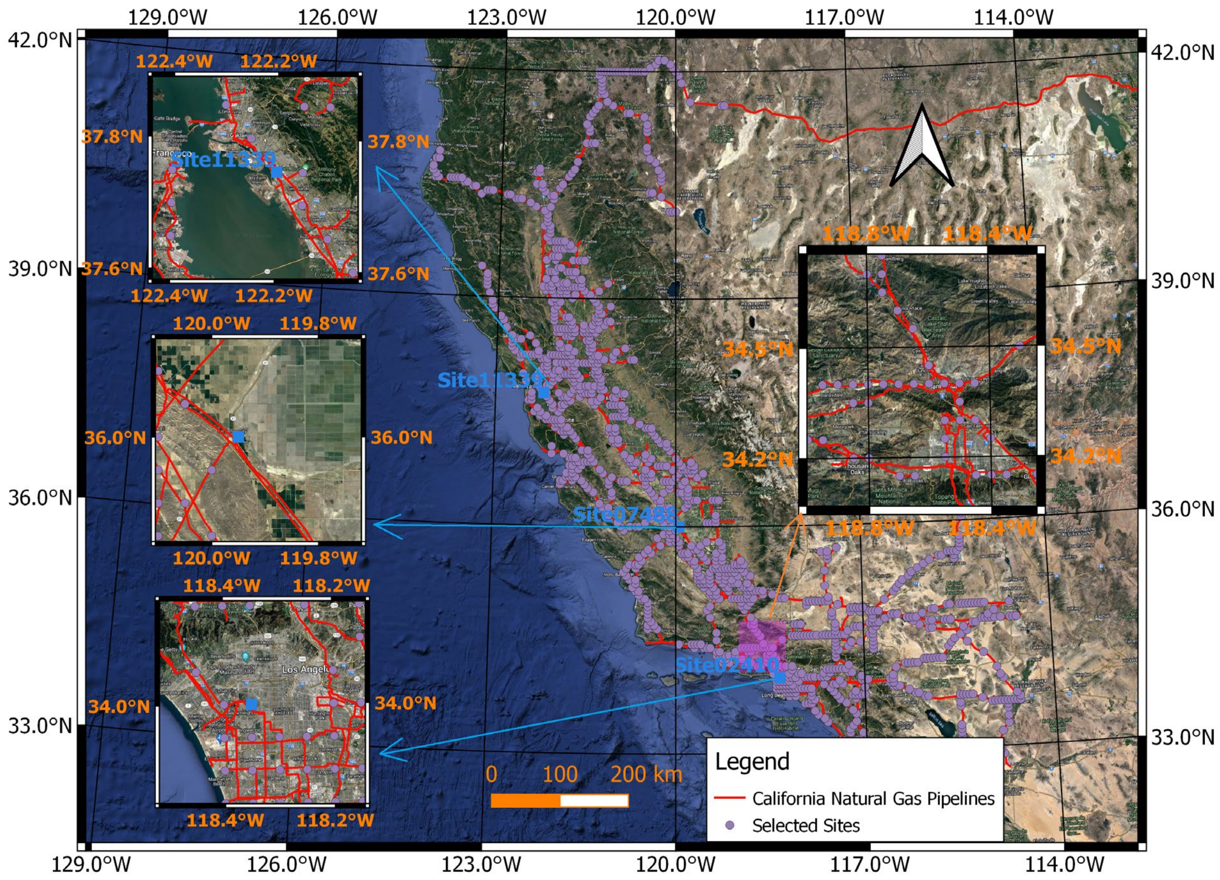


Figure 5.2. A map of selected PSHA sites for selection of hazard-consistent ground motion realizations.

The results provided by Al Atik et al. (2022) include hazard curves for 20 *IM* levels for the *IM* types of PGA and PGV for the reference site condition of V_{S30} (time-averaged shear wave velocities in the upper 30 m of the site) = 760 m/s. We utilize these results for the 1,220 selected sites along with disaggregations at each site for both *IM* types and all considered *IM* levels (the disaggregations are not at specified exceedance rates). Because the analysis framework described in Chapters 3 and 4 takes the ground motions for a given return period range (i.e., 200 years to 2,475 years) as input, we convert the *IM* level range provided from the hazard analysis to a standardized return period range (note the number of selected *IM* levels for the same return period range differs at different sites due to different hazard curves). The corresponding hazard curve segments and disaggregations are then used to develop target hazard vector $\boldsymbol{\lambda}$ and hazard matrix from considered events $\boldsymbol{\Lambda}$ or considered ground motion realizations $\boldsymbol{\Lambda}^{\boldsymbol{GM}}$.

5.1 Hazard-Consistent Scenario Events

Al Atik et al. (2022) considered more than 400,000 fault rupture events and over 2,500 grid point sources for background seismicity from two branches in UCERF3 (Field et al. 2015) (i.e., $N_E > 400{,}000$). As a result, it would be cumbersome to use all events to develop the corresponding event hazard matrix $\mathbf{\Lambda}$, which we instead developed using a subset of important representative events based on disaggregation results. Figure 3.3(a) shows a typical disaggregation result for one *IM* type at one *IM* level (or return period) at one site. The disaggregation reveals that only a small subset of magnitude-distance bins are important (i.e., many events do not contribute significantly to hazard). Moreover, while many events within the UCERF3 model may contribute within a single magnitude-distance bin, they would provide similar ground motion distributions. Therefore, we pre-select a subset of important representative events by taking only one event from each magnitude-distance bin that has a relative contribution larger than a threshold value. This process is repeated for all sites, all *IM* types, and all *IM* levels. The selection of an appropriate threshold relative contribution level is an important consideration. Higher thresholds decrease computation time and decrease the number of pre-selected events (fewer magnitude-distance bins are included), whereas lower thresholds will result in more pre-selected events and better hazard matches to develop $\mathbf{\Lambda}$. We found a threshold of 10% for multi-fault rupture events and 5% for gridded point sources balances goodness-of-match and efficiency for this study region. The outcome of this process for the present study region with 1,220 sites for two *IM* types and multiple *IM* levels for a return period from 200 years to 2,475 years is that 7,700 events are pre-selected (including fault ruptures and point sources). We used these pre-selected events to establish $\mathbf{\Lambda}$ (Eq. 3.4). The calculation of $\mathbf{\Lambda}$ from the pre-selected events has been implemented in an R package, *RPSHA* (Wang, 2022). More specifically, the calculations were conducted using the functions *event_haz_calc* and *events_hazmat_calc*. (Note the annual rate of occurrence of each pre-selected event, v_i, is also required to calculate the final corrected rates). The computation was conducted using High Performance Computing (HPC) infrastructure at Old Dominion University.

