Mining for Hydrologic Features in LiDAR Data

Rebecca Reizner, Eric Shaffer, and Brianna Birman, *University of Illinois at Urbana-Champaign*

***Abstract*—Light Detection and Ranging (LiDAR) can generate 3D point data of terrains with high resolution and accuracy, enabling automated detection of important hydrologic features. This paper describes a method for detecting sinkholes in LiDAR data. Current methods of sinkhole detection are lengthy and labor intensive, requiring hours or days of manual work. The method demonstrated in this study can locate sinkholes in the same LiDAR data within minutes with no need for human intervention.**

I. Introduction

Automated detection of hydrologic features has become increasingly important for geologists. The ability to acquire high-resolution LiDAR data for large swaths of land means that much more data is available for analysis. The increased detail of LiDAR data over USGS topographic maps potentially allow up to 30% more sinkholes to be identified[8]. Unfortunately, traditional, mostly manual methods for landform analysis do not scale well. Sinkhole identification is an operation of particular interest, as sinkholes cause safety hazards to those living and working in areas exhibiting the potential for such formations. This is because sinkholes serve as a direct conduit to the underlying bedrock aquifer in the region creating a high potential for groundwater contamination[7].

II. Previous Work

Sinkhole detection and cataloging has been an important problem for decades. Previous methods have used seismic and acoustic emission/ microseismic(AE/MS) techniques[1], topographic maps, aerial photos[2], contouring[3], and LiDAR data visually inspected for sinkholes. A common approach to identifying sinkholes is to locate closed depression contours[8]. Even when computers are used for the contouring or slope analysis, people are still needed to accurately locate the sinkholes by hand.

A study by Young[5] has attempted to use LiDAR to ocate sinkholes in Jefferson County, West Virginia. He has created a DEM from the data and used a modification of the Terrain Shape Index to attempt to locate sinkholes. His algorithm found 94 sites. They were able to visit 55 of these to determine accuracy. Of these, 16.4% were definitely a sinkhole, 43.6% were probably a sinkhole, 25.5% were depressions, and 14.5% were not sinkholes. The geologists desired greater accuracy then this and when we tried a similar technique, our results were poorer.

While LiDAR data has been effectively used to segment many urban features[4], identifying landforms in LiDAR data has not been researched extensively.

III. Hydrologic Features

Sinkholes are one of the most studied hydrologic land features. They are formed when ground below the surface erodes away causing the land to collapse. This erosion is due to ground water slowly dissolving and washing away the underlying bedrock which is typically limestone or other carbonate rock. Sinkholes can vary in size dramatically from less than a foot deep to thousands of feet across. Shapes vary from circular to elongated to completely irregular. When first formed, the sides tend to be very steep and cylindrical. Over time, erosion cause the sinkholes to flatten out into more of a cone shape. Tools for automatic identification of sinkholes must be sophisticated in order to accurately analyze the immense variety of formations.

IV. Methods

Testing was done on a tract of land 20,000 by 35,000 feet in Waterloo, IL. This area is characterized by thousands of sinkholes. The LiDAR data was acquired by the Illinois State Geological Survey in April 2011. The sampling method had the contractor flying over the same area twice, once with a density of at least 1pt/m^2, and once at a lower altitutde with a point density of at least 4pts/m^2. This was to achieve improved vegetation penetration. LiDAR Class 2 points are classified as ground points. LiDAR Class 8 points are derived from LiDAR Class 2 points and are an interpolation of the key points. A combination of Class 2 and Class 8 points were the basis of the data used for our algorithm.

A digital elevation map (DEM) was created from this data at $\frac{1}{10}$ resolution. This operation effectively generated a regular spatial clustering of the original set of points and enabled interpolation within sparse areas.

As seen in Algorithm 1, an iterative process then segmented out all of the points that were in the lowest 1% of the heights. We created sets of points that were touching. If this set contained more than 20 cells it was temporarily labeled as a sinkhole. The exclusion of the smaller sinkholes prevented noise from the LiDAR data being counted as a sinkhole. The process was then repeated, segmenting out the lowest 2% of ground heights. This time the new sinkholes are compared to the old sinkholes. If one of the new sinkholes covers 2 or more old sinkholes that are larger than 100 cells, it is discarded. If the new sinkhole covers multiple sinkholes that are smaller than 100 cells, the smaller old sinkholes

Rebecca Reizner is with the Department of Computational Science and Engineering, University of Illinois, Urbana, IL 61801, USA (phone: 630-696-2456; email: reizner1@illinois.edu).

