A COMPARATIVE STUDY OF KARST SINKHOLE HAZARD MAPPING USING FREQUENCY RATIO AND ARTIFICIAL NEURAL NETWORK FOR EAST CENTRAL FLORIDA

YongJe Kim
Department of Civil, Environmental, and Construction Engineering, University of Central Florida, 4000 Central Florida Blvd, Orlando, FL 32816, USA, yongkim@knights.ucf.edu

Boo Hyun Nam
Department of Civil, Environmental, and Construction Engineering, University of Central Florida, 4000 Central Florida Blvd, Orlando, FL 32816, USA, boohyun.nam@ucf.edu (corresponding)

Abstract
Sinkholes are one of the most common geohazards occurring in East Central Florida (ECF). Identifying areas prone to sinkholes is vital for land use planning in the ECF area, and thus, sinkhole hazard mapping plays a critical role. The present study presents (1) sinkhole hazard maps of ECF by using frequency ratio (FR) and artificial neural network (ANN) models and (2) a validation and comparison of the performance of two models. An inventory map with a total of 757 sinkhole locations was prepared from Florida subsidence incident reports (FSIR). 70% (530 sinkholes) were randomly selected to calibrate the sinkhole hazard models, and the remaining 30% (227 sinkholes) were used for the model validation. Five sinkhole contributing factors were considered including age of sediment deposition, hydraulic head difference, groundwater recharge rate, overburden thickness, and proximity to karst features. The relationship between sinkhole occurrence and sinkhole contributing factors was investigated through a GIS-based statistical analysis.

Introduction
Karst topography occurs in terrains that contain distinctive landforms and hydrology created primarily from the dissolution of soluble bedrock. The soluble bedrock is typically carbonate rocks such as limestone, dolomite, and gypsum and common features of karst topography include sinkholes, springs, and caverns, etc. According to the US Geological Survey (USGS), karst topography makes up about 20 percent of the Nation’s land surface and extensive karst topography is found in the states of Florida, Texas, Alabama, Missouri, Kentucky, Tennessee, and Pennsylvania (Weary and Doctor, 2014).

A sinkhole is one of the most common and frequent natural geohazards in the karst terrain. It poses a threat not only to public safety but also to property, resources, and the environment. Sinkholes create at least $300 million in damages each year in the United States and the actual damage is probably much higher than this estimate as there is no formal nationwide tracking of damage costs (Weary, 2015). The Florida Office of Insurance Regulation reports that insurers received a total of 24,671 claims for sinkhole damage in Florida between 2006 and 2010 totaling $1.4 billion. The report shows the insurers’ expense has been rising with an increasing trend in both frequency and severity of sinkholes (FLOIR, 2010).

Considering that a number of factors contribute to sinkhole formation in karst areas, geologists and geotechnical engineers are faced with a difficult task when attempting to identify the most important factors that lead to sinkhole development. Past studies on sinkhole formation in Florida show that hydrogeological factors are strongly linked to its occurrence (Wilson and Beck, 1992; Tihansky, 1999; Xiao et al., 2016; Perez et al., 2017). These sinkhole-related factors include hydraulic head difference, groundwater recharge rate, overburden soil thickness, aquitard layer thickness, water table depth, and distance to karst features, as well as geological age (Kim et al., 2018, submitted).

Use of geospatial data and geographic information system (GIS) technologies enables sinkhole hazard assessment and mapping. Various methods have been proposed for geohazard mapping, and can be generally classified into two groups: qualitative methods and quantitative methods. The qualitative methods include field investigation procedures and/or overlaying of different factor maps and applying predetermined weights to develop a hazard map. The quantitative methods contain artificial intelligence techniques such as artificial neural networks (ANN) and fuzzy systems, statistical techniques and deterministic or probabilistic procedures (Bhardwaj and Venkatachalam, 2014).

This study aims to produce sinkhole hazard maps of East Central Florida (ECF), using frequency ratio (FR) and artificial neural network (ANN) methods using GIS as a
platform. In addition, an attempt was made to compare the two methods.