A case study by Wang et al. (202x) found that if magnitude hazard marginal distributions were not incorporated in event selection, their distributions could be poorly preserved. However, the marginal distance distributions were generally acceptable with and without their distributions incorporated into LASSO selection. Considering the significance of event magnitude for geo-hazards analyses and heavy computation demands, we incorporated only magnitude marginal distributions into the LASSO regressions. Thus, the target hazard vector $\boldsymbol{\lambda}'$ and hazard matrix $\boldsymbol{\Lambda}'$ are,

$$\boldsymbol{\lambda}' = \begin{bmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\lambda}_M \end{bmatrix},\ \boldsymbol{\Lambda}' = \begin{bmatrix} \mathbf{\Lambda} \\ \mathbf{\Lambda}_M \end{bmatrix} \tag{5.1}$$

The LASSO regression method was then performed following steps *iii* – *v* described in Chapter 3.3 as implemented by the function *scenario_selection* in the *RPSHA* package. The results provided by this package are a series of selected events and their corresponding rate adjustments $\widehat{\boldsymbol{\beta}_{\boldsymbol{R}}}$.

We use the mean of absolute arithmetic relative errors [in a similar form as the *MHCE* used in Han and Davidson (2012)] to quantitatively measure the goodness-of-match of hazard curves from the selected events, which is defined as

$$err_3 = \frac{1}{N_S \times N_T \times N_X \times (N_M + N_R)} (e_h + e_m + e_d) \tag{5.2}$$

where

$$e_h = \sum_{j=1}^{j=N_S} \sum_{k=1}^{k=N_T} \sum_{l=1}^{l=N_X} \left| \frac{{}^3\lambda_{l,j,k} - \sum_{i=1}^{i=N_E} {}^4\Lambda_{l,j,k,i} * \widehat{\beta_i}}{{}^3\lambda_{b,l,j,k}} \right| \tag{5.3}$$

is the sum of absolute arithmetic errors for hazard curves,

$$e_m = \sum_{j=1}^{j=N_S} \sum_{k=1}^{k=N_T} \sum_{l=1}^{l=N_X} \sum_{b=1}^{b=N_M} \left| \frac{{}^4\lambda_{b,l,j,k} - \sum_{i=1}^{i=N_E} {}^5\Lambda_{b,l,j,k,i} * \widehat{\beta_i}}{{}^4\lambda_{b,l,j,k}} \right| \tag{5.4}$$

is the sum of absolute arithmetic errors for magnitude distribution, and

$$e_d = \sum_{j=1}^{j=N_S} \sum_{k=1}^{k=N_T} \sum_{l=1}^{l=N_X} \sum_{d=1}^{d=N_R} \left| \frac{{}^4\lambda_{d,l,j,k} - \sum_{i=1}^{i=N_E} {}^5\Lambda_{d,l,j,k,i} * \widehat{\beta_i}}{{}^4\lambda_{d,l,j,k}} \right| \tag{5.5}$$

is the sum of absolute arithmetic errors for distance distribution. In this study, since we only consider hazard curves and magnitude marginal distributions, then Eq. (5.2) is simplified as,

$$err_3 = \frac{1}{N_S \times N_T \times N_X \times N_M} (e_h + e_m) \tag{5.6}$$

Two additional error metrics, err_1 and err_2, are also provided by the *RPSHA* package. Although the error values differ among the three error metrics, they show consistent trends. We then plot the err_3 in Figure 5.3 as a function of the number of events. The error decreases as the number of selected events increases and eventually converges to a small value. The errors shown in Figure 5.3 could be further reduced by lowering relative contribution thresholds when pre-selecting representative events from disaggregations. We then select a specific number of selected events based on the error decay curve. This is an admittedly subjective decision, but we select a point where the slope of the curve is relatively flat, which is 599 events. Of these 599 events, 158 are ruptures on mapped faults and 441 are grid point source events.

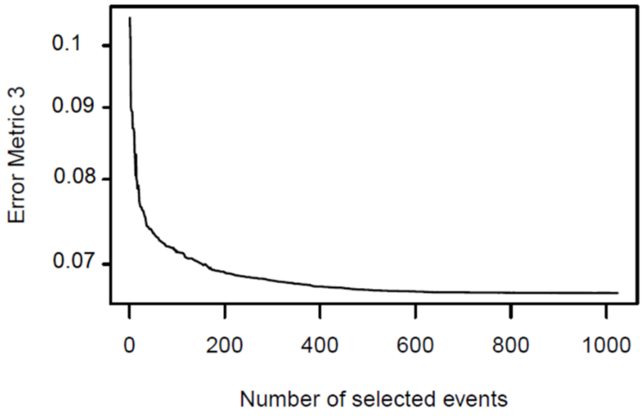


Figure 5.3. Variation of error metric err_3 versus the number of selected events.

Figure 5.4 compares PGA and PGV hazard curves from full hazard calculations to those recovered from the 599 reduced events with adjusted rates. Results are shown for three representative sites (site 11339, 07488, and 02410 in Figure 5) in the Bay Area, Central Valley, and Los Angeles. The results show that over the considered range of rates (from $\frac{1}{200} = 0.005$ to $\frac{1}{2475} = 0.0004$, which are indicated by two purple horizontal lines) the discrepancies in hazard curves are minor. The error term computed using Eq. (5.6) is $err_3 = 0.066$. Comparisons for other sites show the same trend and are not presented here for brevity. We plot the target and recovered hazard curves at all 1,220 sites on the same graph in Figure 5.5. The well-overlapping lines indicate good fitting and unbiasedness.

