CENTER FOR GEOSPATIAL DATA ANALYSIS
INDIANA UNIVERSITY

Prediction of potential areas of sinkhole development in southwestern Indiana using multiple regression analysis

Sally L. Letsinger, Ph.D. (sletsing@indiana.edu)[1], Greg A. Olyphant, Ph.D. (olyphant@indiana.edu)[2]

[1]Center for Geospatial Data Analysis, Indiana Geological Survey, Indiana University, Bloomington, IN, [2]Center for Geospatial Data Analysis, Department of Geological Sciences, Indiana University, Bloomington, IN

Abstract

Results of a multiple regression analysis, including prediction intervals, were used to develop a probability-based map of potential areas of sinkhole development in southwestern Indiana based on landscape-level variables as predictors. The dependent variable used in the analysis was the log of the sinkhole density (number of sinkholes/km^2), derived from an inventory of 154,925 sinkholes. The analysis was conducted for five large (8-digit scale Hydrologic Unit Code) watersheds in southwestern Indiana and northern Kentucky.

The study area was divided into regular units of analysis at a scale of a quarter section in the Public Land Survey System (160 acres or 0.65 km^2). For this analysis, 28,832 analysis cells were used. The conceptual model considers karst development to be more probable near specific geologic materials or features, especially in humid-temperate climates. The final best-fit equation accounted for 67% of the observed variability in the log sinkhole density and contained 14 statistically significant (99% confidence level) independent variables; five were related to lithology (including one interaction term, which combined geology and soils), two were land-use classes, and seven were related to terrain. Influential variables in the conceptual model were limestone, soil parent materials derived from limestone dissolution, and soil moisture and topographic parameters that might describe preferential flow to or from sinkhole features. The spatial distribution of potential sinkhole-development risk was mapped using 99% confidence intervals for the predicted log sinkhole density values.

The resultant distribution of sinkhole-development risk exists within a well-defined karst zone that trends from northwest to southeast along the Mitchell Plateau. The “extremely high” risk category has a high density of existing karst features, covering 2,617 km^2 (14%) of the study area. The “high" risk category borders the extremely high category, covering an area of 2,471 km^2 (13%). The “moderate" risk category (5,027 km^2, 27%) includes sinkholes developed in discontinuous limestone units as well as near faults where offsets have brought limestone closer to the surface. The “low" risk category covers 46% of the study area (8,554 km^2, 46%), and is dominated by noncarbonate bedrock geology, thick soil, or loess cover.

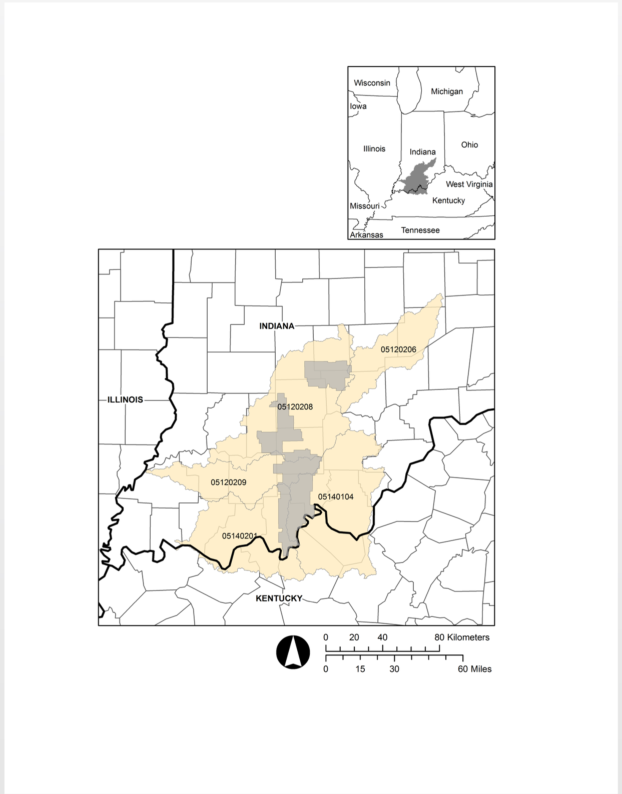

Figure 1. 8-digit watershed boundaries (HUC 8) encompassing the Hoosier National Forest (HNF) in southwestern Indiana. Watersheds are shown in yellow; HNF boundary shown in gray.

