Hydrological Proximity Influences on Sink Area Coverage: A Geographically Weighted Regression Analysis of Perry County, MO

Emily A. Richardson

Western Illinois University

GIS 408: Environmental Applications of GIS

[redacted]

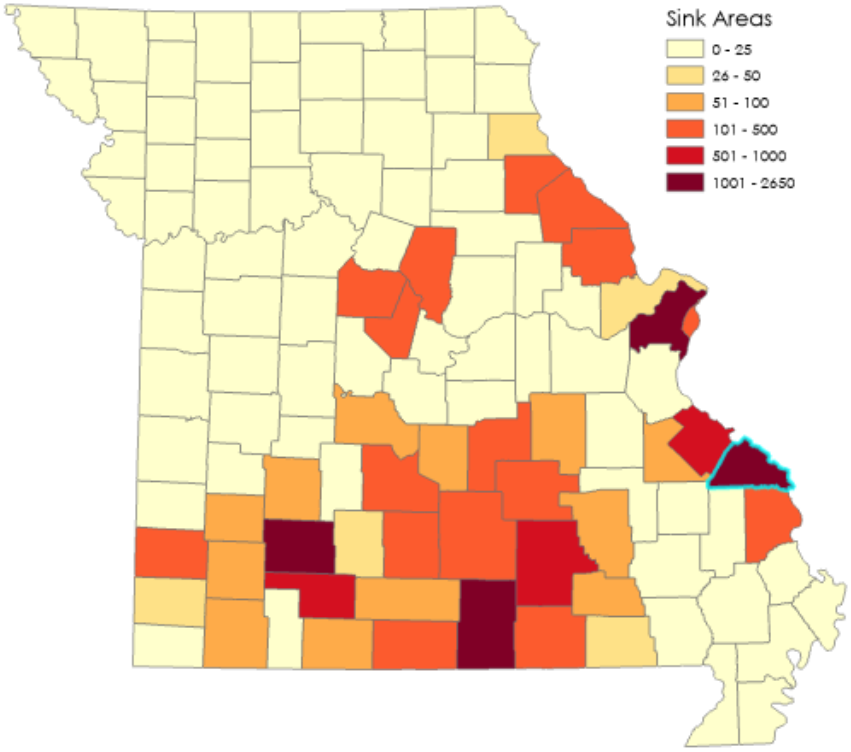
Hydrological Proximity Influences on Sink Area Coverage: A Geographically Weighted Regression Analysis of Perry County, MO

Missouri is known for its karst topography; caves, springs, natural bridges, and sinkholes characterize its landscape. As of 2018, the Missouri Department of Natural Resources estimated a total of over 14,000 sink areas in the state (Missouri Department of Natural Resources 2019). Of 114 counties as well as the independent city of St. Louis, only four had a total of over 1,000 sink areas: Perry, St. Louis, Greene, and Howell. Perry County had the highest estimated total at approximately 2,623, over 1,200 more than the next highest estimated total of 1,387 in St. Louis County. Of course, the high density of sink areas is expected due to Perry County being part of the Salem plateau region that comprises the Ozarks, which is classified as complex[1] and mature[2] karst. This and the area's adjacency to the Mississippi River make Perry County an ideal location for studying the influences of hydrological proximity on sink area coverage. According to the National Wetlands Inventory (U.S. Fish and Wildlife Service n.d.), approximately 24, 884, and 11 square kilometers are classified as wetlands, riverines, and ponds or lakes, respectively.

This study will investigate the influence of proximity of the three categories of hydrological features on sink area coverage within the study area of Perry County, MO. In this study, "wetlands" is a feature class that consists of both freshwater emergent wetlands and freshwater forested/shrub wetlands. The third category groups ponds and lakes together due to the small number of lakes in Perry County as well as the definition of both generally describing a still or standing body of water surrounded by land. The dependent variable, "sink area coverage", has been calculated as the percentage of a 1-kilometer surrounding area occupied by sink areas.

[1] Localized in temperate regions. Wide caves (most more than 15 feet wide), large dissolution sinkholes, scattered collapse features, and many subsidence and buried sinkholes are common (Obi 2016).
[2] "Common in temperate regions… Most fissures have extensive secondary openings with many caves. All kinds of sinkholes (subsidence, collapse and dissolution sinkholes) are found in numbers in the area" (Obi 2016).

This study seeks to determine if the influence of hydrological proximity (i.e., three independent variables) on sink area coverage (i.e., the dependent variable) varies in a statistically significant way throughout geographic space by creating a geographically weighted regression (GWR) model. This study will also evaluate the statistical significance and goodness-of-fit of the GWR model, use the GWR output to create a prediction map, and evaluate predictions by comparing them to observed sink area coverage values at validation points.



Map 1. Sink Area Distribution in Missouri Counties. Perry County is outlined in cyan.