**Study Area**
Florida is one of the most sinkhole prone states in the United States due to its hydrogeology, geomorphologic characteristics, climate conditions, and human activities (e.g. groundwater pumping for drinking water and irrigation). According to the Florida Geological Survey (FGS), about 3,800 sinkholes have been reported in the state since 1954. The study area is defined as Central District established by Florida Department of Environmental Protection (FDEP). See Figure 1 for a detailed map showing the study area location.

The region is characterized by low-lying and flat land surface relatively with karst features such as sinkholes, springs, and caves. The geology of Florida is largely characterized by sedimentary rocks with no major igneous or metamorphic provinces. Limestone is the main bedrock in the study area, and an impervious clay layer overlies the bedrock. The hydrostratigraphic units of ECF consist broadly of a surficial aquifer system (SAS), intermediate aquifer system (IAS), and Floridan aquifer system (FAS), from top to bottom (Miller, 1986). Due to the combination of hydrogeological conditions in ECF, the area is exposed to numerous sinkhole hazards. In the period 1961 – 2017 (July), a total of 954 sinkholes have been recorded in the study area.

**Data Preparation**

**Sinkhole Inventory Map**
The preparation of sinkhole inventory maps that show the location of previously identified sinkholes should be the first step for sinkhole hazard analysis. Florida Subsidence Incident Reports by Florida Geological Survey (FGS) with GIS was utilized to prepare the inventory map and locate the sinkhole positions. After

---

**Figure 1.** Location of the study area and spatial distribution of the reported sinkholes.
identification and removal of incorrect data, a total of 757 sinkholes available for analysis were recognized, 530 (70%) of which were randomly selected for model calibration, and the remaining 227 (30%) were used for model validation. All data layers were transformed into raster format with a thirty-meter resolution.

Thematic Layers
In this study, five key sinkhole contributing factors including age of sediment deposition (i.e. epoch), hydraulic head difference, groundwater recharge rate, overburden thickness, and proximity to other karst features were taken into consideration based on previous studies, data availability, and the hydrogeological conditions of the ECF region (Kim and Nam, 2017). Analysis of sinkhole occurrence and geological age shows that certain categories of epochs are more subject to sinkhole occurrence (Kim et al., 2017). Hydraulic head difference provides a driving force to cause the down-washing of overburden soil particles into carbonate cavities and voids. Groundwater recharge accelerates soil erosion and facilitates soil structure raveling (Stewart and Parker, 1992). The effects of these hydrogeological factors on sinkhole occurrence in East Central Florida (ECF) were investigated and found to be significant (Xiao et al., 2016). Overburden soil thickness is an important factor since it has shown that karst sinkholes mainly occur at sites with the thickness of 25 meters or less (Drumm and Yang, 2005). Proximity to other karst features is also an important factor to be taken into account. Sinkhole frequency increases as distance to karst features, such as caves, springs and sinkholes, decreases (Kromhout, 2017). Each model parameter was divided into a number of classes (Figure 2).

Methods
Frequency Ratio (FR)
In general, it is assumed that sinkhole occurrence is determined by sinkhole-related factors and that future sinkhole events are likely to occur under similar conditions to past sinkhole events. With these assumptions, the frequency ratio (FR) method derives the spatial associations between sinkhole locations and each of the factors contributing sinkhole occurrence in the study area. The FR method has been widely used for geohazard mapping such as for landslides, earthquakes, and sinkholes.

The FR of each class within a certain sinkhole-related factor can be calculated by the ratio of a class' percent area of the total study area and its percent of the total number of sinkholes in the study area (Equation 1). This density-based method holds the principal of conditional probability, in which the relation analysis is the ratio the area where sinkholes occurred to the total area. Therefore, a value of 1 means an average value. If the value is greater than 1, there is a strong correlation, and lower than 1 means a weak correlation between sinkholes and factor classes (Lee and Pradhan, 2007)

\[
FR_{i,j} = \frac{N_{i,j}}{A_{i,j}} / \frac{N_T}{A_T}
\]

where \(FR_{i,j}\) is the frequency ratio for \(j\)th class of the factor \(i\), \(N_{i,j}\) is the number of sinkholes in \(j\)th class of the factor \(i\), \(N_T\) is the total number of sinkholes in the study area, \(A_{i,j}\) is the area of \(j\)th class of the factor \(i\), and \(A_T\) is the total area.