Conceptual Model

The regression analysis is based on a conceptual model that interprets karst development to be more probable in or near specific geologic materials or features, especially in humid-temperate climates, such as that of the study area. The independent variables that were hypothesized to most strongly influence sinkhole development include: limestone bedrock geology, soil parent materials that were derived from limestone dissolution, and soil moisture and topographic parameters that might describe preferential flow to or from sinkhole features. A small subset of sinkholes forms along fault lines in Indiana, so an effort was made to capture the variability related to structural offsets. The null hypothesis is that sinkholes are not related to landscape geologic or land-cover variables.The dependent variable in the study was the log of sinkhole density, which was determined after conducting an inventory of mapped and modeled sinkholes in southern Indiana and northern Kentucky in 2010 and 2011. The resultant database contained 154,925 sinkholes from which sinkhole density was determined.

The number of measureable contributing variables that can be captured in spatial (geographic information system) formats is substantial. For this study, the attributes that were considered are shown in Table 1, accompanied by source and scale information. For each variable evaluated, depending on the source scale of the original data, GIS data layers were created in vector or 30-meter raster formats. The data were aggregated by the regularized analysis cells either by conducting a spatial join to variables in vector formats, or by conducting a zonal statistical analysis to derive a mean value for variables in raster formats. For this analysis, 28,832 analysis cells were used.

Type	Variable	Source	Scale	Units	
Physical Characteristics	Catchment area	National hydrography dataset (NHD)	1 arc-second (30-m)	Sq kilometers	*
	Elevation	National Elevation Dataset (NED)	1 arc-second (30-m)	Meters	*
	Slope	NED derivative	1 arc-second (30-m)	Gradient (dimensionless)	
	Aspect	NED derivative	1 arc-second (30-m)	degrees	*
	Curvature	NED derivative	1 arc-second (30-m)	dimensionless	*
	Compound Topographic Index (CTI)	NED derivative	1 arc-second (30-m)	dimensionless	*
	Direct insolation duration	NED derivative	1 arc-second (30-m)	Watt hours per square meter (WH/m2) - hours	*
	Flow accumulation	NED derivative	1 arc-second (30-m)	Count	
	Local contributing area	NED derivative	1 arc-second (30-m)	Sq meters	
	Floodplain proximity	Federal Emergency Management Agency (FEMA)	500K	Meters	
Hydrologic	Hydrologic variables (modeled) • Recharge • Baseflow Index • Runoff	United States Geological Survey (USGS) SPARROW model	1 arc-second (30-m)	• mm/year • % • in/year	
Geologic	Geology	Indiana Geological Survey (IGS) and Kentucky Geological Survey (KGS)	250K (Indiana); 500K (Kentucky)	---	*
	Karst Proximity	IGS/KGS geology derivative	250K (Indiana); 500K (Kentucky)	Meters	*
	Carbonate	IGS/KGS geology derivative	250K (Indiana); 500K (Kentucky)	Percent	*
	Depth to limestone	IGS/KGS geology and NED derivative	250K (Indiana); 500K (Kentucky)	Meters	
	Fault proximity	United States Geological Survey (USGS)	12K to 500K	Meters	
	Geology and Soil *	IGS and NRCS derivative	12K – 500K	---	*
Soils	Soil Parent Material	Natural Resources Conservation Service (NRCS) SSURGO	12K – 20K	---	*
	Available water storage 150cm	Natural Resources Conservation Service (NRCS) SSURGO	12K – 20K	---	
	Soil proximity	NRCS derivative	12K – 20K	Meters	*
Land Cover	Land cover (15 classes evaluated)	National Land Cover Dataset (NLCD)	24K (30-m)	Area, sq meters	* Forest * Hay

Table 1. Variables considered. Only variables that demonstrated statistical significance (*) were used in the final analysis.

Statistical Model Background

The spatial data were prepared for input into a multivariate model, intended to investigate the predictive power of the landscape variables on sinkhole density, and therefore prediction of sinkhole-development risk. A multivariate method is appropriate for this type of analysis because errors (in this case, omissions) in the measurement of the dependent variable, sinkhole density, are probable, and the multiple regression is appropriate because random error computation is incorporated into the parameter estimates and significance tests.

