Data construction method for the analysis of the spatial distribution of disastrous earthquakes in Taiwan

Hsiao-Fan Wang and Chun-Jung Huang

Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Hsinchu, Taiwan, Republic of China
E-mail: hfwang@ie.nthu.edu.tw

Received 10 October 2006; received in revised form 10 June 2008; accepted 22 July 2008

Abstract

Considering a disastrous earthquake as a rare event, the aim of this study is to apply the proposed data construction method (DCM) to determine the possible distribution pattern of disastrous earthquakes in Taiwan. Owing to the availability of only a limited amount of data and based on the multiset division of DCM, virtual samples have been generated. The procedure is illustrated by a numerical experiment that consists of data from 12 disastrous earthquakes in Taiwan from 1990 to 1999. The results show that the pattern constructed by DCM is geologically consistent with the actual phenomenon, which was caused by the collision of the Philippine Sea Plate and the Eurasian Continental Plate. A case study of disastrous earthquakes in East Taiwan is then conducted by studying three near-source regions. Based on the Kolmogorov–Smirnov test, the constructed spatial distribution has shown its validity and capability in providing useful information for the risk assessment of disastrous earthquakes as rare events.

Keywords: data construction method; plate movement theory; disastrous earthquake; rare events; small sample; virtual sample generation

1. Introduction

Plate movement theory has not only contributed to significant research in the field of Earth Science in the 20th century but also provides a practical explanation as to why earthquakes take place. In particular, Taiwan lies in a junction between the Philippine Sea Plate and the Eurasian Continental Plate, and the collision of these two plates results in frequent earthquakes. The occurrence of an earthquake results in the release of energy. Most bring no harm to the environment and living beings. Then even if it does cause harm, the influence of attenuation is less important locally (Kracke and Heinrich, 2004). However, there are a few of those so-called “disastrous earthquakes” possessing magnitudes ($M$) no less than 5.0, which can cause the death
of human beings and result in the heavy loss of goods. Examples of these disastrous earthquakes are the San Francisco earthquake ($M$: 7.8) in 1906, the Chile earthquake ($M$: 8.6) in 1960, the Mexico earthquake ($M$: 7.8) in 1985 and recently in May 2008, Sichuan earthquake ($M$: 8.9). Because related data are statistically few, such earthquakes are regarded as rare events (Bucklew, 2004). Furthermore, how to characterize the spatial distribution of disastrous earthquakes for the assessment of recurrent risk still remains a question.

Owing to the disaster brought about by the Chi-Chi earthquake ($M$: 7.3) in Taiwan in 1999, considerable research has focused on the allocation of seismic hazard regions with respect to the active faults resulting from plate collisions. Even so, two crucial issues relating to the assessment of recurrent risk remain: (1) When, where, and at what scale will the next disastrous earthquake occur? (2) How does one evaluate the occurrence probability of an earthquake at a specific location? The former involves the forecast of the occurrence of an earthquake, the details of which include the time, location, and scale. The latter, referred to as the so-called Seismic Risk Assessment, requires the analysis of the possible impact of earthquakes in a specific region. Although the inherent difficulty of forecasting the occurrence of an earthquake greatly contributes to the uncertainty of prediction as based on historical data, efforts have been made to analyze the possible impact of earthquakes in order to reduce the degree of damage they bring about.

In general, seismic hazard is expressed as a specific depth, magnitude, intensity, or acceleration. Quantities are often represented on a seismic hazard map (Kayabali, 2002; Kracke and Heinrich, 2004). Depending on the scale of analysis, the maps can be used to evaluate the impacts on existing buildings (Jun et al., 1997) or specific cities (Leydecker and Kopera, 1999). Although some studies had been carried out to show that there is a nonlinear relationship between earthquake magnitude and intensity (Cavallini and Rebez, 1996) and that a negative exponential distribution holds for the magnitude of an earthquake (Scholz, 1998), it is useful to evaluate the possible risk of earthquakes using geological prospect and fault confirmation.

Moreover, except for geological studies, drawing the spatial distribution of disastrous earthquakes is also an effective approach to risk assessment because each disastrous earthquake signifies that a huge energy release brings about the loss of human lives and goods. However, although there are, on average, 2000 perceptible earthquakes per year in Taiwan, only 86 disastrous earthquakes have taken place within 100 years from 1900 to 1999. Therefore, due to the lack of sufficient data, recognizing earthquakes’ complete pattern has been a challenging task.

A systematic approach called the data construction method (DCM) has been shown to be able to generate virtual samples effectively (Huang and Wang, 2008). Therefore, the aim of this study is to apply this method to analyze disastrous earthquakes as a rare event. The remainder of this paper is laid out as follows. In Section 2, the theoretical background of DCM is first introduced. A numerical example of disastrous earthquakes in Taiwan that analyzes the trend relationship between the frequencies of disastrous earthquakes with respect to the longitudes and latitudes is then demonstrated for illustration purposes. A full-scale case study on the three near-source regions that cover East Taiwan is presented in Section 3. Each region has been assigned as a virtual center. Using historical data from 1959 to 1999 as the training data of DCM, virtual samples are generated in the three regions. For validation, the Kolmogorov–Smirnov (K–S) test is used to confirm the three membership functions derived from the virtual samples with respect to the empirical distribution functions of all disastrous earthquakes in three near-source regions in East Taiwan from 1900 to 1999. Finally, the discussion and conclusion are presented in Section 4.
2. Methodology

Before proceeding to analyze the spatial distribution of disastrous earthquakes in Taiwan, the features and procedures of the DCM will be briefly discussed.

2.1. Definitions and properties of DCM

Definitions and basic features of DCM are first given in this section.

Definition 1. Multiset (Blizard, 1989)

A multiset, \( L = (A, m) \), in set theory is defined by a set of ordered pairs as

\[
L = \{(a, m(a)) | a \in A; m : A \rightarrow \mathbb{N}\},
\]

where \( A \) is some set and \( m \) is a function that denotes the multiplicity (number of occurrences) of \( a \) in \( A \times m(a) \). Without confusion, it can be denoted by \( A = \{(a, m(a)) | a \in A; m(a) \in \mathbb{N}\} \).

If the set \( A \) is finite, the size or length of the multiset \( (A, m) \) is the sum over all multiplicities of the elements in \( A \) with the form \(|(A, m)| = \sum_{a \in A} m(a)|\).

For instance, given a set \( A = \{a, b, c, b, c, a\} \), then \( 3, 3, \) and \( 2 \) would be the multiplicities of \( a, b, \) and \( c \), respectively. Thus, \( A \) can be expressed as a multiset in ordered pairs as \( L = \{(a, 3), (b, 3), (c, 2)\} \) with size \(|L| = 8\).

By considering a mode as the highest frequency in appearance, the above example shows multiple modes. When giving a sample shown by the multiset form as \( A_o = \{(a_{i_o}, k_i) | a_{i_o} \in \mathbb{R}, k_i \in \mathbb{N}, 1 \leq i_o, i \leq n\} \), a representative mode can be uniquely defined by

\[
\text{mode}(A_o) = \left\{ \min_{a_{i_o}} \left\{ \max_{k_i} (a_{i_o}, k_i) \right\} | 1 \leq i_o, i \leq n \right\}.
\]

Whereas the mode location will lead to different distribution shapes for the generated virtual sample, if in practice, any additional information about the mode is perceived in advance, e.g. making the data speak through the exploratory data analysis, or in other kind of information, a different reference point can be redefined.