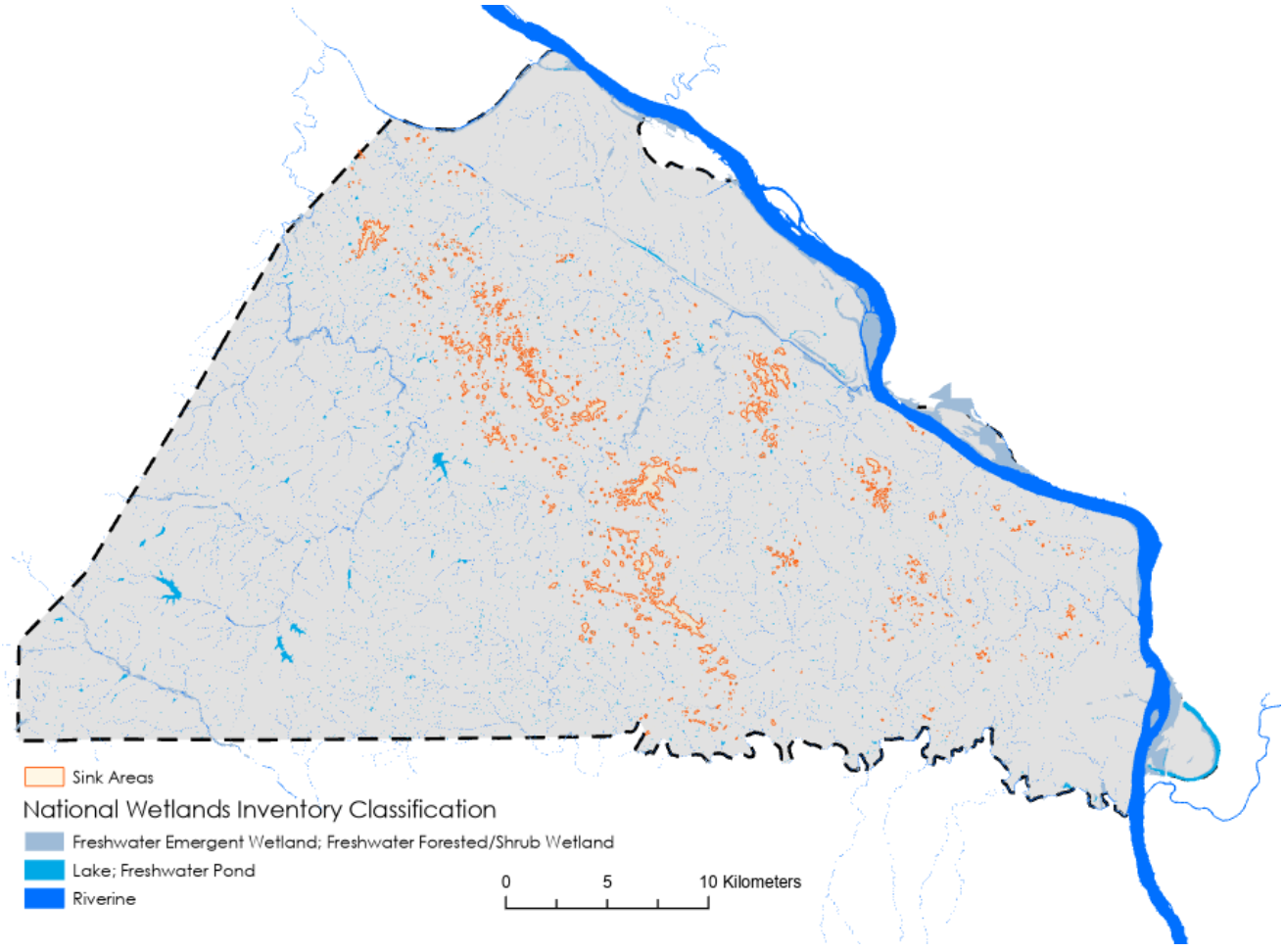
Data

This study utilizes the "MO 2018 Sink Areas" dataset from the Missouri Department of Natural Resources (2019) to calculate the dependent variable. Each of the three independent variables was calculated using the National Wetlands Inventory (NWI) Missouri dataset from the U.S. Fish and Wildlife Service (n.d.). State and county boundary shapefiles were obtained from the United States Census Bureau (2018). Lastly, a high-resolution (30-centimeter) imagery layer

was obtained from Esri's Living Atlas (Esri, Maxar, Earthstar Geographics, and the GIS User Community 2024). All downloaded data layers were projected to an Albers Equal Area Projected Coordinate System (PCS) using the Batch Project tool. (See Table 1.)