The statistical regression model had the following form:

$$Log\ Y = b0 + b1\ x1 + b2\ x2 + b3\ x3 + b4\ x4 + \ldots bm\ xm + e$$

Where Y is sinkhole density, $x1 – xm$ are independent variables that are used to improve the prediction of sinkhole density and $b0 – bm$ are best-fit regression coefficients estimated using ordinary least squares, and e is an error term. The dependent variable was log-transformed because the range of observed density values covered several orders of magnitude, and the transformation provided a better model fit.

Multiple Regression Analysis

Each of the fourteen landscape variables was included in the regression analysis as an independent variable. The magnitude of the student's t statistic indicates the relative importance of each independent variable to the prediction of sinkhole density. The t values (Table 2) indicate that all of the variables were contributing significantly to the prediction at the 99% confidence level. This is not surprising given the unusually large sample size of 28,832 and the dependence of standard error on sample size. Figure 2 shows the independent variables in order of significance. The best-fit model combined variables that elucidated details from the geologic and soils data, with lesser influence from topographic and land-use variables. Although some variables have an equivalent effect on sinkhole development (*e.g.*, soil parent material and proximity to important soil parent materials), the correlation matrix does not show interdependence, and the predictive power of the regression equation is diminished when either variable is removed. Land-use and land-cover classes were included to identify any relationships to the log of sinkhole density; of 15 land-cover variables, only mixed forest classes and hay were significant. Two variables intended to capture sinkhole development processes around faults, proximity to mapped faults and depth to limestone, were not statistically significant.

Variable ID	Variable name	Parameter Estimate, b_i	Standard Error (SE)	t-ratio ($t=b_i/SE_{bi}$)
V_1	Catchment area	5.93E-03	1.84E-04	32.15
V_2	Elevation	-5.02E-03	2.06E-04	-24.38
V_3	Aspect	-2.04E-03	2.01E-04	-10.12
V_4	Curvature	3.13	4.41E-01	7.08
V_5	CTI	9.26E-02	4.23E-03	21.89
V_6	Insolation	1.60E-03	8.82E-05	18.09
V_7	Carbonate	1.62E-02	3.35E-04	48.45
V_8	Geology (Karst)	2.80E-01	3.71E-02	7.56
V_9	Karst proximity	6.58E-05	1.56E-06	42.09
V_{10}	Soil parent	4.05E-01	2.50E-02	16.21
V_{11}	Geology & Soil	1.21E-01	2.56E-02	4.71
V_{12}	Soil proximity	-2.28E-05	4.81E-07	-47.39
V_{13}	Forest	6.16E-07	5.04E-08	12.23
V_{14}	Hay	6.61E-07	5.24E-08	12.60

Table 2. Parameter estimates, standard errors, and t-ratios. The null hypothesis that the regression parameter equals zero is rejected at the 99% confidence level when the calculated t-ratio exceeds the critical value of 2.58 for degrees of freedom > 1000. Total sample size, N = 28,832. R^2 = 0.67. Multiple correlation coefficient, R = 0.82.

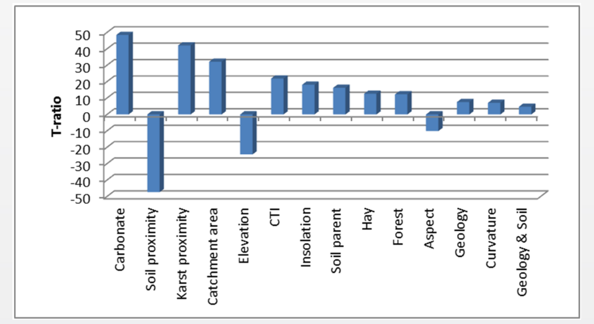

Figure 2. Computed t-ratios of significant parameters used in the multiple regression analysis.