Then, to calculate the Sinkhole Hazard Index (SHI), FR values of each factor are summed (Equation 2). The SHI represents the relative risk of sinkhole occurrence based on past sinkhole data, in which the higher the value, the greater the risk is

\[
SHI = \sum_{i=1}^{n} FR_i
= FR_{age_dpstn} + FR_{head_diff} + FR_{rechrg_rate}
+ FR_{ovrbdn_thk} + FR_{prox_karst}
\]

where \(FR_i\) is the frequency ratio of each contributing factor \(i\), and \(n\) is the number of factors.

Artificial Neural Network (ANN)
An artificial neural network (ANN) is a computational information-processing model that imitates the neural system of the human brain. ANNs, with the capability of acquiring knowledge through learning and storing information within interneuron connections, can extract patterns and detect trends that are too complex to be found by conventional methods (Yilmaz, 2009). The goal of ANN is to find the optimal paths to connect the inputs and outputs of neurons together, and to use them to predict outputs for a given set of inputs. Therefore, there are two stages involved in using ANN for multi-source classification: a training stage and a classifying stage. Compared with other statistical analysis and techniques, the ANN model has many advantages and ability to handle imprecise and fuzzy data. Therefore, it is considered an efficient approach for geohazard assessment and mapping.

The most popular and widely used ANN architecture is multi-layer perception (MLP) network, comprising of an input layer, one or more hidden layers, and an output layer. Input data are fed through the hidden
Figure 2. Sinkhole contributing factor maps: (a) age of sediment deposition; (b) head difference; (c) recharge rate; (d) overburden thickness; (e) proximity to other karst features.
sinkhole locations) were selected at random to be used as reference dataset in the weight adjustment process (Figure 3). 15 sample location data are presented in Table 1.

In this study, the MATLAB Neural Network tool was used to train the ANN model. The learning rate was set at 0.02, and the initial weights were randomly selected between 0.1 to 0.9. The root mean square error (RMSE) goal for the stopping criterion was set to 0.01. Other parameters of the neural network were taken as 10,000 for epochs (or iterations), and 0.9 for momentum factor. The final weights between layers acquired during training of the neural network were used to predict sinkhole hazard.

Results and discussion

Frequency Ratio (FR)

The relationship between the sinkhole occurrence and each sinkhole contributing factor for the study area was determined by the FR and the results are presented in Table 2. FR values greater than 1.0 represents significant correlation and values less than 1.0 indicates no significant correlation with sinkhole occurrence. FR analysis of the age of sediment deposition indicates that sinkholes were observed to occur predominantly in the Eocene, Miocene, and Pliocene sediments. Head difference, recharge rate, and overburden thickness are found to have good association with the sinkhole occurrence. In general, FR values increases when the head difference and recharge rate increase, and the overburden thickness decreases. A large number of sinkholes are likely to occur in areas where the distance to other karst features is within ≤300 meters.

Sinkhole Hazard Index (SHI) was determined by summing FR values of each factor. SHI values of each layer that processes them to obtain the optimal output during training sessions. Each neuron in hidden and output layers processes its inputs by multiplying each input \(x_i\) by a corresponding weight \(w_i\), summing up the products (Equation 3), and then processing the sum (if that exceeds the neuron threshold, then the neuron is activated) using a nonlinear transfer function (Equation 4) to produce a result \(y_i\) (Polykretis et al., 2015).

\[
net = \sum_{i=0}^{n} w_i x_i \quad \text{Eq. 3}
\]

\[
y_i = f(net) \quad \text{Eq. 4}
\]

The proper weights for each input factor are learned gradually, and results are refined by adjusting the internal weights between neurons to reduce the difference between actual and target output values (i.e., errors). After a large number of iterations, at the end of the training stage, the neural network generates an appropriate model that can predict the target value correctly from given input values. A back-propagation (BP) algorithm is typically applied to train the network where the training process continues until the target error is achieved. After the completion of the training stage, the network is used as a feed-forward structure to produce a classification for the entire data set (Paola and Schowengerdt, 1995).