Field	Value
Projected Coordinate System	Perry County Equal Area
Projection	Albers
Authority	Custom
Linear Unit	Meters (1.0)
Central Meridian	-89.81.00278
Standard Parallel 1	37.58844444
Standard Parallel 2	37.88094444
Latitude of Origin	37.73469444
Geographic Coordinate System	NAD 1983

Table 1. Customized PCS coordinate system details.



Map 2. Sink Areas and Hydrological Features, Perry County.

Methodology

The following will summarize the steps of each task. GIS tasks were completed in ArcGIS Pro, while the statistical evaluations were conducted in Microsoft Excel. For specific tool parameters not mentioned in the discussion below, see the Appendix.

Dependent Variable Calculation

The downloaded sink area shapefile was a polyline feature class and this study requires polygons. Before converting the polyline outlines of sink areas to polygons, all primary sink areas were extracted using the Export Features tool, and a new feature class "PrimarySinks" was created. Primary sinks were classified in one of two categories: 1 (primary sink that contains no secondary sinks) or 1M (primary sink that contains at least one secondary sink). Note that primary sinks, and not secondary, tertiary, or quaternary sinks, are the focus of this study. This is because primary sink areas *contain* secondary, tertiary, and quaternary sinks. The complexity of incorporating these additional orders of sink areas is beyond the scope of this study. The export output was used to produce a polygon sink area layer in the Feature to Polygon tool. Unnecessary attribute fields were deleted from "PrimarySinks": (e.g., "HCOLLTEXT"[3], "HACCURACY"[4], "FLD_VERIFY"[5], and "D_SOURCE"[6]).

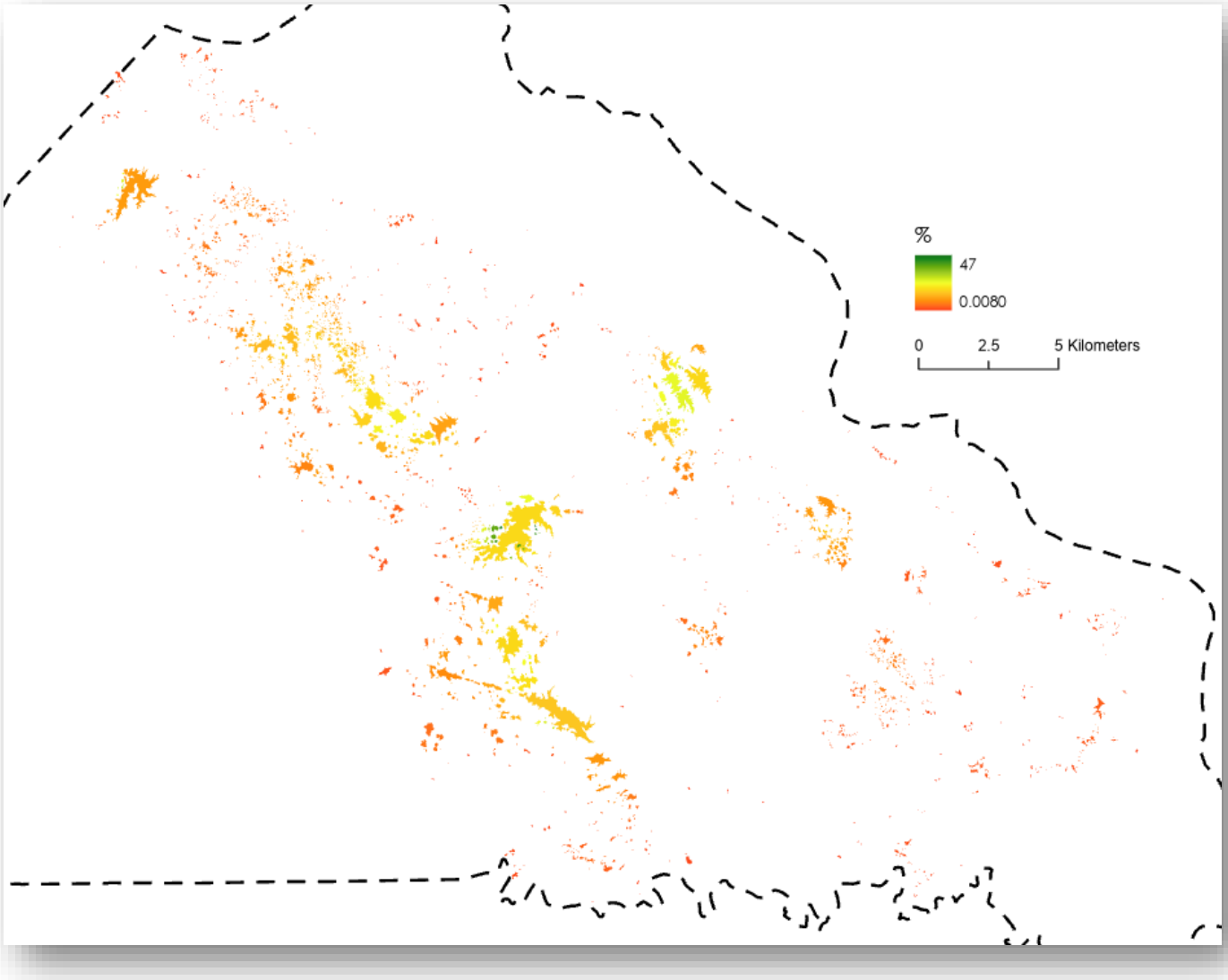
To calculate "sink area coverage percentage" (i.e., percentage of 1-kilometer surrounding area occupied by sink areas), a 1000-meter (1-kilometer) buffer was created around each sink area polygon using the Buffer tool. The 1-kilometer buffer was used for this calculation in order to allow for proper spatial variance of values for GWR calculations (Cahalan and Milewski