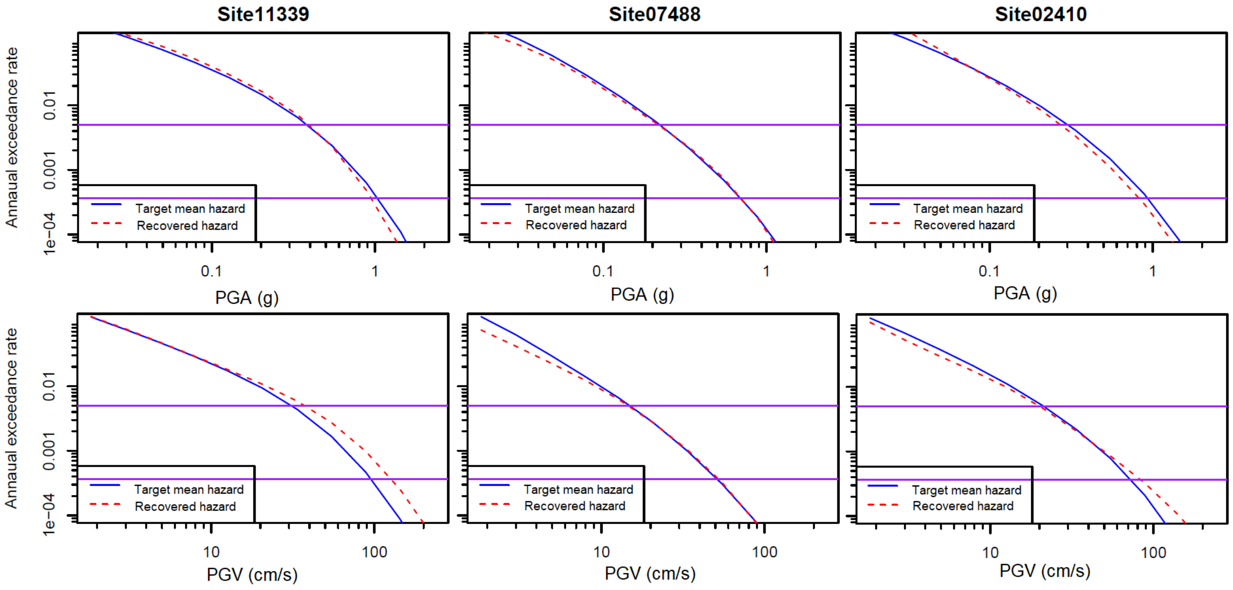


Figure 5.4. Plots of recovered hazard curves by the 599 selected events from LASSO regression (red dashed line) and the target mean hazard from full PSHA (blue solid line) at three representative sites.

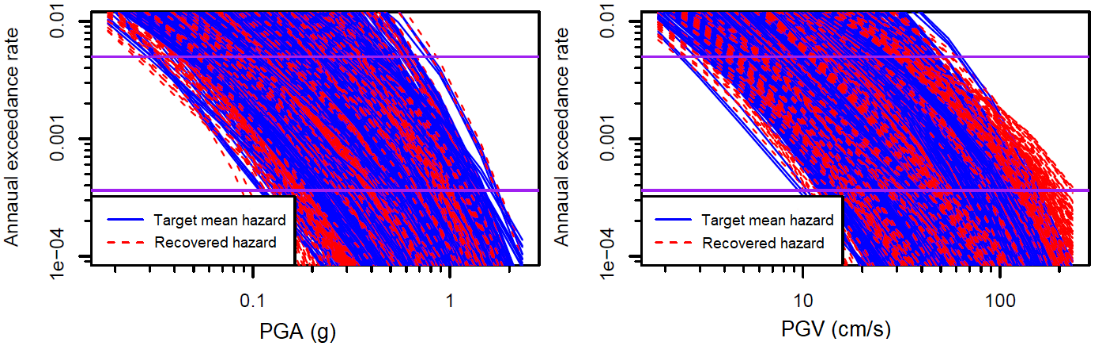


Figure 5.5. Plots of recovered hazard curves from the 599 selected events by LASSO regression (red dashed line) and the target mean hazard from full PSHA (blue solid line) for all 1,220 target sites.

Figure 5.6 shows cumulative marginal magnitude distributions for PGA and PGV at the return period of about 1,000 years at the three representative sites. The blue solid lines are the target distributions based on disaggregation from full PSHA, while the red dashed lines are the calculated recovered cumulative distributions from the reduced subset of 599 selected events. Overall, the red dashed lines match blue solid lines reasonably well for small to large magnitudes but mismatch for extremely large magnitudes (greater than 8.5). The reason is that the annual exceedance rates associated with extremely large magnitude are very small and much smaller than 0.0004 (the smallest annual exceedance rate considered in event selection). The plots for other sites and return periods show a similar trend, so we conclude that the selected 599 events are hazard-consistent and preserve the magnitude distributions.

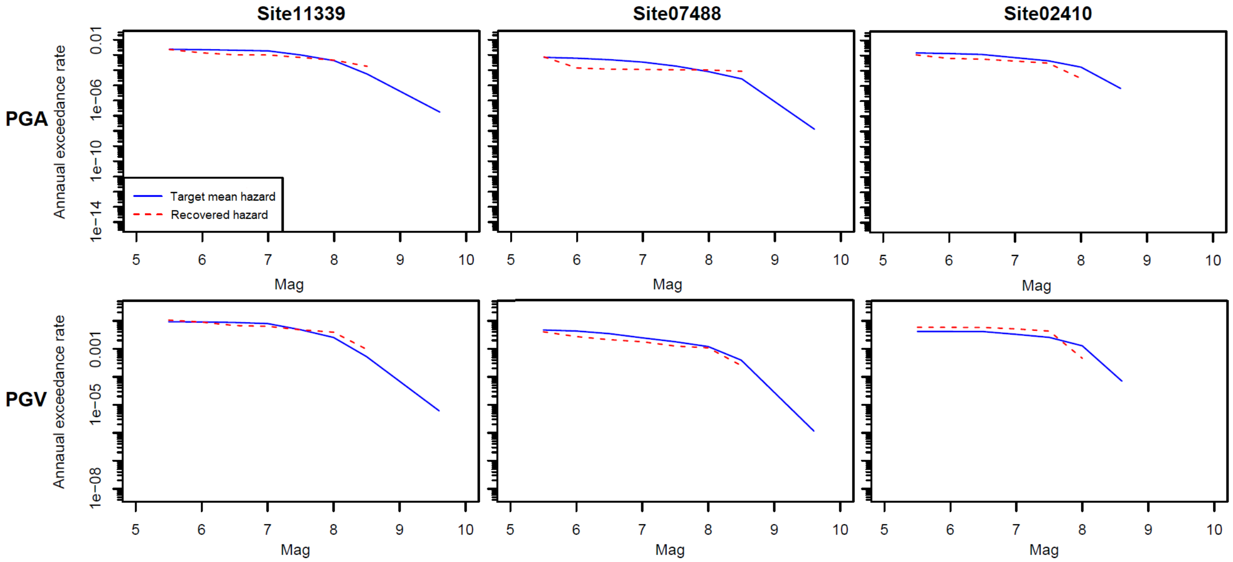


Figure 5.6. Marginal magnitude distribution plots for PGA and PGV hazards at three sites. The results apply for the 1000-year return period hazard level. LASSO regression results (red

dashed line) shown in the figure apply for 599 selected events by LASSO regression configured to match target mean hazard curves for both intensity measures and target magnitude distributions (blue solid line).