Now, based on the selected mode, \( A_o \) can be translated into \( A \) by subtracting each element of \( A_o \) from \( \text{mode}(A_o) \). Then, after re-arranging in ascending order, we have

\[
A = \{(a_i, k_i) | a_i \in \mathbb{R}, k_i \in \mathbb{N}, 1 \leq i \leq n\},
\]

where \( a_i = a_{i_o} - \text{mode}(A_o) \) and \( a_{i+1} > a_i \).

Therefore, \( 0 \in A \) always holds with \( \text{mode}(A) = 0 \). For instance, if \( A_o = \{(16, 3), (17, 1), (19, 5), (20, 2)\} \), then \( \text{mode}(A_o) = 19 \) and \( A = \{(-3, 3), (-2, 1), (0, 5), (1, 2)\} \); or if \( A_o = \{(16, 1), (18, 3), (19, 3), (20, 1)\} \), then \( \text{mode}(A_o) = 18 \) and \( A = \{(-2, 1), (0, 3), (1, 3), (2, 1)\} \).

Because the mode represents a datum that occurs the most frequently in the original sample, it is relatively important and should be retained during the data construction process defined below.

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies
Definition 2. Multiset division

Given two multisets \( A = \{(a_i, k_i) | a_{i+1} > a_i, \ k_i \in \mathbb{N}, \ 1 \leq j \leq m \} \) and \( C = \{(c_j, p_j) | c_{j+1} > c_j, \ p_j \in \mathbb{N}, \ 1 \leq j \leq m \} \), the multiset division of \( A \) by \( C \) is defined by

\[
A(\cdot)C = \{(c_j^{-1}a_i, k_i p_j) | 1 \leq i \leq n, \ 1 \leq j \leq m, k_i \in \mathbb{N}, p_j \in \mathbb{N} \}, \quad \text{where} \ \forall j, \ c_j \neq 0. \tag{3}
\]

While the division operated on single values (single value by single value) results in equal numbers of ratios, division on the multisets (multiset by multiset) will result in a multiset with many more elements. In addition, to ensure that the resultant multiset is bounded from both sides in a division process, apart from \( 0 \subseteq A \), the other condition is assumed, that is, \( \inf(C) = 1 \). For ease of illustration, let us consider the divisor set \( C \) including only two elements \( \{1, c\} \) with \( c > 1 \), then the resultant set is defined by

\[
Z^t = A(\cdot)C^t, \tag{4}
\]

where \( A \) is a multiset defined in equation (2); \( C = \{(1, 1), (c, 1)\} \) with \( c > 1 \) is a given divisor and \( t \) denotes the number of divisions. For instance, \( Z^t \equiv A(\cdot)C(\cdot)C(\cdot)C \).

Then, the resultant multiset \( Z^t \) has the following properties:

**Property 1. Virtual sample generation.**

Given \( A = \{(a_i, k_i) | a_i \in \mathbb{R}, k_i \in \mathbb{N}, 1 \leq i \leq n \} \) with \( 0 \in A \), and \( C = \{(1, 1), (c, 1)\} \) with \( c > 1 \), the size of \( Z^t \) is \( |Z^t| = |C|^t \times |A| = 2^t \times |A|, \ \forall t \in \mathbb{N} \).

**Proof.** See Appendix A1.

Therefore, by applying the \( 2^t \)-multiple generation procedure, we can obtain additional amount of virtual sample for analysis.

**Property 2. Bounded multiset division.**

Given \( A = \{(a_i, k_i) | a_i \in \mathbb{R}, k_i \in \mathbb{N}, 1 \leq i \leq n \} \) with \( 0 \in A \), and \( C = \{(1, 1), (c, 1)\} \) with \( c > 1 \), let \( Z^t \equiv A(\cdot)C^t \), then \( \sup(Z^t) = a_n \) and \( \inf(Z^t) = a_1, \ \forall t \in \mathbb{N} \).

**Proof.** See Appendix A2.

Property 2 says that being \( \inf(C) = 1 \) and \( 0 \in A \), the universal discourse of the constructed multiset \( Z^t \) will always be bounded by \([a_1, a_n]\) no matter how many times the multiset division is conducted. This confirms that the proposed multiset division is bounded.

Because we aim to generate a sufficient number of virtual samples for analysis during the division process, we should know whether \( |Z^t| \) is enough or not. That is, how large \( t \) should be in order to have \( Z^t \) with the desired distribution. Chebyshev’s inequality defined below will provide such criterion:

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies
Lemma 1. Chebyshev’s inequality (Ross, 1987).

Given a random variable x with mean μ and finite variance σ², then for any value k > 1, 
\[ P(|x - \mu| \leq k\sigma) = 1 - \frac{1}{k^2} \]

The importance of Chebyshev’s inequality is that it enables us to estimate the number of 
required data falling in the range of mean within multiple of its standard deviation such that the 
lower bound of probabilities of data occurrences is ensured. Of course, if the actual distribution 
was known, then the desired probabilities could be exactly computed. However, if μ and σ² are 
unknown, when giving a sample with size n the sample mean \( \bar{x} \) and variance \( s^2 \) can be adapted by the 
necessary modification of using \( P(|x - \bar{x}| \leq ks) \geq 1 - ((n + 1)(k^2 + n - 1))^{-1}(nk^2) \) instead 
(Saw et al., 1984). This paves the way for defining the stopping rule for the division process as below:

Definition 4. Stopping rule of multiset division procedure.

Let 
\[ Z^t \equiv A(\cdot):C^t = \{(z_q^t,f_q^t)|z_q^t \in [a_1,a_n], f_q^t \in \mathbb{N}, 1 \leq q \leq Q\}, \quad Q \leq 2^t \times |A| \text{ with } |C| = 2, \forall t \in \mathbb{N}. \]

Then, for any value k > 1, when the probability is greater than or equal to the lower bound of w as below
\[ P(|z_q^t - \bar{z}_t|) \geq 1 - ((|Z^t| + 1)(k^2 + |Z^t| - 1))^{-1}(|Z^t|k^2), \]
the division process stops and the number of divisions, t, is determined where \( \bar{z}_t = \left(\frac{\sum_q z_q^t \times f_q^t}{|Z^t|}\right) \) and \( s_t^2 = \left(\frac{|Z^t|^{-1} \sum_q z_q^t f_q^t - (\bar{z}_t)^2}{2}\right)^{0.5} \).

Once the desired amount of data are generated, the universal discourse of \( Z^T \) is translated back using \( a_i = a_{i_0} - \text{mode}(A_{i_0}), \quad 1 \leq i \leq n \). Thus, supposing \( t = T \), virtual samples of \( Z^{T'} = \{(z_q^{T'},f_q^{T'})|1 \leq q \leq Q\} \) with \( z_q^{T'} = z_q^T + \text{mode}(A_{i_0}) \) will be obtained. This virtual sample generation procedure is summarized in the following algorithm and is illustrated step by step using a numerical example.