[3] The text that describes the method used to determine the coordinates for a point on the earth: Classical Surveying Techniques; GPS, Carrier Phase, Static Mode (SA Off); Interpolation from map; Interpolation from photography.
[4] The estimated accuracy (in meters) of the coordinates.
[5] A discriminator used to indicate whether or not the location has been verified by a field visit.
[6] The text that identifies the source of the data: Environmental Geology Field Map; Environmental Geology Miscellaneous Report; Sinkhole Recognition Project; Mobile GIS; or "Unknown".

2018). The Apportion Polygon tool was used to aggregate sink area within each buffer polygon; this value was assigned to each 1-km buffer feature. This calculation accounted for all sink areas, whether fully or partially contained within the buffer. Note that in the output attribute table, the total sink area calculation is the "Area" field. Via the Join Field tool, the total sink area value within each sink's 1-km buffer was assigned to each sink polygon. This could be accomplished easily due to the copy of the original FID field kept in the output feature class produced by the Apportion Polygon tool. Finally, a new field "buffer_area" was created and populated before making the final dependent variable calculation:

$$\frac{Total\ sink\ area\ within\ surrounding\ 1\ kilometer}{Total\ surrounding\ (i.e., buffer)\ area} \cdot 100$$

$$\Rightarrow Python\ expression: (!SinkAreaBuffer! \ / \ !buffer_{area}!) \cdot 100$$

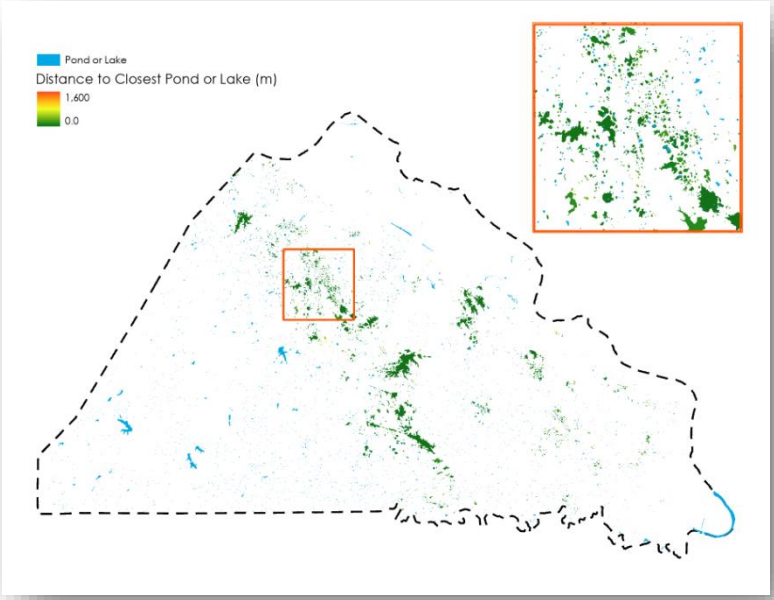


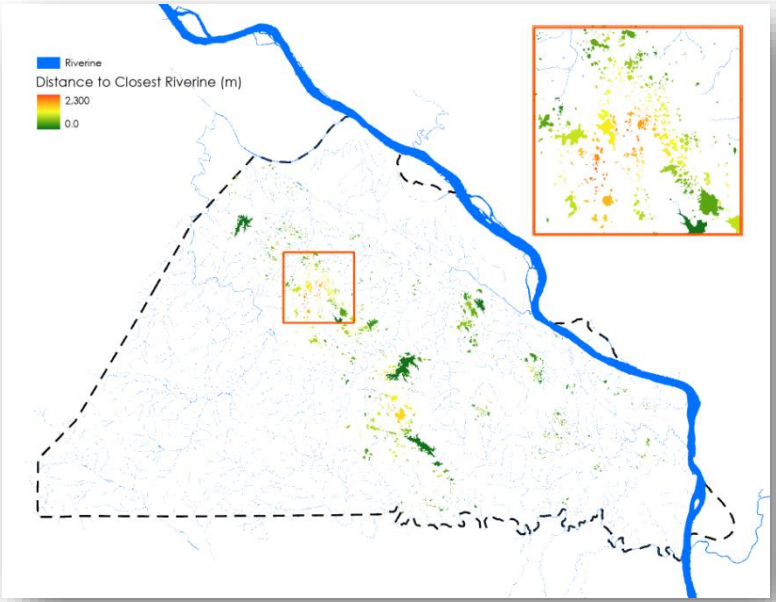
Map 3. Percentage of Surrounding Area Occupied by Sinks. (Dependent Variable)