Eric Shaffer is with the Department of Computational Science and Engineering, University of Illinois, Urbana, IL 61801, USA (phone: 217-372-4190; email: shaffer1@illinois.edu).

Brianna Birman is with the Department of Computer Science, University of Illinois, Urbana, IL 61801, USA (email: birman1@illinois.edu).

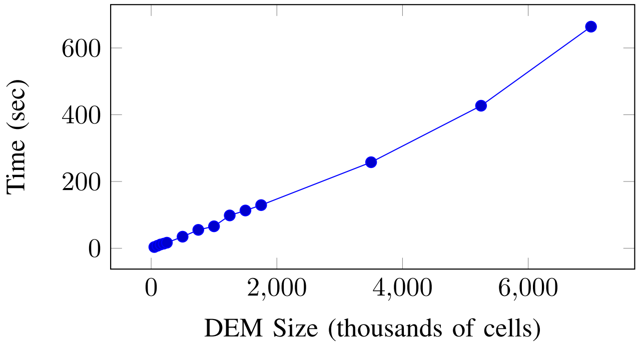

Fig. 1. Time vs Problem Size

are discarded, allowing the newer larger one to effectively absorb them. This value of 100 cells was used to mirror the manual process of segmenting sinkholes as performed by geologists. If a new sinkhole covers only one old sinkhole, the old sinkhole is replaced with the new one. If a new sinkhole does not cover an old sinkhole, it is simply added to the temporary list of sinkholes. This process is repeated up until 99% of the lowest elevation points in the DEM are segmented out and checked for sinkholes.

Algorithm 1 Find Sinkholes

```
Input: DEM
Output: List of sinkholes
Initialize SINKHOLES to empty list
for $i = 0.01$; $i < 1$; $i+ = 0.01$ do
    Flood DEM at $i$
    Add potential sinkholes to SINKHOLES
    if new sinkhole overlaps old sinkhole then
        if old sinkhole is smaller than 100 cells then
            remove old sinkhole
        else if new sinkhole overlaps 2 or more old sinkholes then
            remove new sinkhole
        end if
    end if
end for
return SINKHOLES
```

The algorithm is scalable, requiring linear time in the number of cells in the DEM. This theoretical time-bound has been verified from experimental timings, as seen in Figure 1.

V. Main Results

Our algorithm found 2564 sinkholes in the LiDAR data. The LiDAR data consists of 56 las files creating a total of 15.2 GB of data. The program takes under 10 minutes to complete running serially on a 2.00GHz Intel Xeon CPU with 126GB of memory. Figure 2 shows these sinkholes overlayed on the DEM we created. Segmenting the same data set by hand would require days.