5.2 Hazard-Consistent Correlated Scenario Ground Motion Maps

Given the 599 selected hazard-consistent events, we followed the procedure in Chapter 4.1 to generate 50 correlated PGA-PGV ground motion realizations per event. We applied Eq. (4.16) to convert the generated ground motion realizations to hazard curves. Rank reduction transformation was then implemented to obtain a hazard matrix $\boldsymbol{\Lambda}^{\boldsymbol{GM}}$. Magnitude and distance distributions can be derived from ground motion realizations. However, due to heavy computation demand, we did not undertake this analysis. Instead, we only selected ground motion realizations targeted at matching hazard curves. Through this process, we use the target hazard vector $\boldsymbol{\lambda}$ and replace $\boldsymbol{\Lambda}$ with $\boldsymbol{\Lambda}^{\boldsymbol{GM}}$ in Eqs. (3.3) and (3.4) to conduct ground motion realization selection and obtain their hazard-consistent annual occurrence rates.

Since we did not consider magnitude distribution in LASSO, the err_3 for ground motion realization selection is updated as,

$$err_3 = \frac{e_h}{N_S \times N_T \times N_X} \tag{5.7}$$

where

$$e_h = \sum_{j=1}^{j=N_S} \sum_{k=1}^{k=N_T} \sum_{l=1}^{l=N_X} \left| \frac{{}^{3}\lambda_{l,j,k} - \sum_{p=1}^{p=N_G} {}^{4}\Lambda_{l,j,k,p} * \widehat{\beta_p}}{{}^{3}\lambda_{b,l,j,k}} \right| \tag{5.8}$$

is the sum of the absolute relative error of hazard curves produced by the selected ground motion maps, p is the running column (or realization) index defined by Eq. (4.17), and $N_G = N_E{}' \times n_G$ is the total number of ground motion realizations where $N_E{}'$ is the number of selected events and n_G is the number of simulated realizations per event. The plot of err_3 versus the number of selected ground motion realizations is shown in Figure 5.7. The error decreases as the number of selected ground motion realizations increases. Selection of the preferred number of selected realizations is subjective, but we recommend it be taken from the relatively flat part of the curve. Applying such criterion, about 5,000 realizations would be selected, which is much more than is practical for geo-hazards analyses. In consideration of run times and available human resources in this project, 25 maps were selected. This produces $err_3 \approx 0.8$, which is much larger than 0.066 obtained when conducing event selection. This is expected because the hazard curve produced by an event is a smooth curve, whereas the hazard curve from a ground motion realization is stepped (Figure 4.1). Many more maps would be required to smooth out hazard curves derived from ground motion realizations.

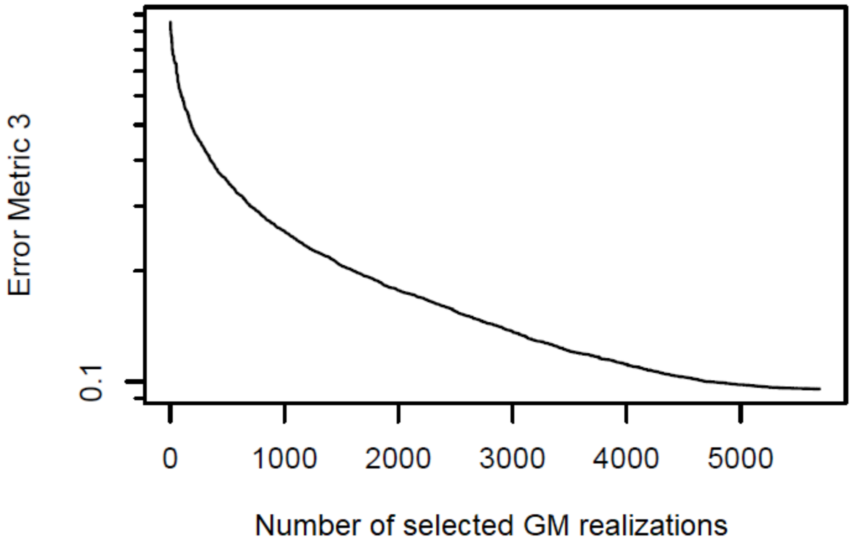


Figure 5.7. Variation of error metric err_3 versus the number of selected realizations.

Figure 5.8 compares PGA and PGV hazard curves from full hazard calculations to those recovered from the 25 selected ground motion scenarios with adjusted hazard-consistent rates for the same three representative sites. The overall mismatches are much larger than those we observed from selected events in Figure 5.4. In addition, the recovered PGA hazard curves are consistently lower than the target mean hazard curves at three sites, while PGV hazard curves show slightly better fitting. Because the maps are selected by minimizing the errors of PGA and PGV hazard curves at 1,220 sites, it is expected that for some sites, the hazard curves are underestimated, and for other sites, the hazard curves are overestimated. We plot hazard curves and the recovered hazard curves from 25 realizations at all 1,220 sites in Figure 5.9. The red dashed lines do not overlap well with the blue solid lines, indicating relatively poor fitting. However, when looking at the mean trends of all red dashed lines, they are roughly aligned with blue solid lines, which implies an overall lack of bias.