In this study, a three-layer feed-forward network trained by a back-propagation (BP) algorithm was selected to predict the distribution of sinkhole-prone areas. The input layer has 5 neurons (age of sediment deposition (i.e. epoch), head difference, recharge rate, overburden thickness, and proximity to other karst features), and the output layer has one neuron. In general, it is not easy to determine the number of hidden layers and the number of neurons in the hidden layer required for a particular classification problem. According to the Kolmogorov theorem (Kurkova, 1992), for a three-layer feed-forward neural network, the neuron number of the hidden layer is \(2n + 1\), if the input layer has \(n\) neurons, and the output layer has \(m\) neurons. Therefore, the hidden layer has \(2 \times 5 + 1 = 11\) neurons, and as a result, a three-layer system consisting of an input layer (5 neurons), one hidden layer (11 neurons) and an output layer (1 neuron) was used as a network structure of \(5 \times 11 \times 1\), with input data normalized within the range of 0.1-0.9 based on the Sinkhole Hazard Index (SHI).

The sinkhole inventory map was reclassified by assigning a value of 1 to the sinkhole location pixels and a value of 0 to the non-sinkhole location pixels. From these two classes (sinkhole and non-sinkhole), 954 training location samples (530 sinkhole locations and 424 non-sinkhole locations) were selected at random to be used as reference dataset in the weight adjustment process (Figure 3). 15 sample location data are presented in Table 1.

\[\text{Figure 3. Back propagation ANN architecture for sinkhole classification.}\]
Table 1. Normalized sinkhole sample data.

<table>
<thead>
<tr>
<th>Sample No</th>
<th>Age dpstn.</th>
<th>Head diff.</th>
<th>Rech. rate</th>
<th>Over. thk.</th>
<th>Prox. karst</th>
<th>Sinkhole occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2</td>
<td>0.9</td>
<td>0.6</td>
<td>0.9</td>
<td>0.9</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0.9</td>
<td>0.6</td>
<td>0.6</td>
<td>0.9</td>
<td>0.9</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0.2</td>
<td>0.9</td>
<td>0.2</td>
<td>0.5</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>0.6</td>
<td>0.6</td>
<td>0.8</td>
<td>0.9</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0.2</td>
<td>0.9</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0.9</td>
<td>0.4</td>
<td>0.6</td>
<td>0.5</td>
<td>0.9</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0.2</td>
<td>0.7</td>
<td>0.4</td>
<td>0.8</td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0.6</td>
<td>0.9</td>
<td>0.6</td>
<td>0.9</td>
<td>0.9</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0.2</td>
<td>0.8</td>
<td>0.4</td>
<td>0.8</td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>0.6</td>
<td>0.7</td>
<td>0.6</td>
<td>0.9</td>
<td>0.9</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>0.9</td>
<td>0.4</td>
<td>0.2</td>
<td>0.8</td>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0.9</td>
<td>0.4</td>
<td>0.6</td>
<td>0.9</td>
<td>0.4</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>0.6</td>
<td>0.1</td>
<td>0.6</td>
<td>0.9</td>
<td>0.4</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>0.8</td>
<td>0.6</td>
<td>0.6</td>
<td>0.9</td>
<td>0.3</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2. Normalized sinkhole sample data.
factor in each grid cell were calculated to construct the sinkhole hazard map of East Central Florida (Figure 4). For better visual interpretation, the map was classified into five categories (very low, low, moderate, high, and very high) by using the natural break method (Table 3). For the validation purpose, the remaining 227 (30%) sinkholes were projected on the GIS-based sinkhole map. 25.2% and 24.9% of the total area are found under very low and low hazard classes. Areas with moderate, high, and very high hazard represent 15.8%, 17.0%, and 17.2% of the total area, respectively. The percentages of the total sinkholes in very low, low, moderate, high, and very high hazard classes are 0.5%, 1.8%, 10.9%, 21.8%, and 65.0%, respectively.