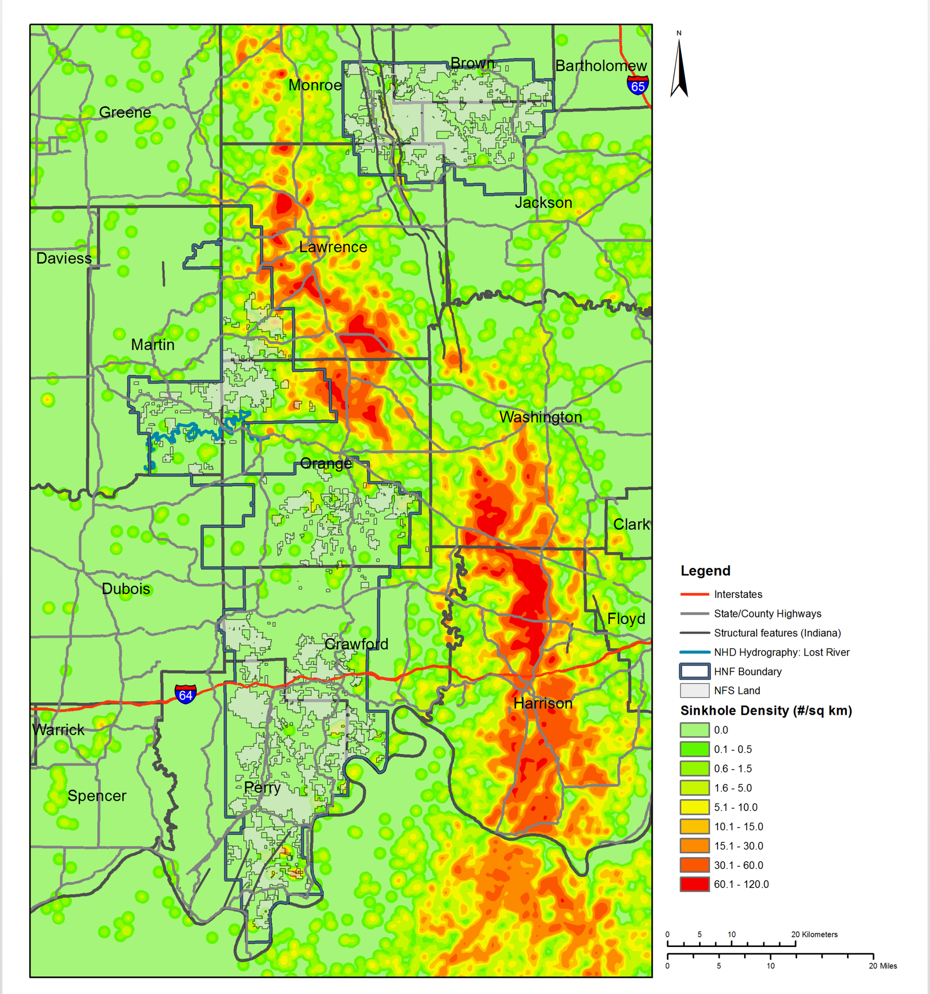

Figure 3. Sinkhole density in number of sinkholes per km^2. A kernal density function was employed with a 1000-m search radius.

Sinkhole development risk prediction and mapping

To map the results of the multiple regression analysis showing the spatial distribution of potential sinkhole-development risk, confidence intervals for the predicted log sinkhole density values were employed. This is because there is enough uncertainty in the analysis to make direct predictions of sinkhole-development locations questionable. The upper bounds of the 99% confidence intervals on sinkhole density predictions were classified into ranges by standard deviation. The resulting sinkhole-development risk categories are presented in Table 3. Note that roughly half of the study area is in low-risk areas and that half of the remaining area is in moderate risk of sinkhole development. Special attention should be given to the high and extremely high risk area in future planning activities.