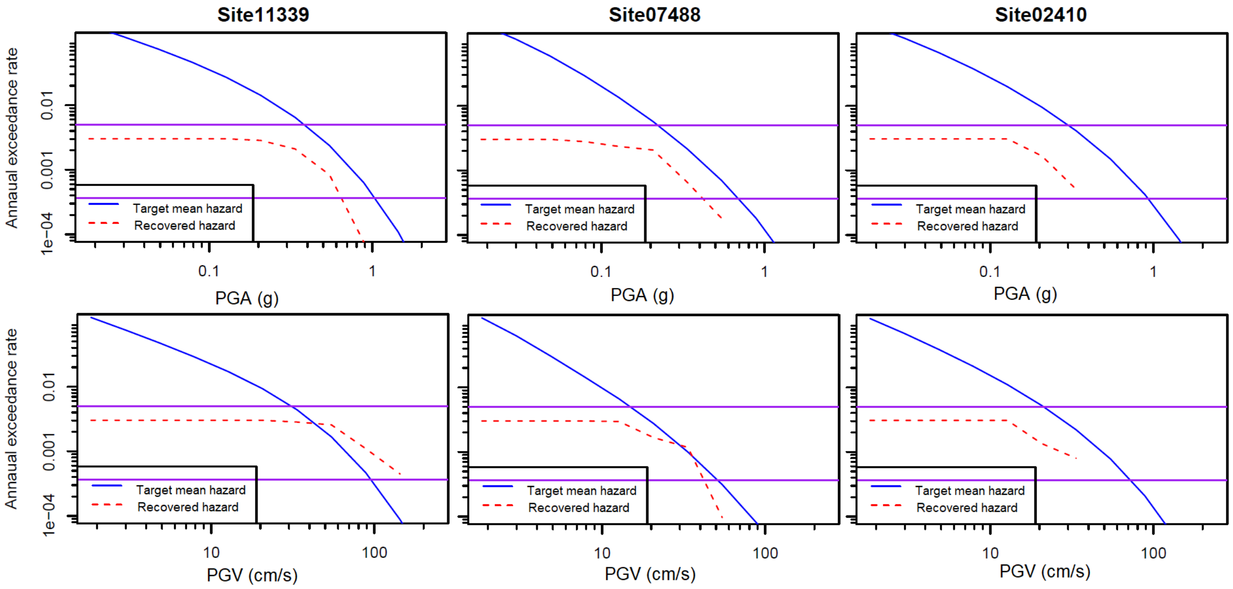


Figure 5.8. Plots of recovered hazard curves by the 25 selected ground motion maps by the LASSO regression method (red dashed line) and the target mean hazard from full PSHA (blue solid line) at three representative sites.

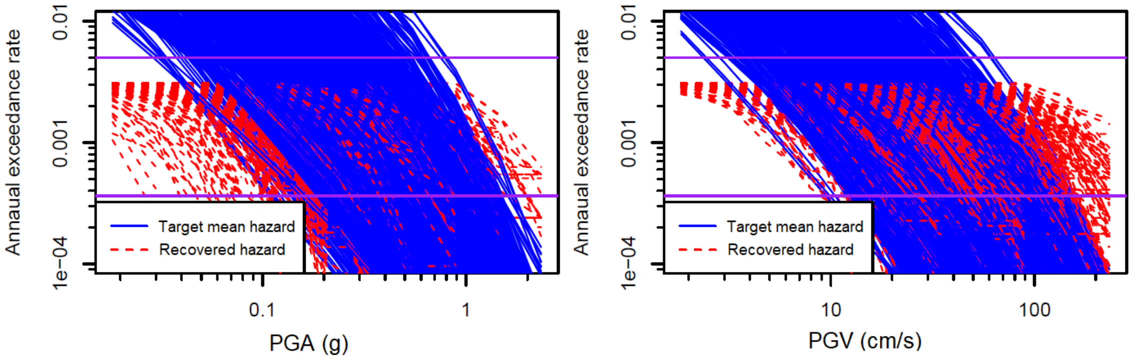


Figure 5.9. Plots of recovered hazard curves by the 25 selected ground motion maps by the LASSO regression method (red dashed line) and the target mean hazard from full PSHA (blue solid line) at 1,220 sites.

The number of selected ground motion realizations, 25, is constrained by the computation demand for the subsequent geo-hazard analyses. If the number of selected ground motion realizations could be larger, the recovered hazard curves fitting would be improved. In Figure 5.10, we plot the recovered hazard curves with 200 ground motion maps to illustrate smaller mismatches.

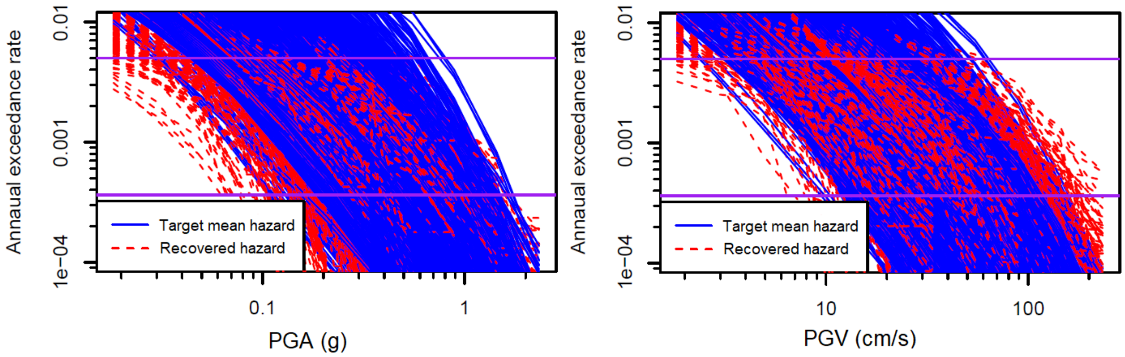


Figure 5.10. Plots of recovered hazard curves from 200 selected ground motion realizations by the LASSO regression method (red dashed line) and the target mean hazard from full PSHA (blue solid line) at 1,220 sites.

5.3 Co-Kriging Interpolation for Hazard-Consistent Correlated Ground Motion Maps

As described above, the ground motions realizations were generated and selected at 1,220 sites (shown in Figure 5.2). For geo-hazard analysis, higher resolution maps (with a spacing of 100 meters) are required, so interpolation is needed. Since PGA and PGV are correlated, we need to consider spatial and cross *IM*s correlations when interpolating. PGA and PGV can be considered as two covariates at each site. Accordingly, when interpolating these *IM*s for a new site, a vector of two covariates needs to be estimated simultaneously. Co-Kriging is a geostatistical tool that can perform correlated multivariate interpolation.

