Geographical Analysis of regression model with application to Sinkhole occurrence

1. Introduction

A sinkhole is a hole formed when the upper layer of soil collapses due to a void created beneath the ground. In South Korea, frequent sinkholes have caused significant casualties and property damage, yet research remains limited. Internationally, sinkholes often result from the dissolution of limestone by groundwater, and extensive studies have been conducted on this phenomenon. However, since limestone areas in South Korea are mostly limited to the Gangwon Province, directly applying such findings is challenging. Additionally, prior studies have proposed installing Internet of Things(IoT) devices in areas susceptible to sinkhole formation, considering factors such as water leakage from supply and drainage systems or groundwater outflow. However, installing IoT devices in all regions is impractical.

2. Objectives

The goal is to predict the probability of sinkhole occurrences in order to prepare for and mitigate human casualties and property loss caused by sinkholes. However, the primary causes of sinkhole formation are likely to vary by region, making it challenging to analyze the entire country. Therefore, this study aims to predict the probability of sinkhole occurrences for each district in Seoul, the most densely populated city in South Korea. To develop a sinkhole occurrence probability prediction model that incorporates the artificial and environmental characteristics of Seoul, which have not been addressed in previous studies. Through this, the installation of Internet of Things(IoT) devices can be prioritized in high-risk districts.

3. Data

- **467** dongs in Seoul divided into districts(Eup, Myeon, Dong) based on the legal boundary
- Sinkhole accident data was obtained from the 'Underground Safety Information System', and a total of 83 accidents existed
- Sinkhole accident data from January 2020 to September 2024
- Flooding traces, traffic volume, building density, population density, presence of subway stations, and groundwater outflow were adopted as variables expected to affect sinkhole accidents. And there are three types of groundwater outflow: subway groundwater outflow, building groundwater outflow, and communication port groundwater outflow.

4. Analysis

A 467×467 distance matrix was constructed based on the administrative boundaries of Seoul's districts (Eup, Myeon, Dong). The matrix was defined such that if the centroid of a district lies within a 2km radius or directly neighbors another district, the corresponding value is 1; otherwise, it is 0.

During the analysis, it was identified that the traffic volume variable contained two missing values. Given that traffic volume is likely associated with neighboring districts, Moran' s I statistic was calculated using the W matrix to assess spatial autocorrelation. Therefore, the missing values were imputed using the average traffic volume of the adjacent districts.

The distance matrix was then multiplied by the independent variables to create spatial interaction variables, which were incorporated as predictors in the regression models. Various models, including Poisson regression, logistic regression, and the Zero-Inflated Poisson(ZIP) model, were applied for analysis.

Model performance was evaluated using Recall, and Accuracy, with higher values indicating better performance. Since accurately identifying areas where sinkholes occur is of primary importance, Recall was considered the most critical evaluation metric.

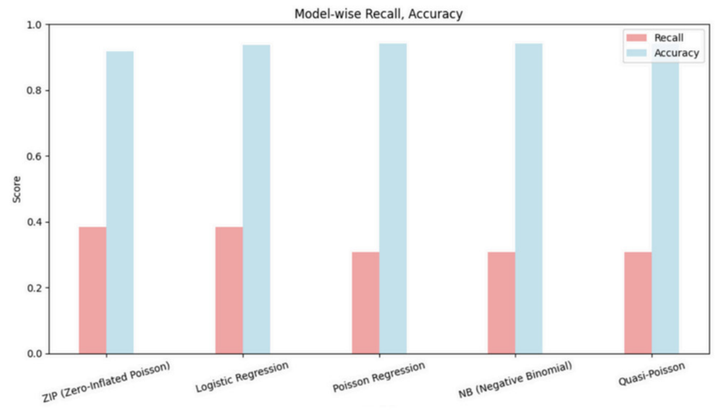
Actual Values \ Predictive Values	False	True
False	TN	FP
True	FN	TP

$$Recall = \frac{TP}{TP+FN}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

pmf of ZIP

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)e^{-\lambda_i} & \text{if } y_i = 0, \\ (1-\pi_i)\frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} & \text{if } y_i > 0. \end{cases}$$