Independent Variable Calculation

When compared to high-resolution imagery, it was revealed that the National Wetlands Inventory (NWI) was much more accurate in this rural county than the National Hydrography Dataset. Thus, three separate layers were exported from the original NWI layer using the Export Features tool: “PondsLakes”; “Riverines”; and “Wetlands”. The Near tool was used three times to calculate the three explanatory variables: proximity to ponds/lakes, riverines, and wetlands. Because of its relatively small size, Perry County is bested suited for the Planar method, which treats distances as straight lines on a flat surface, suitable for analysis at this scale while still

providing accurate results. Note that the appropriate distance values have field titles of "NEAR_PL" for ponds/lakes, "NEAR_R" for riverines, and "NEAR_W" for wetlands.





GWR Model

Recall a 1-kilometer buffer was used for the dependent variable calculation. In order to allow for proper spatial variance of values for GWR calculations and address the issue of spatial autocorrelation and dependence, the fixed GWR distance band will be greater than 1 kilometer. Choosing a neighborhood type of "distance band" and a "golden search" neighborhood selection method means the tool will identify an optimal distance based on the characteristics of the data. After running the tool, a warning stated, "The final model didn't have the lowest AICc encountered in the Golden Search Results." The GWR tool was run a second

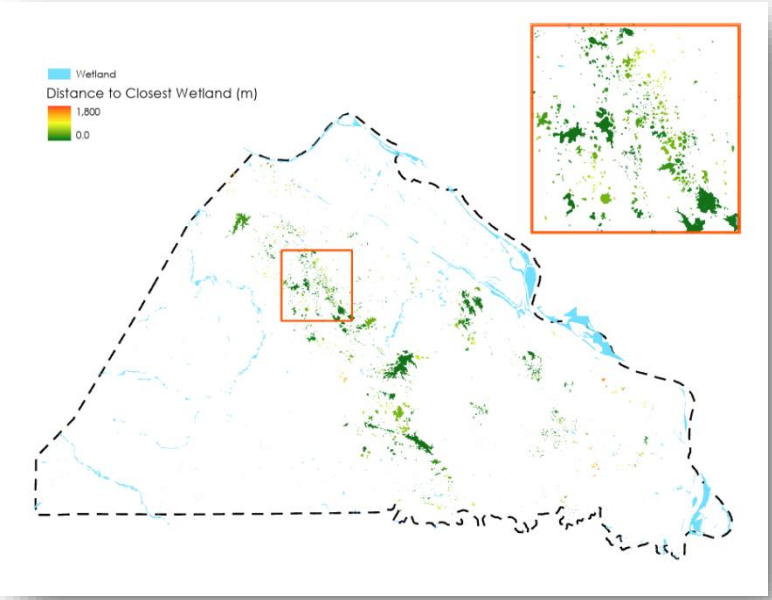
time using the lowest AICc, 13238.5059, which corresponds to a distance band of 4,166.4301 meters. This second run will provide a better fit to the observed data. See Table 2 for the results of the second GWR run.

Analysis Details	
Number of Features	2346
Dependent Variable	SINKPERCENTAGE
Explanatory Variables	NEAR_PL
	NEAR_R
	NEAR_W
Distance Band (Meters)	4166.4301

Model Diagnostics	
R2	0.6682
AdjR2	0.6470
AICc	13238.5059
Sigma-Squared	15.9306
Sigma-Squared MLE	14.9751
Effective Degrees of Freedom	2205.2922
Adjusted Critical Value of Pseudo-t Statistics	3.1202

Table 2. Second GWR run results.