2.2. The algorithm of DCM

**Step 0:** Given \( k > 1 \) and set \( t = 1; \)

**Step 1:** Input a sample \( A_{i_0} = \{(a_{i_0},k_i)|a_{i_0} \in \mathbb{R}, k_i \in \mathbb{N}, 1 \leq i_0 \leq n\} \) and determine an arbitrary set 
\( C = \{(1,1),(c,1)\} \) with \( c > 1; \)

**Step 2:** Take mode\((A_{i_0}) = \left\{ \min_{a_{i_0} \in \mathbb{R}} \left\{ \max_{k_i \in \mathbb{N}} (a_{i_0},k_i) | 1 \leq i_0, i \leq n \right\} \right\} \) and let \( A = \{(a_i,k_i)|a_i \in \mathbb{R}, k_i \in \mathbb{N}, 1 \leq i \leq n\} \) \( \text{where } a_i = a_{i_0} - \text{mode}(A_{i_0}); \)

**Step 3:** Let \( Z^t \equiv A(\cdot):C^t = \{(z_q^t,f_q^t)|z_q^t \in [a_1,a_n], f_q^t \in \mathbb{N}, 1 \leq q \leq Q\}, \quad Q \leq 2^t \times |A|, \forall t \in \mathbb{N}; \)

© 2009 The Authors.

Journal compilation © 2009 International Federation of Operational Research Societies
Step 4: Set a constant $w$ with 

$$w = 1 - (|Z|^2 + |Z_t|^2 - 1)^{-1}(|Z|^2 k^2),$$

$$\bar{x} = \left(\frac{\sum_{q} z_q f_q}{|Z|}\right),$$

and

$$s = \left(\frac{(|Z|^2 - 1)^{1/2}}{\sum_{q} z_q^2 f_q} - (\bar{x})^2\right)^{0.5}.$$ 

If $P(|z_q - \bar{x}| \leq ks) \geq w$, go to Step 5; otherwise, $t = t + 1$ and go to Step 3.

Step 5: Let $t = T$. Calculate $z_q^* = z_q^r + \text{mode}(A_o)$ and output $Z^T = \{(z_q^*, f_q)\}$.

2.3. Numerical example

Taking the data from 12 disastrous earthquakes that took place in East Taiwan from 1990 to 1999 as an example, Table 1 lists their geological locations (epicenters).

The data points, each representing a disastrous earthquake epicenter, are shown as the pair value of latitude and longitude in Fig. 1. Apart from the earthquakes' uniform frequencies, there is no other information revealed from the graph. Therefore, the DCM is used to estimate a possible pattern as an illustrative example. A step-by-step procedure is shown in Section 2.2 as follows.

Step 1: Denote the given sample of size 12 as $A_o = \{(23.9, 121.5), (23.9, 121.8), \ldots, (23.5, 120.4)\}$ with 2D data showing the latitude and the longitude of the location and the third number representing its frequency.

Step 2: The modes for two coordinates can be found by 23.9 and 121.7 respectively. Then, $A_o$ is translated into $A = \{(a_{i1}, a_{i2}, k_i) | i = 1 \sim 12\} = \{(0, -0.2, 1), (0, 0.1, 1), (0.7, -1.6, 1), (0.1, -0.1, 1), (0.8, -0.3, 1), (0.7, -1.2, 1), (0.5, 0.1, 1), (0.3, 0, 1), (0.7, 0, 1), (0.4, -1, 1), (0, -0.9, 1), (0.4, -1.3, 1)\}$.

Step 3: Set $C = \{(1, 1), (2, 1)\}$ and carry out multiset division to generate virtual samples. For example, when $t = 1$, $Z^1 \equiv A(1, C) = \{(a_{i1} / 1, a_{i2} / 1, k_i) | i = 1 \sim 12\} = \{(0, -0.2, 1), (0, 0.1, 1), (0.7, -1.6, 1), (0.1, -0.1, 1), (0.8, -0.3, 1), (0.7, -1.2, 1), (0.5, 0.1, 1), (0.3, 0, 1), (0.7, 0, 1), (0.4, -1, 1), (0, -0.9, 1), (0.4, -1.3, 1)\}$.

Table 1
The disastrous earthquake data for Taiwan from 1990 to 1999

<table>
<thead>
<tr>
<th>No.</th>
<th>Date</th>
<th>Latitude (°N)</th>
<th>Longitude (°E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1990/12/13</td>
<td>23.9</td>
<td>121.5</td>
</tr>
<tr>
<td>2</td>
<td>1990/12/14</td>
<td>23.9</td>
<td>121.8</td>
</tr>
<tr>
<td>3</td>
<td>1991/03/12</td>
<td>23.2</td>
<td>120.1</td>
</tr>
<tr>
<td>4</td>
<td>1992/04/20</td>
<td>23.8</td>
<td>121.6</td>
</tr>
<tr>
<td>5</td>
<td>1992/05/29</td>
<td>23.1</td>
<td>121.4</td>
</tr>
<tr>
<td>6</td>
<td>1993/12/16</td>
<td>23.2</td>
<td>120.5</td>
</tr>
<tr>
<td>7</td>
<td>1994/06/05</td>
<td>24.4</td>
<td>121.8</td>
</tr>
<tr>
<td>8</td>
<td>1995/02/23</td>
<td>24.2</td>
<td>121.7</td>
</tr>
<tr>
<td>9</td>
<td>1995/06/25</td>
<td>24.6</td>
<td>121.7</td>
</tr>
<tr>
<td>10</td>
<td>1998/07/17</td>
<td>23.5</td>
<td>120.7</td>
</tr>
<tr>
<td>11</td>
<td>1999/09/21</td>
<td>23.9</td>
<td>120.8</td>
</tr>
<tr>
<td>12</td>
<td>1999/10/22</td>
<td>23.5</td>
<td>120.4</td>
</tr>
</tbody>
</table>

Data source: Central Geological Survey, MOEA, Taiwan, Republic of China.
Step 4: Set $k = 2$, and $w = 0.95$ (which is no less than $1 - (|Z'_{\text{in}}| + 1)(k^2 + |Z'_{\text{out}}| - 1)^{-1}(|Z'_{\text{out}}|k^2)$), multiset division is repeated as shown in Table 2. After $t = 4$, both attributes satisfy the stopping rule. Then, the division procedure stops and $Z^4 = \{(0, -0.2), 1), ((0, -0.1), 1), ((-0.7, -1.6), 1), \ldots, ((-0.025, -0.0625), 1), ((-0.025, -0.08125), 1)\}$ is obtained.

Step 5: The domain values of $Z^4$ were translated back with 23.9 and 121.7, respectively. Based on Property 1, the virtual sample $Z^4''$ with size 192 was generated.

Let us show $Z^4$, by 3D grid graph as in Fig. 2.

By considering the relative location among Taiwan, the Philippine Sea Plate, and the Eurasian Continental Plate as shown in Fig. 3, it is evident that the derived contour level shown in Fig. 4 is consistent with the relationship between the Philippine Sea Plate and the Eurasian Continental Plate.