The resultant map of sinkhole-development risk (Figure 4) shows a well-defined karst zone trending from northwest to southeast along the Mitchell Plateau. The “extremely high" risk category has a high density of existing karst features (sinkholes, springs, and caves) covering 2,617 km^2 (14%) of the study area. The area is characterized by bedrock geology that contains limestone units in the Mississippian Sanders and Blue River Groups. The most important karst-forming units are the Harrodsburg, Salem, St. Louis, Ste. Genevieve, and Paoli Limestone Formations. These units are dominantly composed of limestone. Above the bedrock surface in the "extremely high" risk category are soils made up of clayey limestone residuum, often with a very thin layer of overlying loess. Sinkhole features are most likely to be located in low elevations, large local catchments or watersheds, on northern slope aspects, and are characterized by forest or hay land-cover classes. The correlation of landscape variables with sinkhole development is not necessarily causal, however. The low elevations and large catchments could be an effect of karst development, not a cause. Similarly, it is likely that developed land-cover classes are not associated with karst areas because of the risk of doing so; therefore, more natural or low-risk land covers dominate karst topography. It is expected that new development of (or exposure of existing) sinkhole features is very likely in this category.

The “high" risk category surrounds the "extremely high" risk category, covering an area of 2,471 km^2. It is characterized by landscape variables similar to the "extremely high" risk category, but some of the most important characteristics are missing. For example, the sinkhole density is lower in this area because they are forming at the edge of the Sanders and Blue River Groups, at the eastern edge of the Mississippian West Baden and Stephensport Groups. These geologic groups have a lower proportion of limestone in the formations that compose the groups. In addition, the thickness of soils overlying these units is greater, with a greater thickness of overlying loess to impede downward water migration to accelerate dissolution. Sinkholes and springs are located in low-lying areas in large local catchments. As the overlying soils are eroded over time, sinkhole development (or exposure of existing, but buried karst features) will increase.

The “moderate" risk category (5,027 km^2) includes sinkholes developed in discontinuous limestone units as well as near faults where offsets have brought limestone closer to the surface. The “proximity to karst” and “proximity to important soils” variables are important in the moderate risk class because in some areas where predominantly clastic geologic units (*e.g.*, Mississippian Borden Group or Pennsylvanian Raccoon Creek Group) dominate the land surface, limestone units are not too far below the surface, or are exposed on hill slopes and in stream valleys. Similarly, locally eroded soils can provide exposures that can host sinkhole development. Springs are very common in the moderate risk category, perhaps a precursor to sinkhole development.

The “low" risk category covers 46% of the study area (8,554 km^2), and is dominated by non-carbonate bedrock geology, thick soils, soils derived from parent materials other than those from limestone units (*e.g.*, sandstone residuum, alluvium, lacustrine sediments), or have a thick loess cover. Although some sinkholes have been mapped (or identified) in the low-risk category, the density of features is very low and rapid expansion of sinkhole development is not expected.

Upper Limit of 99% Confidence Interval for prediction of sinkhole density, #/sq km	Standard Deviation Range	Area (km^2)	% Area	Risk Category
0.3 - 2.5	> 1.5	2,616.55	14%	Extremely high
-1.0 - 0.3	0.5– 1.50	2,471.51	13%	High
-2.4 - -1.0	-0.50 – 0.50	5,026.54	27%	Moderate
-3.6 - -2.4	< 0.5	8,554.12	46%	Low

Table 3. Confidence interval classification by standard deviation ranges to establish risk categories for potential sinkhole development in southwestern Indiana and northern Kentucky (total study area of 18,669 km^2).