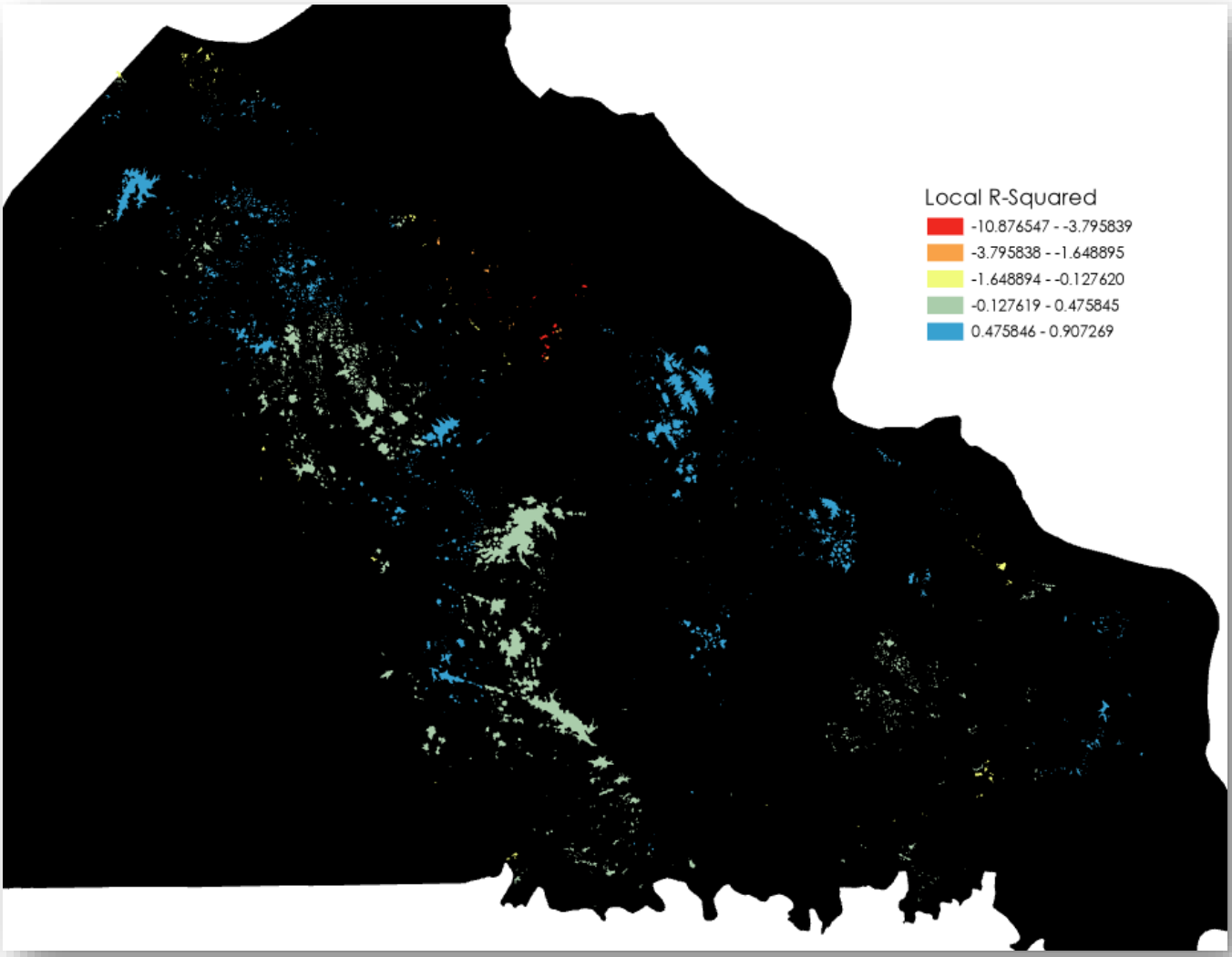
The first two diagnostics are the R-squared and adjusted R-squared values, which are both a measure of goodness of fit and may range "from 0.0 to 1.0, with higher values being preferable" (Esri n.d.). Also according to Esri documentation (n.d.), the adjusted R-squared value is a better measure because it considers the number of predictors in the model by normalizing the numerator and denominator by their degrees of freedom. However, the R-squared value reveals that the GWR model accounts for



approximately 67% of the variance of the dependent variable. This is a relatively good, or solid, value but it is clear that the model is lacking at least one significant explanatory variable.

Are the relationships between hydrological feature proximity and sink area coverage nonstationary (i.e., dynamic and change over space)? This can be answered by visually examining the local R-squared values for each sink area. With a maximum local R-squared value of 0.91 (very good—the model fits extremely well) and a minimum of -10.88 (very bad—the model does not fit well—this value should not go below 0[7]), it's clear that the strength of the relationships between the explanatory variables and the dependent variable vary greatly over space. The influence of hydrological features of sink area coverage is strongest in areas closest to the Mississippi River, Brazeau Creek in the southeastern part of the county, and Saline Creek in the northwestern part of the county. Areas with weaker relationships are along U.S. Route 61 and in and around Perryville, while the weakest relationships exist in near and along State Highway MO-E in (roughly) the northcentral part of the county.

[7] Because this is a weighted regression based on geography, negative numbers *are* possible.