![Fig. 1. The scatter and three-dimensional plots of 12 disastrous earthquake data.](image)

Table 2
The results of the illustrative example by data construction method (DCM) algorithm

| $t$ | $|Z'|$ | $(x'_{\text{in}} - 2 \times s', x'_{\text{in}} + 2 \times s')$ of the 1st | $P\{\|z_{\text{in}} - x'_{\text{in}}\| \leq 2s'\}$ | $P\{\|z_{\text{in}} - x'_{\text{in}}\| \leq 2s'\} = 0.95$ |
|-----|------|---------------------------------------------------------------|------------------|------------------|
| 1   | 24   | $(-0.85007, 0.65007)$, $(-0.85007, 0.65007)$                  | 0.9166           | No                |
|     |      | $(-1.39869, 0.59870)$, $(-1.39869, 0.59870)$                  | 0.9166           |                  |
| 2   | 48   | $(-0.66381, 0.51381)$, $(-1.10682, 0.50682)$                  | 0.9166           | No                |
|     |      | $(-0.66381, 0.51381)$, $(-1.10682, 0.50682)$                  | 0.9375           |                  |
| 3   | 96   | $(-0.52082, 0.40832)$, $(-0.87715, 0.42715)$                  | 0.9479           | No                |
|     |      | $(-0.52082, 0.40832)$, $(-0.87715, 0.42715)$                  | 0.9479           |                  |
| 4   | 192  | $(-0.40959, 0.32521)$, $(-0.69520, 0.35770)$                  | 0.9531           | Yes               |
|     |      | $(-0.40959, 0.32521)$, $(-0.69520, 0.35770)$                  | 0.9531           |                  |

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies
Plate, and that the energy resultant from the plate movement diffuses inwards and decreases progressively.

3. Functional estimation

To reveal the properties of the data set, its distribution function is desired. This section will demonstrate how to use the resultant sample data in order to estimate a function, and its validity will be evaluated.

First, the two properties of the resultant data set by DCM are summarized as follows.

Property 3. The distribution of \( \{ (z'_q, f_q) \mid 1 \leq q \leq Q \} \) is multimodal (Huang and Wang, 2007).

Therefore, the resultant data distribution includes more than one local maximum in density (Silverman, 1981; Li and Lin, 2006).

Property 4. The generated virtual samples may contain noise.
This property is a general drawback of re-sampling procedures (Efron, 1993). Although our proposed procedure is not a kind of re-sampling, this situation still remains. Therefore, eliminating the noises will be aimed for when the function estimation procedure is developed as below.

3.1. Procedure of density function estimation

Given an amount of virtual samples \( \{(z'_{q,t}, f_{q,t})|1 \leq q \leq Q\} \) after \( t = T \) multiset divisions, suppose \( M \) modalities were found with \( 2 \leq M \leq |Z^T| \), denote each modality as \( \{(z_m, f_m)|z_m \in Z^T, f_m \in \} \)

Fig. 4. The contour graph of \( Z^H \) with map of Taiwan.
\[ \mathbb{N}, \ 1 \leq m \leq M \] with \( z_m < z_{m+1} \), then a function \( \Pi(x), x \in \mathbb{R} \) is calculated by using the least-squares method to fit \( \{z'_m, f'_m\}, \{z_m, f_m\}, \forall m \). Then, supposing \( \int_{-\infty}^{\infty} \Pi(x) dx = \theta \), a probability density function denoted by \( f_{DCM}(x) \) is defined by

\[
f_{DCM}(x) = \theta^{-1} \Pi(x).
\]

To show the validity of \( f_{DCM}(x) \), based on K–S test (Ross, 1987), a conventional approach of using empirical distribution function defined below will be adopted for evaluation.

**Definition 5.** Empirical distribution function (Ross, 1987).

Let \( x_1, x_2, \ldots, x_n \) denote a random sample from an unknown population. Then, the empirical distribution function \( F_n(x), -\infty < x < \infty \), is defined by

\[
F_n(x) = \frac{\text{number of } i : x_i \leq x}{n}.
\]

### 3.2. Case study

Based on plate movement theory, the movement of two plates, the Eurasian Continental Plate and the Philippine Sea Plate, along the edge of East Taiwan explains why earthquakes happen so often in Taiwan. The Philippine Sea Plate moves toward the Eurasian Continental Plate at a rate of 7–8 cm/year. At an uncertain cycle time, earthquakes will occur whenever the crust can no longer endure further stresses that are brought about by the movement of these two plates.

The example in Section 2.3 demonstrates the data trend of the locations of disastrous earthquakes in Taiwan. Owing to the division created by the Central Mountain and other different geological structures, East and West Taiwan must be considered as two seismic regions. However, because the eastern side is situated in the breach of the two plates, the resultant hazard is greater compared with that on the western side. Therefore, further investigation was carried out on the sub-regions of East Taiwan.

From the literature (Kim et al., 2004), three near-source regions in East Taiwan are shown in Fig. 5. For ease of investigation, we locate the first region with four vertices of (24.3°N, 121.6°E), (24.8°N, 121.6°E), (24.3°N, 122.3°E), and (24.8°N, 122.3°E); the second region with four vertices of (23.5°N, 121.4°E), (24.2°N, 121.4°E), (23.5°N, 122.7°E), and (24.2°N, 122.7°E); and the third region with four vertices of (21.7°N, 120.8°E), (23.4°N, 120.8°E), (21.7°N, 122.3°E), and (23.4°N, 122.3°E).

Following the same procedure of DCM as illustrated in Section 2.3, sufficient virtual samples are generated for each region. Because the latitude and longitude are in different scales, normalization is adopted and each region induces a virtual center using Step 2 of the DCM algorithm as the mode.

In practice, the virtual center means the location with the highest risk in the defined region because of its highest frequency of occurrence. According to the virtual centers, all data are translated into relative distances (one unit of \( (\Delta \text{longitude}^2 + \Delta \text{latitude}^2)^{0.5} \) is approximately 111.5 km). Therefore, finding the distribution patterns of the generated virtual samples is...
equivalent to finding the frequencies of disastrous earthquakes that happened within relative distances with respect to the virtual centers.

3.2.1. The first near-source region
From 1900 to 1999, 15 disastrous earthquakes happened within the first region (see Table 3). Using data from 1959 to 1999 as the training data of DCM, the modes for two coordinates are 24.4 and 122.1. Therefore, the virtual center of this region is (24.4°N, 122.1°E), which is also regarded as a high-risk location in practice.

While assigning (24.4°N, 122.1°E) as the virtual center as Mode1, \((A_1)_o = \{(24.4, 122.1), 2\}, ((24.5, 121.9), 1), ((24.4, 121.8), 1), ((24.6, 121.7), 1)\} can be translated into \(A_1 = \{((0, 0), 2), ((0.1, -0.2), 1), ((0, -0.3), 1), ((0.2, -0.4), 1), ((0.1, -0.18), 1), ((0.1, 0.16), 1)\}\). After setting \(c = 2, k = 2, \) and \(w = 0.95\), the division process is stopped at \(t = 2\), and \(Z_t = \{((0, 0), 8), ((0.1, -0.2), 3), ((0, -0.3), 1), ((0.2, -0.4), 1), ((0.05, -0.1), 3), ((0, -0.15), 2), ((0.025, -0.05), 1)\}\).
Finally, the virtual samples, \( Z \), were obtained.

