2950 Niles Road, St. Joseph, MI 49085-9659, USA
269.429.0300 fax 269.429.3852 hq@asabe.org www.asabe.org

An ASABE Meeting Presentation

Paper Number: 131619828

Terrain analysis and data mining techniques applied to location of classic gully in a watershed

Laurimar Gonçalves Vendrusculo[1,2], Amy Kaleita[1]

[1] Agricultural and Biosystems Engineering Department – Iowa State University - 126 Davidson Hall, Ames, IA, 50011, USA

[2] Embrapa Informatics - Av. André Tosello, 209 - Campinas, SP, 13083-886, Brasil.

Written for presentation at the

2013 ASABE Annual International Meeting

Sponsored by ASABE

Kansas City, Missouri

July 21 – 24, 2013

Abstract. Gullies are an extreme form of soil erosion that degrade diverse environments trough the siltation of streams and water bodies. Indirectly, gully erosion compromises crop productivity working as a link to watercourse allowing movement of detached topsoil particles from agricultural fields during heavy storm events. Furthermore, studies found reduction of the catchment area when active gullies are present. This complex process involves multiple factors and it demands to be studied consistently in order to locate the areas prone for gully erosion. The determination of gullies areas depends upon topographical, geological, and hydrological characteristics; however its location is mainly controlled by the high capacity of overland flow to cut the channel. We hypothesize that identification of gully in agricultural landscape can be performed from high-resolution elevation data products and unsupervised clustering approaches. In order to examine this hypothesis we have used variables resultant from of LiDAR-based terrain analysis as input of a three clustering techniques. A k-means, fuzzy k-means, and CLARA clustering algorithms were used to carry out the cluster analysis. The results of the cluster analysis suggested that 8 classes were optimal for group areas in the watershed. Elevation data from one field-scale watershed near Treynor in Pottawattamie County, IA, was used to calibration purpose and terrain analysis using slope, flow accumulation, plan convexity, topographic wetness Index, and stream power index were calculated. The cluster analysis has shown highest concordance with percentage of corrected classified pixels that approach based in medoid (CLARA) has obtained the best agreement of points within gullied area (30.1%). The results of this research might speed up gullies field surveys and also can serve as input in conservation planning framework

Keywords. *data mining, soil erosion, Lidar data, feature classification*

Introduction

Classic gullies are an extreme form of soil erosion which degrade diverse environments trough the siltation of streams and water bodies. Indirectly, gully erosion compromises crop productivity working as a link to watercourses. It allows movement of detached topsoil particles from agricultural fields during heavy storm events.

Many studies (e.g. Vandaele et al., 1996, Vandekerckhove et al., 1998, Poesen et al., 2003, Poesen et al., 2011) indicated that location of gullies areas depends upon topographical, geological, and hydrological characteristics; however its location is mainly controlled by the high capacity of overland flow to cut the channel. Nevertheless, despite of its significance, only a restricted erosion models account for channel sediment losses in its procedures, for instance: AGNPS (Annual Agricultural non-point source pollution), WEPP (Water Erosion Prediction Project) model, EGEM (ephemeral gully erosion model), and CREAMS (Chemicals, Runoff and Erosion from Agricultural Management Systems). Most of these models compute sediment production of small or ephemeral gullies. Furthermore, those models need also the user guidance to indicate the gully location. Thus, an accurate and no-time consuming methodology to indicate the gully location in the landscape is needed.

The concept of a topographic threshold is applied, recently, to predict location in the landscape where gullies are prone to develop (Momm et al. 2013, Poesen et al., 2011). A negative relationship between drainage area and watershed slope to associate channel incision was found by Begin and Schumm (1979).

Other approach applied to describe potential flow erosion is the unit stream power index (SPI). This measure of erosive power of overland flow is detailed by Moore et al. (1993) and its equation given as:

$$Stream\ Power\ Index(SPI) = Ln(\ \frac{A_S}{S}) \quad (1)$$

The A_S is the local upslope contributing area per unit width of contour line and *S* is the local slope.

The topographic wetness index (TWI) is often employed to simulate the soil moisture conditions in a watershed and also takes in account both a local slope geometry and site location in the landscape. Validation studies of this index were done by Beven and Kirkby, 1979 and Deng and Li, 2002. The TWI is usually used to describe the long term soil moisture at each point of a drainage basin.

$$Topographic\ Wetness\ Index\ (TWI) = Ln(\ \frac{A_S * Pixel\ area}{Tan(\frac{S * \pi}{180})}) \quad (2)$$

With availability of high resolution spatial data and data mining techniques, tasks of grouping objects with similarity are able to be performed repeatedly and applicable to large data sets.