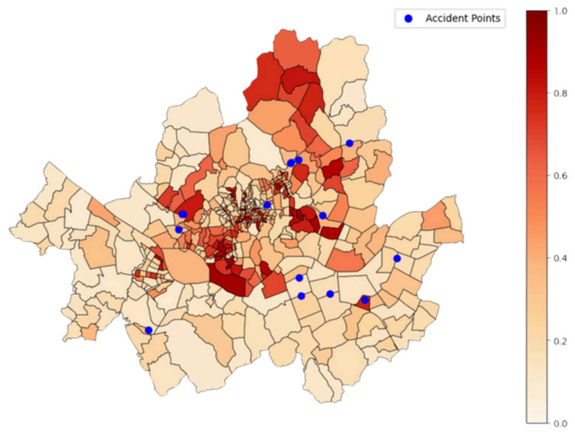


5. Result

This figure represents the result of smoothing using a weighted average, with a 7:3 ratio between Y and WY, where WY is the average of Y values from neighboring districts. The smoothing is applied to the results of the ZIP model. The 7:3 ratio was chosen because it yielded the best recall value during evaluation. The figure on the right visualizes the smoothing result, with blue dots representing sinkhole occurrences in 2024.

fitted model

$$\begin{cases} \log\left(\frac{P(Y_i=0)}{1-P(Y_i=0)}\right) = 0.0175 - 0.5034 \cdot \text{subway\_bi} - 0.1652 \cdot \log(\text{building\_den}) - 0.3915 \cdot \text{out\_num\_sum\_weighted} \\ \log(\lambda_i) = -1.4858 + 0.5790 \cdot \text{subway\_bi} + 0.1043 \cdot \log(\text{building\_den}) + 0.0611 \cdot \text{out\_num\_sum\_weighted} \end{cases}$$



Actual Values \ Predictive Values	False	True
False	327	127
True	3	10

Recall : 0.769

Accuracy : 0.721

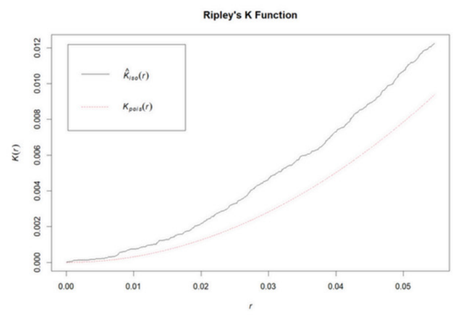
6. Conclusion

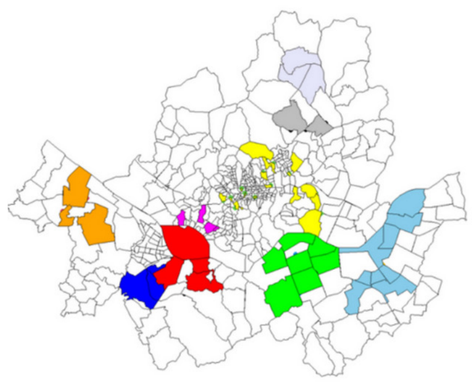
- The high-risk areas for sinkhole occurrence in Seoul identified by the ZIP model were Hwagok-dong, Jayang-dong, Yeoksam-dong, Geoyeo-dong, and Bangbae-dong.
- The risk of sinkhole occurrence increases with the presence of subway stations, higher building density, and higher frequency of groundwater leakage.
- Based on these results, it can be proposed to local governments to install Internet of Things (IoT) sensors in the above five areas for sinkhole detection.

7. Suggestions

By the primary causes of sinkholes, water leakage from underground water pipes was identified, but data could not be used due to restrictions from local governments. If data on the age of water supply and drainage pipes could be incorporated into the study, the effectiveness of the research could be further enhanced.

Since the sizes of Seoul's districts vary significantly, it may be necessary to adjust for differences in spatial proximity. Based on Ripley's K analysis, spatial clustering was confirmed, suggesting the need to expand the analysis scope in future research using a clustering-based approach.





8. Future work

Since sinkholes typically form as inverted cones, bowl shape or cylinder(Kim, 2017) and we have width, length and depth of each accident, the incident size of inverted cones can be expressed as:

$$\text{Sinkhole Volume} = \frac{1}{3} \times \left(\frac{\text{Width}}{2}\right) \times \left(\frac{\text{Length}}{2}\right) \times \text{Depth} \times \pi$$

A model predicting the scale of sinkhole incidents could be developed to alert local governments where the largest predicted sinkhole sizes are expected, urging them to take preventive measures.

Sources

SUNGHYOHYUN, & AHNSEJIN (2017). GIS Based Sinkhole Susceptibility Analysis in Karst Terrain: A Case Study of Samcheok-si.

Jae Heum Lee, Heyng Sub Lee, & In-Hwan Lee (2015). IoT-based Convergence Technology for Urban underground sinkhole Prediction.

Kim, C. (2017). A study on the causes and countermeasures for sinkholes in urban areas. Master's thesis, Graduate School of Construction and Industry, Kyonggi University, Gyeonggi-do, Republic of Korea.