Because \( \text{Mode}_1 \equiv (24.4, 122.1)^\circ \) was the virtual center of the first region, \( Z \) was translated into \( f \) by computing relative distances (km) to \( (24.4, 122.1)^\circ \). For instance, \( (24.5, 121.9) \) was translated into \( (24.93, 3) \) owing to \( ((24.5 - 24.4)^2 + (121.9 - 122.1)^2)^{0.5} \times 111.5 = 24.93 \). Based on Property 3 and equation (6), three local maximum in density of \( Z \), \( f \), were found, and \( f_{\text{DCM}}(x_{i1}) = 0.0384 \times \exp(-0.0384x_{i1}), 0 \leq x_{i1}, \) was calculated by fitting \( \{0, 8, (12.47, 3), (24.93, 3), (49.86, 1)\} \) with \( R^2 = 0.9258 \), where \( x_{i1} \) means the relative distances to \( (24.4, 122.1)^\circ \).

The method of least squares, also known as regression analysis, is used to model the numerical data by adjusting the parameters of the model to obtain optimal fit of the data. Different model forms such as linear, polynomial, and exponential functions have been considered, and based on the highest \( R^2 \) value, the best one is adopted to construct \( f_{\text{DCM}} \). In this case, the linear function with \( R^2 = 0.7501 \), the polynomial function by the power 2 with \( R^2 = 0.9018 \), and the exponential function with \( R^2 = 0.9258 \) were obtained. Therefore, the exponential function was adopted to describe \( f_{\text{DCM}} \) with higher \( R^2 \).

For comparison and illustration of the influence of different modes chosen, with prior information from Central Bureau, another location with \( \text{Mode}_1 \equiv (24.4, 122.1)^\circ \) was used. Then, \( (A_1)_0 = \{(24.4, 122.1), 2\}, (24.5, 121.9) \), \( (24.4, 122.1) \), \( (24.6, 121.7) \) were translated into \( A_1 = \{(0, 0.3), 2\}, (0.1, 0.1), (0, 0) \), \( (0.2, -0.1) \), \( (0.1, 0.12) \), \( (0.1, \)

\begin{table}
\centering
\caption{The disastrous earthquake data from 1900 to 1999 of region 1 \( \equiv (24.3^\circ, 121.6^\circ \text{E}), (24.8^\circ, 121.6^\circ \text{E}), (24.3^\circ, 122.3^\circ \text{E}), (24.8^\circ, 122.3^\circ \text{E}) \)}
\begin{tabular}{lllll}
\hline
Year & Latitude (\(^\circ \text{N}\)) & Longitude (\(^\circ \text{E}\)) & Relative distance (km) to \text{Mode}_1 = \{(24.4, 122.1)^\circ \} & Relative distance (km) to \text{Mode}_1 = \{(24.4, 121.8)^\circ \} \\
\hline
1901 & 24.7 & 121.8 & 47.31 & 33.45 \\
1909 & 24.4 & 121.8 & 33.45 & 0.00 \\
1918 & 24.6 & 121.9 & 31.54 & 24.93 \\
1922 & 24.5 & 122.2 & 15.77 & 45.97 \\
1922 & 24.6 & 122.3 & 31.54 & 60.04 \\
1922 & 24.6 & 122.3 & 31.54 & 60.04 \\
1922 & 24.6 & 122 & 24.93 & 31.54 \\
1922 & 24.6 & 122.1 & 22.30 & 40.20 \\
1923 & 24.6 & 122 & 24.93 & 31.54 \\
1923 & 24.5 & 121.8 & 35.26 & 11.15 \\
1963 & 24.4 & 122.1 & 0.00 & 33.45 \\
1963 & 24.5 & 121.9 & 24.93 & 15.77 \\
1967 & 24.4 & 122.1 & 0.00 & 33.45 \\
1994 & 24.4 & 121.8 & 33.45 & 0.00 \\
1995 & 24.6 & 121.7 & 49.86 & 24.93 \\
\hline
\end{tabular}
\end{table}

Data source: Central Geological Survey, MOEA, Taiwan, Republic of China.
Using the same values of $c = 2$, $k = 2$, and $w = 0.95$, the division process stopped at $t = 3$. In the same way, $f_{DCM}(x_{i1}) = 0.0294 \times \exp(-0.0294x_{1i})$, $0 \leq x_{1i}$; $R^2 = 0.8617$ was obtained where $x_{1i}$ means the relative distances to $(24.4^\circ N, 121.8^\circ E)$.

Because different centers result in different functions, the K–S is carried out to compare two cases with the empirical distribution function of 15 disastrous earthquake data from 1900 to 1999 in the first region as shown in Table 3.

After these data were translated into relative distances, $x_i$, $i = 1–15$, respectively according to two defined virtual centers, the K–S was applied to validate the derived functions with the hypothesis of K–S test stated below:

**$H_0$:** the cumulative distribution function of $f_{DCM}$ is equivalent to $F_{15}(x_i)$, whereas the alternative hypothesis is

**$H_1$:** the cumulative distribution function of $f_{DCM}$ is not equivalent to $F_{15}(x_i)$.

With the test statistic,

$$D = \max_{i} \left\{ \left| F_{15}(x_{i-1}) - \int_{0}^{x_i} f_{DCM} \right|, \left| F_{15}(x_i) - \int_{0}^{x_i} f_{DCM} \right| \right\}, \quad i = 1, 2, \ldots, 15,$$

the null hypothesis is rejected when $D$ is large.

Thus, by setting a significance level $\alpha = 0.05$, a threshold $d$ is given such that $\text{Probability}_{H_0}(D \geq d) = \alpha$. Whenever $D$ is larger than $d$, we will adopt the conclusion that $H_0$ is rejected. The results of K–S tests are listed in Tables 4 and 5.

From the test results, we can find that $f_{DCM}(1')$ is better than $f_{DCM}(1)$. It means that the rule in selecting the mode (deciding on the virtual center) in this study does not apply to the case wherein extra information about the mode is perceived in advance. If any other location in the first region is further verified with the highest risk by related geology studies, the virtual center will instead be redefined by $(24.4^\circ N, 121.8^\circ E)$ (Fig. 6).

| $x_i$ | $f_{DCM}$ | $F_{15}(x_i)$ | $|F_{15}(x_{i-1}) - f_{DCM}|$ | $|F_{15}(x_i) - f_{DCM}|$ | $D$ | $d$ | Test Result |
|------|-----------|--------------|----------------------------|----------------|-----|-----|-------------|
| 0.00 | 0.00      | 0.13         | 0.32                       | 0.13           |     |     |             |
| 15.77| 0.45      | 0.20         | 0.38                       | 0.25           |     |     |             |
| 22.30| 0.58      | 0.27         | 0.35                       | 0.31           |     | 0.38 | 0.304 Reject $H_0$ |
| 24.93| 0.62      | 0.47         | 0.24                       | 0.15           |     |     |             |
| 31.54| 0.70      | 0.67         | 0.06                       | 0.04           |     |     |             |
| 33.45| 0.72      | 0.80         | 0.06                       | 0.08           |     |     |             |
| 35.26| 0.74      | 0.87         | 0.03                       | 0.12           |     |     |             |
| 47.31| 0.84      | 0.93         | 0.08                       | 0.10           |     |     |             |
| 49.86| 0.85      | 1.00         | 0.15                       |                |     |     |             |
### 3.2.2. The second and third near-source regions

From 1900 to 1999, 24 and 15 disastrous earthquakes happened within the second and the third regions, respectively (see Tables 6 and 7). Using the data from 1959 to 1999 of 9 and 8, respectively, as the training data of DCM, the modes for the two coordinates shall be defined by 23.9 and 121.6 for the second region; those for the third region shall be defined by 22.1 and 121.2. Therefore, we assigned these two regions as the virtual centers of Mode 2[^1] (23.9°N, 121.6°E) and Mode 3[^1] (22.1°N, 121.2°E), respectively.