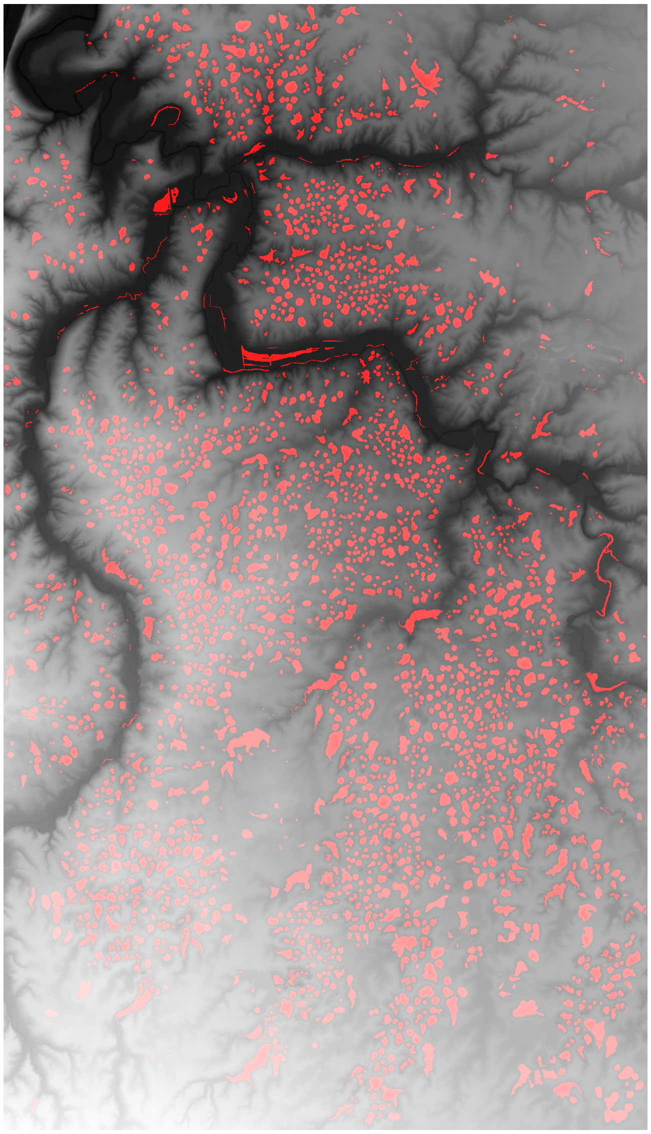

Fig. 2. Detected Sinkholes

To verify our results, we obtained shapefiles from geologists at ISGS that contained data for 2451 sinkholes found using the same LiDAR data. In comparison, this data took them several days to manually generate. To compare our results to the geologists', we filtered out the sinkholes that they found with a bounding box less than 2000ft^2. This is so they would be comparable to the sinkholes we found which only includes sinkholes that cover at least 20 DEM cells. It is necessary to have this lower bound to prevent larger error rates due to differences in interpolation between the LiDAR points. Using this method 83% of the sinkholes identified by the geologists were found with our algorithm. Furthermore, 96% of the sinkholes we found were sinkholes that geologist also found. Further refinement needs to be done in tandem with the geologists to clarify the properties of sinkholes and determine if our algorithm needs to be more or less selective.

VI. Further Filtering

After reviewing our sinkholes, we learned that our algorithm was identifying sections of streambed as sinkholes. We determined that one characteristic differentiating streambed from actual sinkholes is aspect ratio, because thin, long depressions are more frequently streambeds. A second differentiating metric is the fraction of the bounding box around the sinkhole is filled, with curving streambed depressions filling less of their bounding box. These metrics are scale-invariant, allowing them to be applied generally to the initial set of detected hydrologic features.

To employ these metrics as filters, we needed to determine threshold values for each that differentiate sinkholes from streambed. To do this, we manually created a training dataset with sinkholes and streambeds labeled and fed this data into Weka's[6] decision tree algorithm. We used the decision tree to determine the cutoff points for each of these ratios, and then used the learned ratios to perform streamed filtering on the rest of the data. The filtering algorithm proved quite effective, with a sampling of our results before and after streambed filtering shown in Figure 3 and Figure 4 respectively. This brought our false positives from 3.9% to 2.7%. However, this filtering also lowered the number of professionally identified sinkholes that our algorithm found from 84.5% to 83.3%.

Table I shows which of the sinkholes our algorithm found were also identified by the geologists with varying filters. The first is with no filtering. The second is with filtering out sinkholes that are smaller than 20 DEM cells. The third is with the same filter and the streambed filter. These are the same filters represented in Tables II, III, and IV. These three tables represent how many of the geologists sinkholes were found with our algorithm. Table II shows this data in reference to all of the geolgists' sinkholes. Table III represents only the geologists' sinkholes that have a bounding box greater than 2000m^2. Table IV shows only the geologists' sinkholes that have a bounding box greater than 4000m^2.