The authors are solely responsible for the content of this meeting presentation. The presentation does not necessarily reflect the official position of the American Society of Agricultural and Biological Engineers (ASABE), and its printing and distribution does not constitute an endorsement of views which may be expressed. Meeting presentations are not subject to the formal peer review process by ASABE editorial committees; therefore, they are not to be presented as refereed publications. Citation of this work should state that it is from an ASABE meeting paper. EXAMPLE: Author's Last Name, Initials. 2013. Title of Presentation. ASABE Paper No. ---. St. Joseph, Mich.: ASABE. For information about securing permission to reprint or reproduce a meeting presentation, please contact ASABE at rutter@asabe.org or 269-932-7004 (2950 Niles Road, St. Joseph, MI 49085-9659 USA).

Clustering is one of these tasks, which encompass methods to form analogous groups from observations or samples into classes (clusters). This means that samples that have similar attribute values are close together in a multidimensional feature space and consequently form a distinct cluster (Han and Kamber, 2006; Pakhira et al., 2005). On the other hand, the use of a statistical clustering technique can also produce boundaries based on minor attribute difference, or, reflect data noise (e.g. measurement error, caused by sampling bias, etc). The clustering techniques, generally, group samples into distinct classes with discrete boundaries. However, soil or topographic attribute vary gradually over space, for example, and this representation may increase errors associate to inappropriate boundaries (Burrough et al. 1998)

K-means is one example of discrete or hard clustering method (MacQueen, 1967). It is the simplest and a fast unsupervised learning algorithm based on fixed (k) number of clusters. Once initial points are assigned as a centroid (points in space that represent the center of the cluster), then each sample is labeled comparing its distance to the centroid. As disadvantage, k-means with different initial partitions can result in distinct final cluster configuration.

In order to overcome the problem of class overlapping, Bezdek et al. (1984) proposed the fuzzy k-means (FCM) approach and extended by De Gruijter and McBratney (1988). This technique seems suitable to environmental sciences (Burrough et al. 2000). This is because the degree with a sample belongs to a given class is expressed not in terms of a binary “yes” or “No”, but rather than a continuous membership value that ranges, for instance, from 0 to 100. The method FCM clustering allows points to belong to more than one cluster, as result is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{c} u_{ij}^{m} \left|\left|x_i - c_j\right|\right|^2, \; 1 \leq m < \propto \qquad (3)$$

Where m is any real number greater than 1, u_{ij} the degree of membership of xi in the cluster j, xi is the ith of d-dimensional measured data, cj is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the centroid.

Instead to tackle clustering with centroid points Kaufman and Rousseeuw (1990) created the medoid approach to group large data set. CLARA (Clustering LARge Applications), a partitioning method, is one of the medoid variations which find medoids for a sample from the data set.

We hypothesize that identification of gully in agricultural landscape can be performed from high-resolution elevation data products and unsupervised clustering approach. In order to examine this hypothesis we have used outcomes variables from LiDAR-based terrain analysis as input of a three clustering techniques

The main goal of this study is: Analyze the efficiency of three unsupervised clustering techniques to identify potential gullies zones trough terrain analysis variables in a field-scale watershed.

Material and Methods

Description of study area

The study area selected has a computed drainage area of 330,574 m^2 (81.6 Acre or 33.05

hectare) and is a field-scale watershed (41° 9' 44.54"N, 95° 38' 19.94" W) near Treynor in southern Pottawattamie County, IA (Figure 1). Entitled as Watershed #1, this field is one of four study areas established by the U.S. Department of Agriculture Research Service (USDA-ARS) and since 1965 was instrumented to provide measurements of runoff, base flow and sediment concentration. These measurements were quantified using broad-crested V-notch weirs located at the base of each watershed where the gullies channels are located. Precipitation was measured by rain gauges placed in the watershed perimeter. Four soil types occur in watershed #1, with the predominant soil being Monona silt loam (fine-silty, mixed, superactive, calcareous, mesic). Other soil types found in the watershed were Ida silt loam (fine-silty, mixed, superactive, nonacid, mesic), Marshall Silt clay loam (fine-silty, mixed, superactive, calcareous, mesic). The slopes in this sites range from 2% to 4% at the ridges and from 14 % to 20% for valleys.