After setting \(k = 5, c = 2,\) and \(w = 0.95,\) the division processes of the second and third regions stopped at \(t = 4\) and \(t = 3,\) respectively.

In the same way, \(f_{DCM}(x_2) = 0.0235 \times \exp(-0.0235x_2),\) \(0 \leq x_2,\) with \(R^2 = 0.9264,\) and \(f_{DCM}(x_3) = 0.0071 \times \exp(-0.0071x_3),\) \(0 \leq x_3,\) with \(R^2 = 0.5837\) were obtained, where \(x_2\) and \(x_3\) represent the relative distances with respect to Mode 2 \(\equiv (23.9°N, 121.6°E)\) and Mode 3 \(\equiv (22.1°N, 121.2°E),\) respectively. Then, the testing data were translated into relative distances (see Tables 6 and 7).

---

**Table 5**

The result of Kolmogorov–Smirnov (K–S) test of the first near-source region with the virtual center Mode 1[^1] (24.4°N, 121.8°E)

| \(x_i\) | \(f_{DCM}(x_i)\) | \(F_{15}(x_i)\) | \(|F_{15}(x_{i-1}) - f_{DCM}\)| | \(|F_{15}(x_i) - f_{DCM}\) | \(D\) | \(d\) | Test Result |
|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | 0.13 | 0.15 | 0.13 | | | |
| 11.15 | 0.28 | 0.20 | 0.17 | 0.08 | | | |
| 15.77 | 0.37 | 0.27 | 0.25 | 0.10 | | | |
| 24.93 | 0.52 | 0.40 | 0.20 | 0.12 | | | |
| 31.54 | 0.60 | 0.53 | 0.09 | 0.07 | 0.25 | 0.304 | Do not reject \(H_0\) |
| 33.45 | 0.63 | 0.73 | 0.04 | 0.11 | | | |
| 40.20 | 0.69 | 0.80 | 0.06 | 0.11 | | | |
| 45.97 | 0.74 | 0.87 | 0.04 | 0.13 | | | |
| 60.04 | 0.83 | 1.00 | | | | | |

---

![Fig. 6. The graph of Kolmogorov–Smirnov test on the first near-source region.](image-url)

---


© 2009 The Authors.

Journal compilation © 2009 International Federation of Operational Research Societies
The K-S test was used to test the cumulative distributions of the derived functions with $z = 0.05$. The results show that $H_0$ is not rejected in the second region due to $D = 0.23 < 0.242 = \hat{d}$ (see Table 8 and Fig. 7), but it is rejected in the third region because of $D = 0.37 > 0.304 = \hat{d}$ (see Table 9).

Based on the result shown in Table 9, it is concluded that $H_0$ is rejected at $z = 0.05$. Although the reason for rejecting $H_0$ is not clear, the very wide area of the third near-source region can cause mis-estimation of the distribution from the training data (see Fig. 8). Therefore, we tried to use different $c$ values to run the DCM algorithm.

Based on Table 10, it can be observed that the division procedure is sensitive to the value of $c$. The training performance can be improved by adjusting the value of $c$. In this case, although adopting $c = 50$ helps in obtaining a better solution, setting a very large value for $c$ will result in an exactly opposite effect (Table 11; Fig. 9).

3.2.3. Summary and discussion
According to the previous sections, the disastrous earthquakes that took place from 1959 to 1999 were adopted as the source of training data for the DCM algorithm. There were 5, 9, and 8 data
with respect to the three defined near-source regions. The parameters of DCM were set by $k = 2$, $c = 2$, and $w = 0.95$ for initial data construction, and with further setting $c = 50$ for valid data generation of the third near-source region.

After 3, 4, and 4 division times, there were 40, 144, and 128 virtual samples obtained with respect to three regions, and the respective density functions were derived from least-squares fittings. In this study, linear function, polynomial function by the power 2, and exponential function were considered for distribution patterns of which the larger the value of $R^2$, the better is

### Table 7
The disastrous earthquake data from 1900 to 1999 of region 3 $\equiv (21.7^\circ N, 120.8^\circ E), (23.4^\circ N, 120.8^\circ E), (21.7^\circ N, 122.3^\circ E), (23.4^\circ N, 122.3^\circ E)$

<table>
<thead>
<tr>
<th>Year</th>
<th>Latitude (°N)</th>
<th>Longitude (°E)</th>
<th>Relative Distance (km) to Mode$_3$ = (22.1°N, 121.2°E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1923</td>
<td>22.8</td>
<td>121.1</td>
<td>78.84</td>
</tr>
<tr>
<td>1935</td>
<td>22.5</td>
<td>121.5</td>
<td>55.75</td>
</tr>
<tr>
<td>1936</td>
<td>22</td>
<td>121.2</td>
<td>11.15</td>
</tr>
<tr>
<td>1943</td>
<td>22.5</td>
<td>121.5</td>
<td>55.75</td>
</tr>
<tr>
<td>1951</td>
<td>23.1</td>
<td>121.2</td>
<td>111.50</td>
</tr>
<tr>
<td>1951</td>
<td>23.2</td>
<td>121.4</td>
<td>124.66</td>
</tr>
<tr>
<td>1955</td>
<td>21.8</td>
<td>120.9</td>
<td>47.31</td>
</tr>
<tr>
<td>1959</td>
<td>21.7</td>
<td>121.3</td>
<td>45.97</td>
</tr>
<tr>
<td>1959</td>
<td>22.3</td>
<td>121.2</td>
<td>22.30</td>
</tr>
<tr>
<td>1959</td>
<td>22.1</td>
<td>121.7</td>
<td>55.75</td>
</tr>
<tr>
<td>1959</td>
<td>22.1</td>
<td>121.2</td>
<td>0.00</td>
</tr>
<tr>
<td>1965</td>
<td>22.5</td>
<td>120.8</td>
<td>63.07</td>
</tr>
<tr>
<td>1972</td>
<td>22.5</td>
<td>122.3</td>
<td>130.51</td>
</tr>
<tr>
<td>1978</td>
<td>23.3</td>
<td>122.1</td>
<td>167.25</td>
</tr>
<tr>
<td>1992</td>
<td>23.1</td>
<td>121.4</td>
<td>113.71</td>
</tr>
</tbody>
</table>

Data source: Central Geological Survey, MOEA, Taiwan, Republic of China.