Given a simulated correlated PGA (denoted as k_1) and PGV (denoted as k_2) realization g at J sites, the estimated or interpolated ground motions for k_1 and k_2 at a new site j_0 can be written as

$$\begin{bmatrix} \hat{z}_{k_1,j_0} \\ \hat{z}_{k_2,j_0} \end{bmatrix} = \begin{bmatrix} \hat{\mu}_{k_1,j_0} \\ \hat{\mu}_{k_2,j_0} \end{bmatrix} + \begin{bmatrix} \tilde{\eta}_{k_1} \\ \tilde{\eta}_{k_2} \end{bmatrix} + \begin{bmatrix} \widehat{\delta W}_{k_1,j_0} \\ \widehat{\delta W}_{k_2,j_0} \end{bmatrix} \tag{5.9}$$

where $\hat{z}_{k_1,j_0}$ is the interpolated ground motion for k_1 at a new site j_0, which is the sum of median logarithmic ground motion $\hat{\mu}_{k_1,j_0}$ (predicted by a GMM), event term $\tilde{\eta}_{k_1}$ (the same simulated value used by the other simulated J sites), and interpolated within event residual $\widehat{\delta W}_{k_1,j_0}$ (different from the simulated $\widetilde{\boldsymbol{\delta W}}_{\boldsymbol{k_1}}$ at other J sites). The second row for $\hat{z}_{k_2,j_0}$ is evaluated in the same manner. For the interpolated within event residual, we need to solve a Co-Kriging system,

$$\boldsymbol{\Sigma w} = \boldsymbol{c} \tag{5.10}$$

where

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{\boldsymbol{k_1}} & \mathbf{\Sigma}_{\boldsymbol{k_1,k_2}} & \mathbf{1} & \mathbf{0} \\ \mathbf{\Sigma}_{\boldsymbol{k_2,k_1}} & \mathbf{\Sigma}_{\boldsymbol{k_2}} & \mathbf{0} & \mathbf{1} \\ \mathbf{1}^T & \mathbf{0}^T & 0 & 0 \\ \mathbf{0}^T & \mathbf{1}^T & 0 & 0 \end{bmatrix} \tag{5.11}$$

is the covariance matrix of PGA and PGV within event residuals among the given J sites, which contains the $\mathbf{\Sigma}_{\boldsymbol{k_1}}$ and $\mathbf{\Sigma}_{\boldsymbol{k_2}}$ terms defined in Eq. (4.14) in Chapter 4.1 and vectors of $\mathbf{1}$ (a column vector of J elements all equal to 1), $\mathbf{0}$ (a column vector of J elements all equal to 0), $\mathbf{1}^T$ (a row vector of J elements all equal to 1), $\mathbf{0}^T$ (a row vector of J elements all equal to 0), and scalars 0. The $\boldsymbol{w}$ weight vector in Eq. (5.10) is defined as

$$\boldsymbol{w} = \begin{bmatrix} \boldsymbol{w}_{\boldsymbol{k_1}} \\ \boldsymbol{w}_{\boldsymbol{k_2}} \\ -\lambda_{k_1} \\ -\lambda_{k_2} \end{bmatrix} \tag{5.12}$$

where $\boldsymbol{w}_{\boldsymbol{k_1}} = \left\{w_{k_1,1}, w_{k_1,2}, \cdots, w_{k_1,J}\right\}$ and $\boldsymbol{w}_{\boldsymbol{k_2}} = \left\{w_{k_2,1}, w_{k_2,2}, \cdots, w_{k_2,J}\right\}$ for $\boldsymbol{\delta\widetilde{W}}_{\boldsymbol{k_1}}$ and $\boldsymbol{\delta\widetilde{W}}_{\boldsymbol{k_2}}$ at J sites respectively to interpolate $\widehat{\delta W}_{k_1,j_0}$, and λ_{k_1} and λ_{k_2} are Lagrange multipliers (that will not be used in interpolation). The covariance vector $\boldsymbol{c}$ in Eq. (5.10) is defined as,

$$\boldsymbol{c} = \begin{bmatrix} \boldsymbol{c}_{\boldsymbol{k_1}}(j_0) \\ \boldsymbol{c}_{\boldsymbol{k_2}}(j_0) \\ 1 \\ 0 \end{bmatrix} \tag{5.13}$$

where $\boldsymbol{c}_{\boldsymbol{k_1}}(j_0)$ is the covariance between the new site j_0 and simulated J sites for k_1 and $\boldsymbol{c}_{\boldsymbol{k_2}}(j_0)$ is the covariance between the new site j_0 and simulated J sites for k_2. By using the developed spatial and cross-*IM*s correlation model (e.g., Loth and Baker, 2013), we can establish $\mathbf{\Sigma}$ and $\boldsymbol{c}$ and then solve for the weight vector $\boldsymbol{w}$ by,

$$\boldsymbol{w} = \mathbf{\Sigma}^{-1}\boldsymbol{c} \tag{5.14}$$

where the exponent -1 of $\mathbf{\Sigma}$ indicates inverse of the matrix. Once the weight vector is solved, we can interpolate within event residual $\widehat{\delta W}_{k_1,j_0}$ for k_1 at the new site j_0 as,

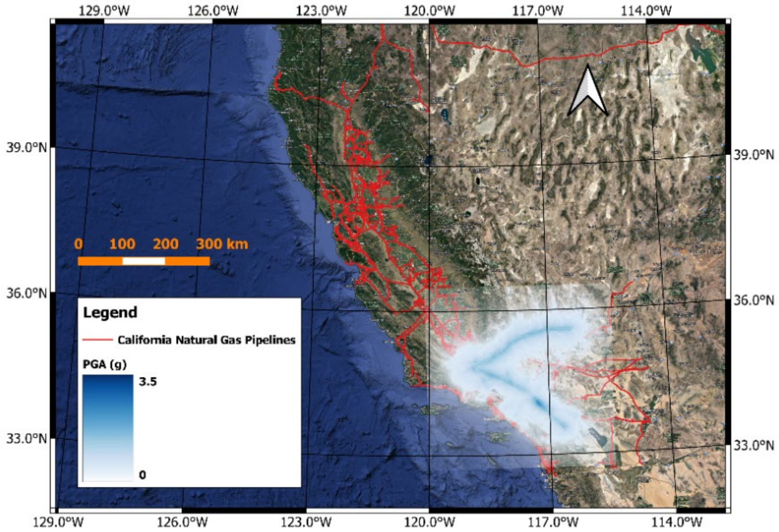
$$\widehat{\delta W}_{k_1,j_0} = \sum_{j=0}^{j=J} \widehat{\delta W}_{k_1,j} w_{k_1,j} + \sum_{j=0}^{j=J} \widehat{\delta W}_{k_2,j} w_{k_2,j} \tag{5.15}$$

After that, we need to switch the position of the corresponding covariances for k_1 and k_2 in Eqs. (5.11) and (5.13) to obtain a new $\mathbf{\Sigma}$ and $\boldsymbol{c}$ and solve a new weight vector $\boldsymbol{w}$. The new weight vector $\boldsymbol{w}$ can be entered into Eq. (5.15) again to interpolate within event residual $\widehat{\delta W}_{k_2,j_0}$ for k_2.