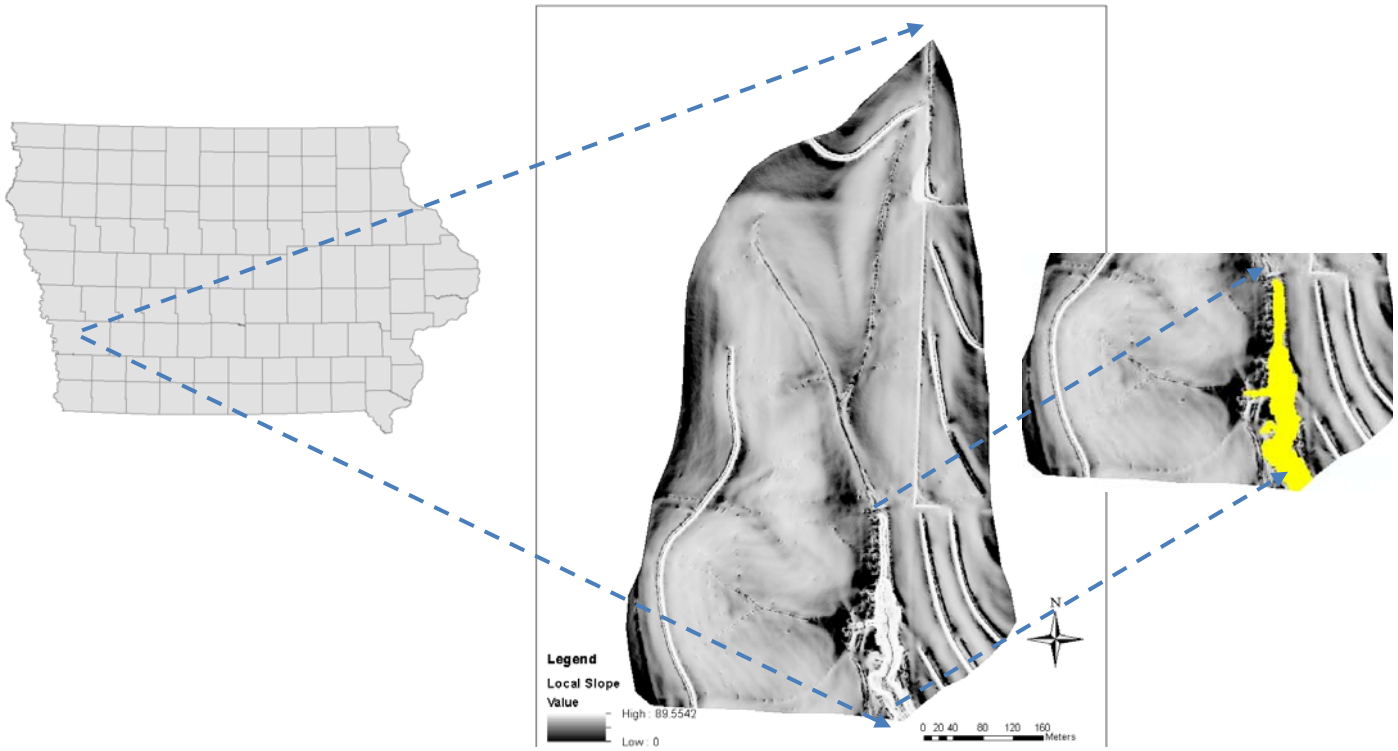

Fig. 1 - Topography map of watershed #1 in Western Iowa generated from LiDAR outcomes and, detail of gullied area.

Methods

Analysis involved completion of four main tasks: acquisition of spatial databases and creation of a digital elevation model (DEMs) of the study area, computation of terrain indices, classification of distinct areas and accuracy assessment.

Spatial database and terrain model creation

Spatial data for Pottawattamie County, including soil survey attributes and aerial imagery (national aerial imagery program, NAIP), were acquired from Iowa Department of Natural Resources Data Gateway (http:// http://www.igsb.uiowa.edu/webapps/nrgislibx/) and LiDAR (LIgth Detection And Ranging) data was obtained by Iowa LiDAR mapping project (

http://geotree2.geog.uni.edu/lidar/). The integration of the spatial data was performed by Arc Map version 10.

Using LiDAR survey Iowa statewide for the watershed #1 were identified 192,000 points inside of the study area. This raw LiDAR point cloud was processed to produce digital elevation model (DEM) and triangulated irregular network with spatial resolution of 1 meter. From the DEM, consequent maps of topographic and hydrological variables were computed (slope, drainage network and aspect) implemented by ArcGis 10 (Jenson and Dominguez, 1988). The gullied area was determined by previous surveys.