### Table 8
The result of Kolmogorov–Smirnov (K–S) test of the second near-source region

| $x_i$ | $\int_0^{x_i} f_{DCM}$ | $F_{24}(x_i)$ | $|F_{24}(x_{i-1}) - \int_0^{x_i} f_{DCM}|$ | $|F_{24}(x_i) - \int_0^{x_i} f_{DCM}|$ | $D$ | $d$ | Test result |
|-------|-----------------|--------------|-----------------------------------------|-----------------------------------------|-----|-----|-------------|
| 11.15 | 0.23            | 0.21         | 0.10                                    | 0.02                                    |     |     |             |
| 15.77 | 0.31            | 0.29         | 0.12                                    | 0.02                                    |     |     |             |
| 22.30 | 0.41            | 0.38         | 0.07                                    | 0.03                                    |     |     |             |
| 24.93 | 0.44            | 0.63         | 0.10                                    | 0.18                                    |     |     |             |
| 31.54 | 0.52            | 0.71         | 0.14                                    | 0.18                                    |     |     |             |
| 35.26 | 0.56            | 0.79         | 0.14                                    | 0.23                                    | 0.23| 0.242| Do not reject $H_0$ |
| 44.60 | 0.65            | 0.83         | 0.17                                    | 0.18                                    |     |     |             |
| 45.97 | 0.66            | 0.88         | 0.18                                    | 0.21                                    |     |     |             |
| 49.86 | 0.69            | 0.92         | 0.01                                    | 0.23                                    |     |     |             |
| 100.35| 0.91            | 0.96         | 0.01                                    | 0.05                                    |     |     |             |
| 127.13| 0.95            | 1.00         | 0.05                                    | 0.05                                    |     |     |             |
the fitting performance. Table 12 shows that the exponential function was best-fit to \( f_{\text{DCM}} \), which was further verified by K–S tests.

Similar to the example in Section 2.3, the spatial distributions of the disastrous earthquakes of three near-source regions in East Taiwan were depicted in Fig. 10 by mapping the generated virtual samples into 2D contour graphs. The shaded areas in the graph designated different frequencies; the darker the shade, the lower were the frequencies. Thus, the black zone represents the lowest density. According to the density distribution, the zone with a higher frequency indicated a relatively higher risk.

Finally, to test the prediction capability of the spatial distributions, five disastrous earthquakes that happened in East Taiwan from 2000 to 2006 were examined Table 13. It can be noted that among them, except for the disastrous earthquake that happened at (24.6°N, 121.9°E) on May 15, 2002, four exactly fell into the zone of the drawn spatial distribution with the highest frequency. Therefore, Fig. 10 has been able to provide useful information about the possible risk among three near-source regions.

![Graph of Kolmogorov–Smirnov test on the second near-source region.](image)

**Table 9**
The result of Kolmogorov–Smirnov (K–S) test of the third near-source region

| \( x_i \) | \( f_{\text{DCM}} \) | \( F_{15}(x_i) \) | \( |F_{15}(x_i) - \int_0^{x_i} f_{\text{DCM}}| \) | \( |F_{15}(x_i) - \int_0^{x_i} f_{\text{DCM}}| \) | \( D \) | \( d \) | Test result |
|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | 0.13 | 0.01 | 0.13 | | | |
| 22.30 | 0.15 | 0.20 | 0.08 | 0.05 | 0.37 | 0.304 | Reject \( H_0 \) |
| 45.97 | 0.28 | 0.33 | 0.01 | 0.05 | | | |
| 55.75 | 0.33 | 0.53 | 0.17 | 0.21 | | | |
| 63.07 | 0.36 | 0.73 | 0.18 | 0.37 | | | |
| 113.71 | 0.55 | 0.87 | 0.26 | 0.31 | | | |
| 130.51 | 0.60 | 0.93 | 0.24 | 0.33 | | | |
| 167.25 | 0.70 | 1.00 | 0.30 | | | |
4. Summary and conclusion

The collision between the Philippine Sea Plate and the Eurasian Continental Plate not only contributed to the birth of Taiwan but also explains why earthquakes frequently happen in Taiwan. Among the earthquakes, there are a few which are considered to be “disastrous earthquakes.” From 1900 to 1999, only 86 disastrous earthquakes occurred in Taiwan, and each one resulted in the heavy loss of human lives and goods. Owing to insufficient knowledge and data, assessment of the recurrence risk of disastrous earthquakes is difficult. Therefore, based on our proposed DCM, the generated data have shown to be valid in discovering the pattern of earthquakes’ distribution so that the frequencies of the occurrence of possible seismic epicenters can be predicted for risk assessment.

To illustrate the method, the properties and procedure of the DCM for virtual sample generation have been outlined. Then a numerical example including 12 disastrous earthquake data in Taiwan was given to illustrate the step-by-step procedure and to show the validity of the method by mapping the generated virtual samples into a contour graph. The observed trend of the generated virtual samples is geologically consistent with the actual phenomenon of the collision of the Philippine Sea Plate and the Eurasian Continental Plate. Finally, a case study that investigated a complete pattern of the disastrous earthquake distribution in East Taiwan was performed by defining three near-source regions and their respective virtual centers.
Table 11
The result of Kolmogorov–Smirnov (K–S) test of the third near-source region with $c = 50$

| $x_i$ | $\int_0^{x_i} f_{DCM}$ | $F_{15}(x_i)$ | $|F_{15}(x_i) - \int_0^{x_i} f_{DCM}|$ | $|F_{15}(x_i) - \int_0^{x_i} f_{DCM}|$ | $D$ | $d$ | Test result |
|------|-----------------|-------------|-----------------|-----------------|-----|-----|------------|
| 0.00 | 0.00            | 0.13        | 0.09            | 0.13            |     |     |            |
| 22.30| 0.22            | 0.20        | 0.21            | 0.02            |     |     |            |
| 45.97| 0.41            | 0.33        | 0.14            | 0.07            |     |     |            |
| 55.75| 0.47            | 0.53        | 0.02            | 0.06            | 0.22| 0.304| Do not reject $H_0$ |
| 63.07| 0.51            | 0.73        | 0.01            | 0.22            |     |     |            |
| 113.71| 0.73        | 0.87        | 0.09            | 0.14            |     |     |            |
| 130.51| 0.77        | 0.93        | 0.08            | 0.16            |     |     |            |
| 167.25| 0.85        | 1.00        | 0.15            |                 |     |     |            |

Fig. 9. The graph of Kolmogorov–Smirnov test on the third near-source region with $c = 50$. 