Artificial Neural Network (ANN)

A back-propagation (BP) multi-layer artificial neural network with three layers and threshold of sigmoid function was carried out. The network structure is $5 \times 11 \times 1$, input, hidden, and output layers, respectively where the input data is Sinkhole Hazard Index (SHI). Parameters of the neural network were taken as 0.02 for learning rate, and 0.9 for momentum factor. One of the ANN outputs represents the weight of sinkhole contributing factors. The weights of factors were taken on average value to obtain the best result and used to update the SHI values.

As a result, proximity to other karst features had the highest weight, 1.612, meaning that it is the most influencing factor on sinkhole events in East Central Florida (ECF). The weights of recharge rate and the age of sediment deposition are 0.845 and 0.757, which are the second and third largest contributing factors to sinkhole formation, respectively. The weight of overburden thickness is 0.455, and head difference had the lowest weight of 0.158. Based on the result of ANN model, sinkhole hazard map of ECF was produced (Figure 5).

The sinkhole hazard map produced by ANN was also grouped into five classes using natural break method (Table 3). According to the model, 30.9% of the study area is exposed to a very low hazard, and 22.9%, 16.3%, 13.7%, and 16.2% occupies low, moderate, high, and very low, respectively. It is observed that 0% and 5.0% of the total sinkholes falls in the very low and low hazard classes, respectively. Moderate, high, and very high hazard classes represent 11.4%, 27.7%, and 55.9% of the sinkholes, respectively.

<table>
<thead>
<tr>
<th>Model</th>
<th>Sinkhole hazard class</th>
<th>Area (%)</th>
<th>Sinkhole (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>Very low</td>
<td>25.2</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>24.9</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>Moderate</td>
<td>15.8</td>
<td>10.9</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>17.0</td>
<td>21.8</td>
</tr>
<tr>
<td></td>
<td>Very high</td>
<td>17.2</td>
<td>65.0</td>
</tr>
<tr>
<td>ANN</td>
<td>Very low</td>
<td>30.9</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>22.9</td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td>Moderate</td>
<td>16.3</td>
<td>11.4</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>13.7</td>
<td>27.7</td>
</tr>
<tr>
<td></td>
<td>Very high</td>
<td>16.2</td>
<td>55.9</td>
</tr>
</tbody>
</table>

Table 3. Comparison of predicted sinkhole hazard class and observed sinkholes.
Figures 4 and 5 show that the sinkhole hazard maps of FR and ANN are slightly different in the spatial distribution of hazard classes, although both methods are similar in the percentage of each hazard class in relation to the study area. According to the FR hazard map, the very low to low hazard class and the high to very high hazard class show 50.1% and 34.2% of the study area, respectively. Similarly, from the ANN hazard map, 53.8% and 29.9% of the entire area are found to be of the very low to low hazard class, and the high to very high hazard class, respectively.

About 86.8% and 83.6% of the sinkholes fall into the high and very high classes of the hazard map by FR and ANN, respectively. The results show that the FR method tends to be more conservative as it produces less low hazard, and more high hazard zones compared to the ANN method. Considering the results of this study, both FR and ANN methods are adequate for producing a regional scale sinkhole hazard map. The FR method has advantages over the ANN method, especially in terms of simplicity and time efficiency. However, it is difficult to analyze the correlation between factors that can be overcome by the ANN method.

Conclusions
In this study, two mathematical models that are frequency ratio (FR) and artificial neural network (ANN) methods were used to construct sinkhole hazard maps, and the susceptible areas for East Central Florida (ECF) are identified. Five key sinkhole contributing factors were selected based on the findings from relevant studies and data availability of the study area, which includes age of sediment deposition (i.e. epoch), hydraulic head difference, groundwater recharge rate, overburden soil thickness, and proximity to other karst features. Both hazard maps were divided into five classes by the natural break classification method (very low, low, moderate, high, and very high). According to the results, both FR and ANN models show a similarly good performance. High percentages of test sinkholes, with 86.8% and 83.6% for FR and ANN respectively, fall in the classes of high and very high. The results of this study can be used to assist local authorities and decision makers for proper site selection and planning. It is important to note that both FR and ANN models in the present study were developed under the assumption of no seasonal effects (e.g. extreme weather events such as drought or heavy rain associated with tropical storms or hurricanes). A sinkhole hazard map accounting for the seasonal effect is currently under development.

References