Following this co-Kriging interpolation method, we loop across each grid point to obtain high-resolution correlated maps.

One final step before finalizing the ground motion realizations is correcting for site effects. The PSHA results and subsequent steps were all based on reference site conditions (i.e., V_{S30} = 760 m/s). Therefore, we need to estimate site-specific V_{S30} and apply site amplification to obtain the ground motions that reflect local site conditions. Site-specific V_{S30} values were estimated using multiple proxy-based approaches, including topographic slope (Wald and Allen, 2007), geomorphic terrain classifications (Yong et al., 2012; Yong, 2016), surface geology (Wills et al., 2015), and a Kriging-based interpolated map (Thompson et al., 2014; Thompson, 2018). The manner in which these proxies are combined is described by Wang (2020). For the site amplification correction, we adapt the ergodic site response model by Seyhan and Stewart (2014), which includes both linear (F_{lin}) and nonlinear (F_{nl}) site responses.

Figures 5.11 and 5.12 show the fifth out of 25 selected correlated PGA and PGV interpolated maps. The event considered in scenario 5 is a multi-fault rupture with a magnitude of 8.02. The participating faults include the Ft. Tejon segment of the San Andreas Fault (in the north-south direction) and the Garlock fault (in the east-west direction). The corrected hazard-consistent annual occurrence rate for the map is 0.0005474 (the annual occurrence rate of this selected event is 2.27E-08 originally modeled in UCERF3 and is corrected to 0.0003005 for hazard-consistency in the subset of selected events). All 25 maps and their associated metadata can be found in a shared OneDrive folder[1].



[1] https://olddominion-my.sharepoint.com/:f:/g/personal/p1wang_odu_edu/EqPt2q8qMmBDsksyaLcMuq4B1FZwrOc39zAe_kTliUrpjQ?e=ohgxk2

Figure 5.11. Spatial correlated PGA map, the 5th out of 25 selected maps (corrected hazard-consistent annual occurrence rate is 0.0005474)

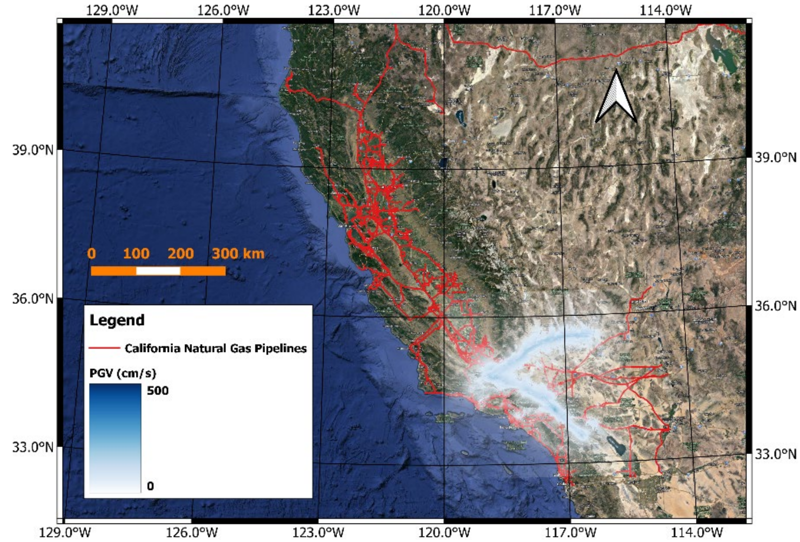


Figure 5.12. Spatial correlated PGV map, the 5th out of 25 selected maps (corrected hazard-consistent annual occurrence rate is 0.0005474).

6 Summary

6.1 Conclusion

In this report, we have presented efficient LASSO regression-based methods for the selection of scenario events and correlated ground motion realizations, which are useful for seismic risk assessments of spatially distributed infrastructure. The methods have the flexibility of matching hazard curves across multiple sites and for multiple intensity measures or alternatively for those objectives combined while preserving the magnitude and distance marginal distributions from disaggregation. The framework provides a series of possible results for alternate values of tuning parameter γ, which gives the user the flexibility to select their preferred subset of events or realizations in consideration of fit error.

The proposed method was applied to select correlated multi-*IM*s (PGA and PGV) realizations for statewide natural gas pipelines. To reduce the computation demand, 1,220 target sites close to the pipelines were selected as the hazard control target sites. Taking the conventional point-based PSHA results as input (completed by Al Atik et al., 2022), we first pre-selected 7,700 important scenarios based on disaggregations and then conducted LASSO regression to select 599 scenario events that achieve hazard-consistency and preserve the magnitude marginal distributions from disaggregations. We then generated 50 correlated PGA and PGV maps for each of the 599 selected events, which resulted in 29,950 realizations. After ground motion realization conversion and tensor transformation, these realizations were formulated as a hazard matrix that can be incorporated into LASSO regression for selection. Given computation constraints in subsequent geo-hazard analyses, 25 correlated PGA and PGV realizations were selected together with their hazard-consistent annual occurrence rates. Co-Kriging and site amplification corrections were implemented to interpolate and obtain high-resolution correlated ground motion maps under site-specific conditions. A publicly accessible self-contained R package was also developed to perform the necessary calculations.