TABLE I
ACCURACY OF SINKHOLES

Filters	Total	Ours Verified	False Positives	Percent Accurate
None	2636	2315	321	87.8
> 20	2162	2077	85	96.1
> 20 & SF	2113	2056	57	97.3

TABLE II
COMPLETENESS OF ALL SINKHOLES

Filters	Found	Total	Percent Found
None	1837	2283	80.5
> 20	1658	2283	72.6
> 20 & SF	1628	2283	71.3

VII. MOVING TOWARDS SINKHOLE CHARACTERIZATION

The ability to identify sinkholes in LiDAR data effectively allows the creation of a digital catalog of sinkholes. A next step is to look at what can be learned about sinkholes through analysis of such a catalog. Our software can compute some basic geometric characteristics of sinkholes such as perimeter and depth. We can also extract information about vegetation locations from LiDAR data. With this data, one can define multiple classes, such as dividing perimeter lengths into three classes of *small*, *medium*, and *large* and similar classes for depth. One interesting question is then how being in one class influences the probability of being in another class. We chose to use a Naive Bayesian Classifier to answer such questions. Clearly, there may be confounding variables that spoil the assumption of conditional independence. So, we must proceed understanding that high probabilities may be simply be indicative of the existence of such a confounding variable. The discovery of such a variable would be interesting in and of itself, making the investigation a worthwhile pursuit.

As an initial inquiry, we examined the relationship between the maximum relative depth (distance from the lowest point of the sinkhole to the top of the sinkhole) and the perimeter using a set of 2366 sinkholes. The perimeter characteristic is divided into three buckets: 0 - 60 feet is

TABLE III
COMPLETENESS OF > 2000 SINKHOLES

Filters	Found	Total	Percent Found
None	1703	1930	88.2
> 20	1630	1930	84.5
> 20 & SF	1608	1930	83.3

TABLE IV
COMPLETENESS OF > 4000 SINKHOLES

Filters	Found	Total	Percent Found
None	1625	1807	89.9
> 20	1602	1807	88.6
> 20 & SF	1583	1807	87.6

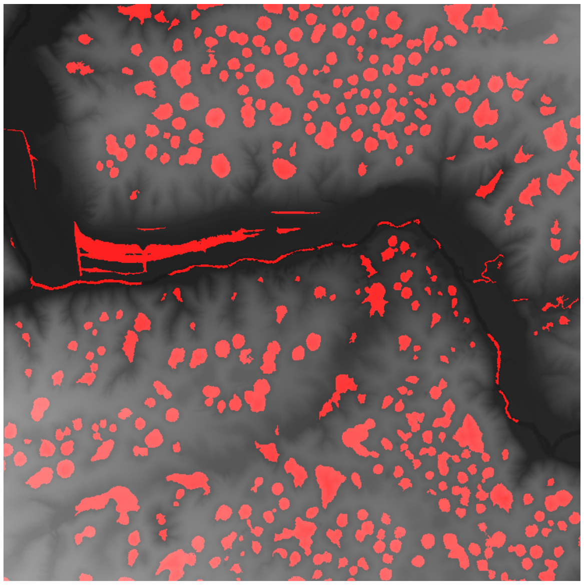

Fig. 3. Before Streambed Filtering

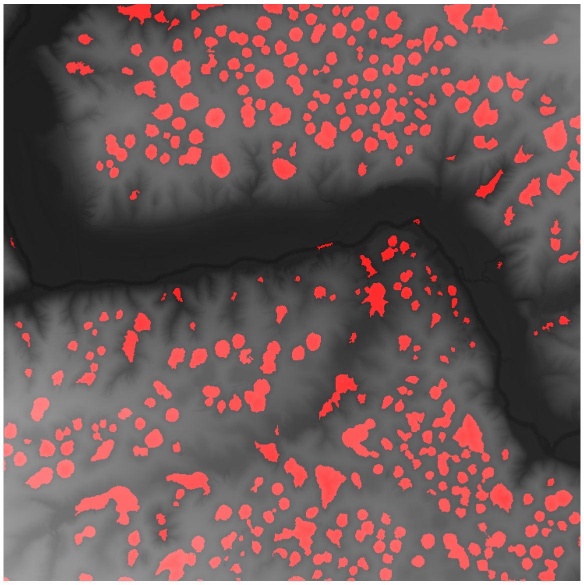

Fig. 4. After Streambed Filtering

small, 60 to 95 is *medium*, and greater than 95 is *large*. Depth is divided into the following buckets: 0 to 15 feet is *shallow*, 15 to 22 is *moderate*, and greater than 22 meters is *deep*. We then calculate the likelihood of a certain depth given the perimeter, producing the results in Table V.