Table 12
The summary of $R^2$ value with different function forms

<table>
<thead>
<tr>
<th>Region</th>
<th>The parameters of DCM</th>
<th>Function form</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>$k = 2$, $w = 0.95$, and $c = 2$</td>
<td>Linear</td>
<td>0.4095</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Polynomial by the power 2</td>
<td>0.4306</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exponential</td>
<td>0.8617</td>
</tr>
<tr>
<td>Second</td>
<td>$k = 2$, $w = 0.95$, and $c = 2$</td>
<td>Linear</td>
<td>0.5903</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Polynomial by the power 2</td>
<td>0.7083</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exponential</td>
<td>0.9264</td>
</tr>
<tr>
<td>Third</td>
<td>$k = 2$, $w = 0.95$, and $c = 50$</td>
<td>Linear</td>
<td>0.2843</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Polynomial by the power 2</td>
<td>0.4846</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exponential</td>
<td>0.7948</td>
</tr>
</tbody>
</table>

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies
Using the disastrous earthquake data in East Taiwan from 1959 to 1999 as the training data of the DCM algorithm, 40, 144, and 128 virtual samples were obtained respectively in three regions. Based on the generated virtual samples, three probability density functions, $f_{DCM}(x)$ were derived. Using K–S statistics for the validity test ($\alpha = 0.05$), the results showed that $f_{0}^{v}$/$f_{DCM}$ were all

Fig. 10. The spatial distributions of disastrous earthquakes of three near-source regions in East Taiwan (●, the disastrous earthquakes from 2000 to 2006).

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies
equivalent to the empirical distribution functions, $F_n(x_i)$. This confirms the validity of the generated virtual samples in representing the patterns of three near-source regions in East Taiwan. Furthermore, to show the prediction capability of the spatial distributions, five disastrous earthquakes that happened in East Taiwan from 2000 to 2006 were examined. The results showed that among these disastrous earthquakes, four exactly fell within the high-density zone of the derived spatial distribution of disastrous earthquakes. Therefore, we may conclude that the spatial distribution derived by the generated virtual samples is able to provide useful information for risk assessment.

In summary, based on the results of this study, we conclude that in theory, DCM is an efficient tool in deriving a sufficient amount of virtual samples from a 2D small data set for density function estimation; and in practice, the resultant spatial distributions have been shown to be useful for the risk assessment of disastrous earthquakes.

Some may argue that the method proposed here is not rigorous in terms of normal accepted statistical methods. However, we make no claim that it is rigorous in this sense, but offer this method as a more intuitive approach. Further studies should therefore be undertaken to develop a systematic method of determining the value of $c$ as well as to scan the entire map shown in Fig. 10 for information on active faults. The purpose of this is to develop a prototype for wider use of inspection, which can in turn reduce the loss of lives and property.

**Acknowledgement**

The authors appreciate the constructive suggestions from anonymous referees, and also acknowledge the financial support from the National Science Council, ROC with project number NSC94-2213-E007-018.

**References**

Appendix A

A1. $|Z'| = 2^t \times |A|$.

(a) If $t = 1$ and 2, then

$$|Z^1| = |A| + \frac{|A|}{c} = |A| + |A| = 2 \times |A|, |Z^2| = |A| + \frac{|A|}{c} + \frac{|A|}{c^2} = 2^2 \times |A|.$$ 

(b) Assume $t = T$ is true, then $|Z^T| = 2^T \times |A|$ is true where

$$|Z^T| = |A| \times \left(1 + \frac{T}{1!} + \frac{(T - 1)T}{2!} + \cdots + \frac{T(T - 1) \cdots 1}{T!}\right).$$ 

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies
(c) As \( t = T + 1 \),
\[
Z^{T+1} \equiv A(;)C^{T+1} = \{(a_i, k_i), \{(c^{-1}a_i, k_i), \ldots, (c^{-1}a_i, k_i)\}, \{(c^{-2}a_i, k_i), \ldots, (c^{-2}a_i, k_i)\}, \ldots, (c^{-T}a_i, k_i)\}.
\]

Hence,
\[
|Z^{T+1}| = |A| \times \left(1 + \frac{T + 1}{1!} + \frac{(T + 1)T}{2!} + \cdots + \frac{(T + 1)T(T - 1)\cdots 1}{(T + 1)!}\right), \quad \text{and}
\]
\[
|Z^{T+1}| - |Z^T| = |A| \times \left(1 + \frac{T + 1}{1!} + \cdots + \frac{T(T - 1)\cdots 1}{T!}\right) - |A|
\]
\[
\times \left(1 + \frac{T}{1!} + \cdots + \frac{T(T - 1)\cdots 1}{T!}\right) = |A| \times \left(1 + \frac{T}{1!} + \frac{T(T - 1)}{2!} + \cdots + \frac{T(T - 1)\cdots 1}{T!}\right) = 2^T \times |A|
\]

Therefore, \(|Z^{T+1}| = 2^{T+1} \times |A|\). \( \blacksquare \)

A2. Sup\((Z^t) = a_n\) and inf\((Z^t) = a_1, \forall t \in \mathbb{N}\).

(a) \( Z^1 \) can be shown as \( Z^1 \equiv A(;)C = \{(a_i, k_i) , (c^{-1}a_i, k_i)\} \). Because of \( 0 \in A \) and for all \( i a_i < a_{i+1} \), regardless of \( k_i \),
\[
\sup(Z^1) = \max_{1 \leq i \leq n} \{a_i, c^{-1}a_i\} = a_n \quad \text{and} \quad \inf(Z^1) = \min_{1 \leq i \leq n} \{a_i, c^{-1}a_i\} = a_1.
\]

(b) \( Z^2 \equiv ((A(;)C)(;):C) = \{(a_i, k_i), (c^{-1}a_i, 2k_i), (c^{-2}a_i, k_i)\} \). Because of \( 0 \in A, \forall i, a_i < a_{i+1} \), and \( 1 < c < e^2 \),
\[
\sup(Z^2) = \max_{1 \leq i \leq n} \{a_i, c^{-1}a_i, c^{-2}a_i\} = a_n, \quad \inf(Z^2) = \min_{1 \leq i \leq n} \{a_i, c^{-1}a_i, c^{-2}a_i\} = a_1.
\]

(c) If \( t = T \) is assumed to be true, then \( \sup(Z_T) = a_n \) and \( \inf(Z_T) = a_1 \) and
\[
Z^T \equiv A(;)C^T = \left\{(a_i, k_i), (c^{-1}a_i, Tk_i), \left(c^{-2}a_i, \frac{T(T - 1)}{2!}k_i\right), \ldots, (c^{-T+1}a_i, Tk_i), (c^{-T}a_i, k_i)\right\}
\]
\[
= \left\{(c^{-1}a_i, \left(\frac{T}{i}\right)k_i)\text{)}\right\}.
\]

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies.
Then when $t = T+1$,
\[
Z^{T+1} = Z^T(C) = \left\{ (c^{-j}a_i, \left( \begin{array}{c} T_j \\ k_i \end{array} \right) \left( k_i \right) \left\{ 0 \right\} : (1,1), (c,1) \right\}
\]
\[
= \left\{ (c^{-j}a_i, \left( \begin{array}{c} T_j \\ k_i \end{array} \right) \left( c^{-j-1}a_i, \left( \begin{array}{c} T_j \\ k_i \end{array} \right) \left\{ 0 \right\} = \left\{ (c^{-j}a_i, \left( \begin{array}{c} T_j \\ k_i \end{array} \right) \left( c^{-j-1}a_i, \left( \begin{array}{c} T_j \\ k_i \end{array} \right) \left\{ 0 \right\} + 1 \right\}
\]

Because of $0 \in A$ and $1 < c < c^2 < \cdots < c^T < c^{T+1}$, $\sup(Z^{T+1}) = a_n$, inf $(Z^{T+1}) = a_1$. ■