6.2 Limitations

The application of the methodologies presented in this report has several limitations, as follows:

1. The hazard-consistent events and correlated maps were conducted based on 1,220 target sites close to pipelines, not 19,316 sites evenly distributed throughout the state. As a result, the selected events and ground motion spatial realizations are best suited for seismic assessment of the pipelines and may not be applicable to any other regions in the state.
2. Only PGA and PGV, two *IM*s, and a relatively narrow return period range (from 200 to 2,475 years) were considered for event and realization selection. If other *IM*s or return periods are required, the selected maps may not be applicable.
3. Given the computation constraints in the subsequent geo-hazard analyses, only a maximum of 25 maps were selected. As a result, the hazard curve mismatches are relatively large. These mismatches likely inflate uncertainties in subsequent seismic risk assessments.
4. Magnitude and distance marginal distributions from disaggregations were not considered in ground motion realization selection, so the fitting could be poor. Therefore, the selected maps may produce some biases for applications that require preserving marginal distributions of either magnitude or distance.

6.3 Future Work

Future work can overcome the limitations described in Chapter 6.2. Such work could include:

1. Validation to check if hazard can be preserved at sites that are not incorporated in the LASSO regression. We anticipate the hazard preservation may be acceptable if the sites are relatively close to the target sites within a distance threshold (e.g., 5 km), but the mismatch will grow with increasing distance. To better manage the selection of statewide target sites (not just for pipelines), we need to understand the relation between the hazard mismatch and distance.
2. Different future applications are likely to require different *IM*s and return periods than those considered here. Accordingly, further study is needed to investigate how to efficiently select correlated multi-*IM*s realizations for a broader return period range. For example, how much the mismatch of PSA at 1 second is if we only match PGA and PGV?
3. It is almost impossible to use a small set of maps to preserve both hazard curves and disaggregations well for a large region. More selected maps are mandatory to improve the accuracy of final seismic risk analysis. If future work of the geo-hazards group can consider a larger number of ground motion realizations, this can be provided.
4. LASSO regression is a very efficient method to conduct selection with a large number of inputs. However, we are still limited to incorporating magnitude and distance marginal distributions when selecting correlated maps. We will investigate if there exists more efficient LASSO solvers or different implementation approaches to allow more inputs.

References

Al Atik L., Gregor N., and Bozorgnia Y. (2022) Probabilistic seismic hazard analysis for the state of California, Natural Hazards Risk & Resiliency Research Center B. John Institute for the Risk Sciences.

Bazzurro P. and Cornell C.A. (1999) Disaggregation of seismic hazard, Bull. Seis. Soc. Am. 89: 501–520.

Boore D.M., Stewart J.P., Seyhan E., Atkinson G.M. (2014). NGA-West2 equations for predicting PGA, PGV, and 5% damped PSA for shallow crustal earthquakes. Earthq. Spectra. 30(3): 1057–1085.

Campbell, K. and Seligson, H. (2003). Quantitative method for developing hazard-consistent earthquake scenarios. Proc. of the Technical Council of Lifeline Earthquake Engineering, ASCE, ed. J. Beavers (ASCE, Long Beach, CA), pp. 829–838.

Chang, S. E., Shinozuka M., and Moore J. E. (2000). Probabilistic earthquake scenarios: extending risk analysis methodologies to spatially distributed systems. Earthquake Spectra, 16(3): 557–572.

Field E.H., Arrowsmith R.J., Biasi G.P., Bird P., Dawson T.E., Felzer K.R., Jackson D.D., Johnson K.M., Jordan T.H., Madden C., Michael A.J., Milner K.R., Page M.T., Parsons T., Powers P.M., Shaw B.E., Thatcher W.R., Weldon R.J., Zeng Y. (2015). Uniform California earthquake rupture forecast, version 3 (UCERF3)--The time-independent model. Bull. Seismol. Soc. Am., 104 (3): 1122–1180.

Han Y. and Davidson R.A. (2012). Probabilistic seismic hazard analysis for spatially distributed infrastructure. Earthq. Eng. Struct. Dyn., 41 (15): 2141–2158.

Jayaram N. and Baker J.W. (2009). Correlation model for spatially-distributed ground-motion intensities. Earthq. Eng. Struct. Dyn., 38 (15): 1687–1708.

Loth C. and Baker J.W. (2013). A spatial cross-correlation model of spectral accelerations at multiple periods. Earthq. Eng. Struct. Dyn., 42 (3):397–417.

McGuire R.K. (2004). Seismic hazard and risk analysis. Earthquake Engineering Research Institute Monograph MNO-10, 2004.

Seyhan E. and Stewart J.P. (2014). Semi-empirical nonlinear site amplification from NGA-West2 data and simulations. *Earthquake Spectra*. 30 (3): 1241–1256.

Stewart J.P., Assimaki D., Rathje, E.M., Lavrentiadis G., Ojomo O., Wang P., and Zimmaro P. (2023) Regional analysis of spatially distributed ground failure displacement hazards in California, Natural Hazards Risk & Resiliency Research Center B. John Institute for the Risk Sciences.

Thompson, E.M., (2018). An Updated Vs30 Map for California with Geologic and Topographic Constraints: U.S. Geological Survey data release. DOI:10.5066/F7JQ108S.

Thompson, E.M., Wald, D.J., Worden, C.B. (2014). A Vs30 Map for California with Geologic and Topographic Constraints. Bulletin of the Seismological Society of America. 104: 2313–2321.

Tibshirani R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58 (1): 267–288.

Wald, D.J., and Allen, T.I. (2007). Topographic slope as a proxy for seismic site conditions and amplification. Bulletin of the Seismological Society of America. 97: 1379–1395.

Wang P. (2020). Predictability and repeatability of non-ergodic site response for diverse geological conditions. Ph.D. thesis. Civil & Environmental Engineering Dept., UC Los Angeles.

Wang P. (2022). Regional Probabilistic Seismic Hazard Assessment (RPSHA) Package. Available at https://github.com/wltcwpf/RPSHA. Last updated March, 2022.

Wang P., Liu Z., Brandenberg S.J., Zimmaro P., and Stewart J.P. (202x). Regression-based scenario earthquake selection for regional hazard-consistent risk assessments. *Earthquake Spectra*. (*Accepted*).

Wills, C.J., Gutierrez, C.I., Perez, F.G., and Branum, D.M. (2015). A Next Generation VS30 Map for California Based on Geology and Topography. Bulletin of the Seismological Society of America. 105: 3083–3091.

Yong, A. (2016). Comparison of measured and proxy-based VS30 values in California. Earthquake Spectra. 32: 171–192.

Yong, A, Hough, S.E., Iwahashi, J., and Braverman, A. (2012). A terrain based site conditions map of California with implications for the contiguous United States. Bulletin of the Seismological Society of America. 102: 114–128.