TABLE V
Probability of Depth Given the Perimeter

	Small Perimeter	Medium Perimeter	Large Perimeter
Shallow	0.640083	0.376623	0.327273
Medium	0.287795	0.345083	0.246753
Deep	0.072122	0.278293	0.4259744

The table shows that some generalizations can be made about the geometric structure of sinkholes. A shallow depth is most likely for a sinkhole with a small perimeter, while *deep* is the least likely. The depth probabilities for a medium perimeter sinkhole are much less pronounced. Shallow and moderate depths are more likely than deep, but not by as much as it was for small perimeter sinkholes. A sinkhole with a large perimeter is most likely deep, and least likely of moderate depth, but like medium perimeter, the results are not as pronounced as the depth likelihoods for small perimeters.

A more intriguing exercise is to look at the relationship between vegetation and sinkholes. As part of pre-classified LiDAR data, points that are determined to be vegetation are divided into three categories: *low*, *medium*, and *high*. These vegetation points occur above points that are classified as bare earth. By projecting vegetation points to the bare earth level, we can determine if that vegetation is covering a sinkhole.

The results in Table VI show that vegetation of every type is more likely on a sinkhole than it is on other land. An overlay image of vegetation and sinkholes, seen in Figure 5 shows a portion of the analyzed area. It implies that this strong relationship stems from land without sinkholes being more often cleared for development.

TABLE VI
Probability of Vegetation Level Given Sinkhole Existence

	Over Sinkhole	Not Over Sinkhole
Low Vegetation	0.965995	0.888883
Medium Vegetation	0.987562	0.926441
High Vegetation	0.995783	0.939347

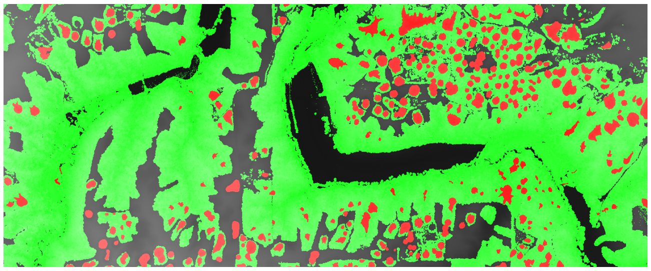

Fig. 5. Sinkholes with Detected Vegetation Overlaid

VIII. Conclusions and Future Directions

The algorithm presented in this paper provides an accurate, efficient, and automated way of identifying sinkholes from LiDAR data. This allows sinkholes to be cataloged and monitored, providing important information for land planning strategies. Future work will attempt to characterize the risk of sinkhole formation in an area through correlations between sinkholes and soil type. It may also be possible that the geometric pattern of emergent sinkholes, by exposing underlying geological lineation, can be used to predict where new sinkholes are likely to form.

Acknowledgment

The authors would like to thank Donald Luman and Samuel Panno from the Illinois State Geological Survey.

References

[1] Hardy, H. R., Jr., Belesky, R. M., Mrugala, M., Kimble, E. J., & Hager, M. E. 1986, Pennsylvania State Univ. Report

[2] Mukherjee, Arindam, Pavlowsky, Robert T., and Gouzie, Douglas, "GIS Database for Sinkhole Hazard Assessment in Christian County, Missouri" Joint South-Central and North-Central Sections, both conducting their 41st Annual Meeting (1113 April 2007)

[3] Seale, L. Don, Brinkmann, Robert, and Vacher, H.L. "GIS Database for Sinkhole Hazard Assessment in Christian County, Missouri" Joint South-Central and North-Central Sections, both conducting their 41st Annual Meeting (1113 April 2007)

[4] Golovinskiy, Aleksey, Kim, Vladimir G., and Funkhouser, Thomas "Shape-based Recognition of 3D Point Clouds in Urban Environments"

[5] Young, John, "Using LiDAR to map sinkholes in Jefferson County, West Virginia" Eastern Panhandle West Virginia GIS Users Group Meeting (2009)

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[7] Merrick & Company, "Merrick Utilizes LiDAR in Large Sinkhole Plain" (March 07, 2013)

[8] Jacoby, Doug CMS, GISP; Luman, Donald PhD; "Sinkhole ID" (November, 1, 2012)