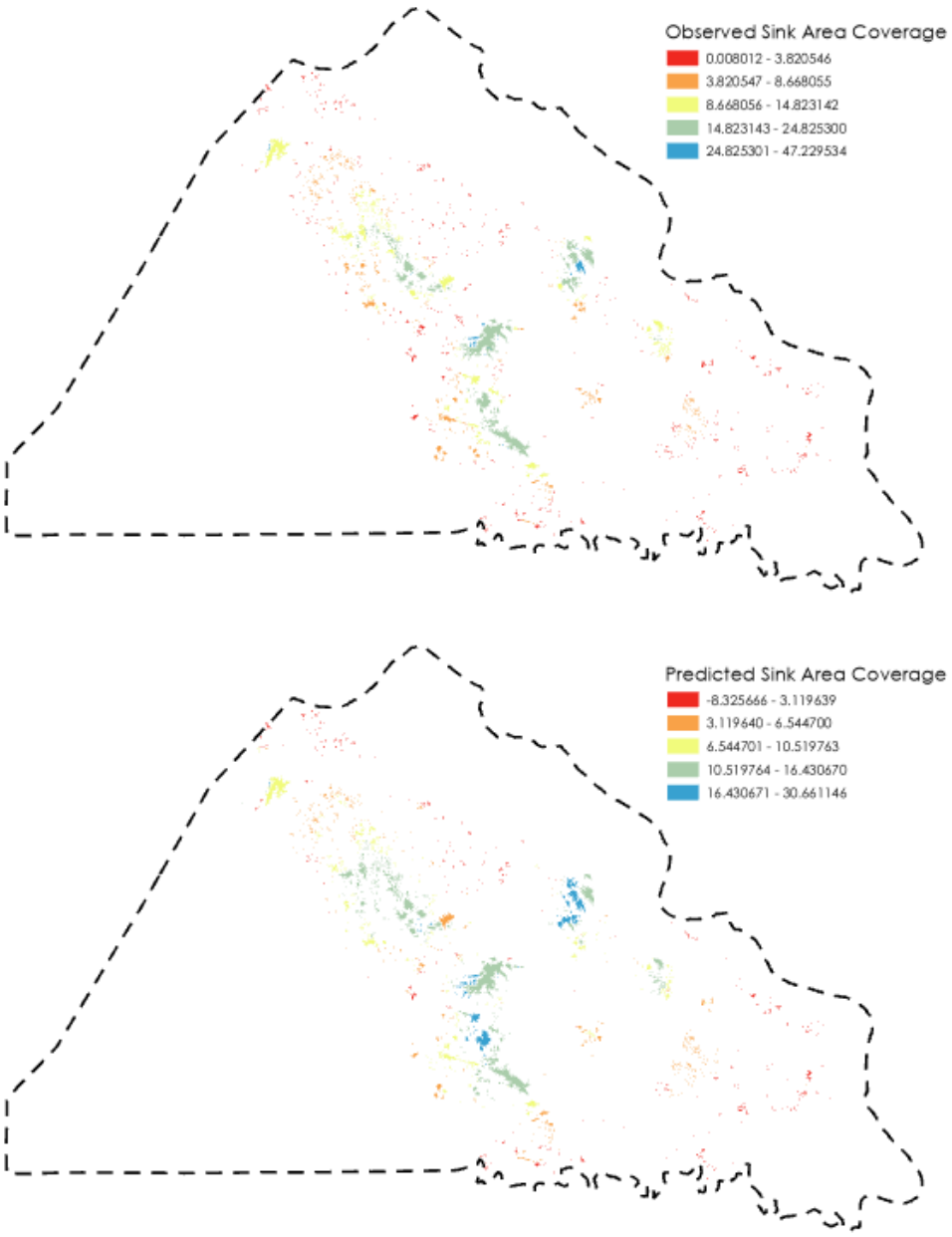


Map 4. GWR results: local R-squared comparison.

The GWR results indicate that, as expected, areas closer to hydrological features exhibit stronger relationships. This could potentially be due to increased groundwater flow and the potential for increased soluble rock dissolution. Areas closer to major roads and built-up areas (e.g., Perryville) experience a weaker relationship. Explanatory variables in these areas likely are related to human activity (e.g., diverting surface water or artificially creating ponds, mining, and other activities that influence hydrological processes and ground stability).

Predictions & Conclusions

The GWR tool produces a new field in the output feature class with the predicted values for the dependent variable. The original observed dependent variable value is retained in the new feature class. See Map 5 for a visual comparison of the measured and predicted sink area coverage values.



Map 5. Observed (top) and Predicted (bottom) sink area coverage values.