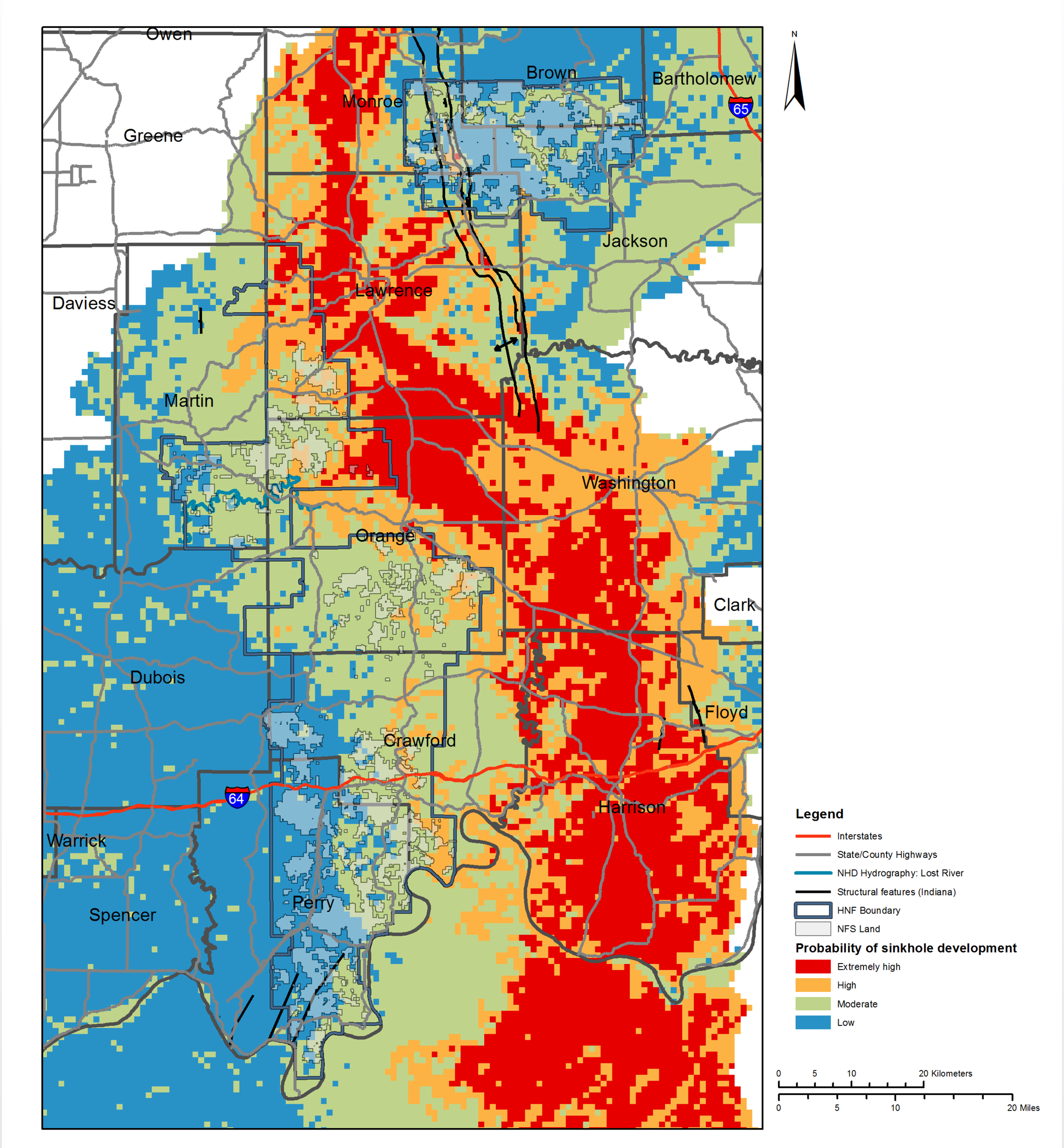

Figure 4. Probability of sinkhole development in southwestern Indiana and northern Kentucky. Probability based on 99% confidence intervals from a multiple regression analysis.

Conclusion

Although the multiple regression analysis did not explain all of the variability in the characteristics correlated to sinkhole development, roughly two-thirds of the variability in the spatial distribution of sinkholes can be explained by such a model using selected landscape-level variables. This suggests that the conceptual model of sinkhole development being controlled by geologic, topographic, and land-cover (including soils) factors is fundamentally correct at the scale of the analysis. Local or large-scale variables or processes might explain the remaining variability in the data. In addition to a more robust sinkhole inventory, future work to improve the analysis would involve constructing new, more spatially detailed data layers that could describe processes at the local catchment (field) scale.

References

Brock, T. D., Passman, F., and Yoder, I., 1973, Absence of obligately psychrophilic bacteria in constantly cold springs associated with caves in southern Indiana: American Midland Naturalist, vol. 90, no. 1., pp. 240-246.

Hughes, J. H. 1951, The geology of the Deer Creek Fault area, Perry County, Indiana: Bloomington, Ind., Indiana University, M.S. thesis, 31 p.

Taylor, C. J., and Nelson Jr., H. L., 2008, A compilation of provisional karst geospatial data for the Interior Low Plateaus physiographic region, central United States: U.S. Geological Survey Data Series 339, 26 p.