The Export Table to CSV tool made it possible to produce a CSV and calculate several error metrics to compare the predicted to the observed sink area coverage (dependent variable) values. Two model performance measures were calculated: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Both measure the accuracy of the GWR model by quantifying the difference between the observed and predicted values; in both cases, a smaller value indicates a better performing model. The RMSE is the square root of the average of the squared differences between the observed and predicted values; it is, at times, sensitive to outliers. The MAE is the average of the absolute values of the errors (i.e., differences); it is less sensitive to outliers. For this GWR model, the RMSE and MAE values are 3.87 and 2.5, respectively. This indicates that, overall, the predicted and observed values do not agree; the model may need to be refined. Additional variables will likely improve the accuracy of this model as well.

The influence of hydrological feature proximity on sink area coverage clearly varies throughout geographic space, which is to be expected. In some parts of Perry County, the model fits well, particularly in areas closest to the Mississippi River. In other parts, the model does not fit well and indicates a need for further study and additional explanatory variables, perhaps those related to human development and activity (e.g., proximity to roads or buildings).

References

Cahalan, M., and A. Milewski. 2018. Sinkhole Formation Mechanisms and Geostatistical-Based Prediction Analysis in a Mantled Karst Terrain. *CATENA.* 165:333–344. doi: 10.1016/j.catena.2018.02.010

Esri, Maxar, Earthstar Geographics, and the GIS User Community. 2024. High Resolution 30cm Imagery. https://hub.arcgis.com/datasets/esri::high-resolution-30cm-imagery (Accessed 17 April 2024).

Esri. (n.d.). How Geographically Weighted Regression (GWR) works—ArCGIS Pro | Documentation. https://pro.arcgis.com/en/pro-app/3.1/tool-reference/spatial-statistics/how-geographicallyweightedregression-works.htm

Missouri Department of Natural Resources. 2019. MO 2018 Sink Areas. https://data-msdis.opendata.arcgis.com/datasets/MSDIS::mo-2018-sink-areas (Accessed 16 April 2024).

Missouri Department of Natural Resources. 2019. MO 2018 Sink Areas. https://data-msdis.opendata.arcgis.com/datasets/MSDIS::mo-2018-sink-areas (Accessed 16 April 2024).

Obi, J. C. 2016. Geophysical Imaging of Karst Features in Missouri. https://scholarsmine.mst.edu/doctoral_dissertations/2485 (Accessed 2 May 2024).

U.S. Fish and Wildlife Service. National Wetlands Inventory (NWI) - Missouri. https://www.fws.gov/program/national-wetlands-inventory/data-download (Accessed 17 April 2024).

United States Census Bureau. 2018. 2018 TIGER/Line Shapefiles (machine readable data files). *U.S. Department of Commerce*. https://www.census.gov/cgi-bin/geo/shapefiles/index.php (Accessed 16 April 2024).

Appendix

Tool Parameters

Dependent Variable Calculation	
Export Features	
Input Features	SinkAreas
Output Feature Class	PrimarySinkPolylines
Filter (SQL expression)	S_ORDER = '1' Or S_ORDER = '1M'
Feature to Polygon	
Input Features	PrimarySinkPolylines
Output Feature Class	PrimarySinkholes
Buffer	
Input Features	SinkholePolygons
Output Feature Class	Sink_1000mBuffer
Distance	1000 meters
Side Type	Full
Method	Planar
Dissolve Type	No Dissolve
Apportion Polygon	
Input Polygons	SinkholePolygons
Fields to Apportion	Area (sq. m)
Target Polygons	Sink_1000mBuffer
Output Feature Class	Apportioned
Apportion Method	Area
Maintain target geometry?	Yes
Join Field	
Input Table	SinkholePolygons
Input Field	OBJECTID_1
Join Table	Apportioned
Transfer Method	Select transfer fields
Transfer Fields	“Total Sink Area within 1000 m”
Index Join Fields	Do not add indexes

Independent Variable Calculation	
Export Features	
Input Features	Wetlands
Output Feature Class	Run #1: “PondsLakes: Run #2: “Riverines” Run #3: “Wetlands2”
Filter (SQL expression):	Run #1: WETLAND_TY = 'Freshwater Pond' Or WETLAND_TY = 'Lake'

	Run #2: WETLAND_TY = 'Riverine' Run #3: WETLAND_TY = 'Freshwater Emergent Wetland' Or WETLAND_TY = 'Freshwater Forested/Shrub Wetland'
Near	
Input Features	SinkholePolygons
Near Features	Run #1: “PondsLakes” Run #2: “Riverines” Run #3: “Wetlands”
Method	Planar
Distance Unit	Meters

GWR Model	
GWR	
Input Features	SinkholePolygons
Dependent Variable	Percentage of Surrounding Area Occupied by Sink Areas
Model Type	Continuous (Gaussian)
Explanatory Variable(s)	NEAR_PL, NEAR_R, NEAR_W
Output Features	GWR
Neighborhood Type	Distance band
Neighborhood Selection Method	User defined
Distance band	4166.4301 meters

Figure 1

GWR Output Charts