Local slope (%) was calculated by eight-direction (D8) algorithm (Greenlee, 1987; Jenson and Dominguez, 1988), and later the percent slop was divided by 100 to obtain slope in m/m unit. In the same way flow direction and flow accumulation grids were computed. Flow accumulation for each cell represents the sum of all upstream elements (pixels) draining to the watershed outlet. The plan curvature (m/100) was computed by Zevenbergen and Thorne(1987) method where positive values represent a convex surface and negative a concave surface. Zero values indicate flat topography. Profile curvature is towards the maximum slope and plan curvature is perpendicular to the direction of the maximum slope (ESRI, 2013)

Creation of terrain-analysis indices

Initially, the stream power and topographic wetness indices were computed trough raster calculator function available at ArcGis. The drainage area present in both equations was calculated by the flow accumulation procedure. An automated model was created using model builder application (ArcGis 10) to calculate and consolidate topographic indices in watershed #1 data. The squared boxes represent the ArcGis functions (e.g. raster calculator, transform raster to point) and the rounded boxes symbolize the initial and intermediate input/output data as depicted in Figure 2.

Fig. 2 - Flow chart showing the process to calculate topographic indices.

Once all the topographic, hydrologic features and indices were computed for each cell for its correspondent geography coordinates, they were consolidated in one text file to be used in the classification task. Table 1 lists descriptive statistics for the study area.

Other important spatial layers were analyzed in this study such as the soil data from SSURGO database which contains diverse soil parameter (e.g, soil series, available water capacity, surface and subsoil bulk density, parent material, texture, sand, silt, and clay content). From SSURGO soil properties analysis we could confirmed areas with soils types prone to erosion as well.

Table 1 – Summary statistics of the quantitative independent variables used in the model.

Variable	Max	Min	Mean	Standard Deviation	Units
Local slope	40.6	0.0003	5.3	3.5	%
Curvature	211	-238	0.05	7.8	rad/meter
Plan curvature (tangential)	109.2	-160	0.02	3.6	m/100
Profile curvature	145.6	-120.6	-0.03	5.7	m/100
Flow Accumulation	701	1	10.6	23.4	Pixels (m^2)
Stream Power Index	7.4	-1.2	2.8	1.3	pixels
Topographic wetness Index	14.8	0.15	4	1.54	pixels

Creation of distinct areas in the watershed and accuracy assessment

Three unsupervised classification approach were used to analyze the data. K-Means (centroid approach), Fuzzy K-means (membership centroid approach) and CLARA (medoid approach) clustering techniques were performed on the dataset by R package (e1071, cluster and clusterSim functions). The within-cluster variation was the measure used to define the optimal number of clusters which is required to each grouping technique. In this case given the number of cluster *K*, the clustering algorithm minimizes the within-cluster variation:

$$W = \sum_{k=1}^{K} \sum_{C(i)=k} ||X_i - X_k||_2^2 \quad (4)$$

Over clustering assignments C, where X_k is the average of points in group *K* and

$$X_k = \frac{1}{n_k} \sum_{C(i)=k} X_i \quad (5)$$

As a way to measure the accuracy of the method to classify correctly the pixels in gullied or non-gullied area, we compared all gully pixels in the reference map to those pixels classified by each clustering technique. Thus, the percentage of accuracy was calculated as total correct pixels in the gully area divided by total test pixels, multiplied by 100. The equation of the percentage of corrected classified pixel index (PCCP) is given by

Percentage of Corrected Classified Pixels ($PCCP$) = ((Pixels within gully area)/(total area classified as gully at watershed #1))* 100 (6)

Results and Discussion

The flow accumulation has the highest variability (s.d. 23.4 pixels) and local slopes vary from 0 to 40 % suggesting that abrupt breaks in the landscape occurs locally in the gullied area.

The sum squared distance within clusters curve displayed local minima at the set of eight (8) classes, which suggested that this number of classes is optimal for partitioning the set of observed areas in the watershed.

Among the three clustering techniques, CLARA (CLustering LARge Applications) obtained the best classification rate in the gully feature (Tab. 2). CLARA is based on medoid (median) approach instead of centroid method of K-means and Fuzzy K-means (Fig. 3). The resulting plot of the eighth classes by each classifier method is shown in Figure 3.

Table 2 – Calculation of percentage of corrected classified pixels (PCCP)

Clustering technique	Pixels within gully area (m^2)	total area classified as gully at watershed #1(m^2)	PCCP (%)
K-means	708	2544	27.83
Fuzzy K-means	3162	59977	5.27
CLARA	982	3173	30.95

It can noted graphically in Figure 3 that fuzzy k-means has the worst performance of classification including not only points in the gullied area but in curves lines. Even though, CLARA model final results classified more spurious points outside of the gullied area it appear that linear features prone to high concentrated overland flow were included in its analysis. It worth mentioned that the lower rates of correct classification are due to the inclusion of pixels in the contour lines.

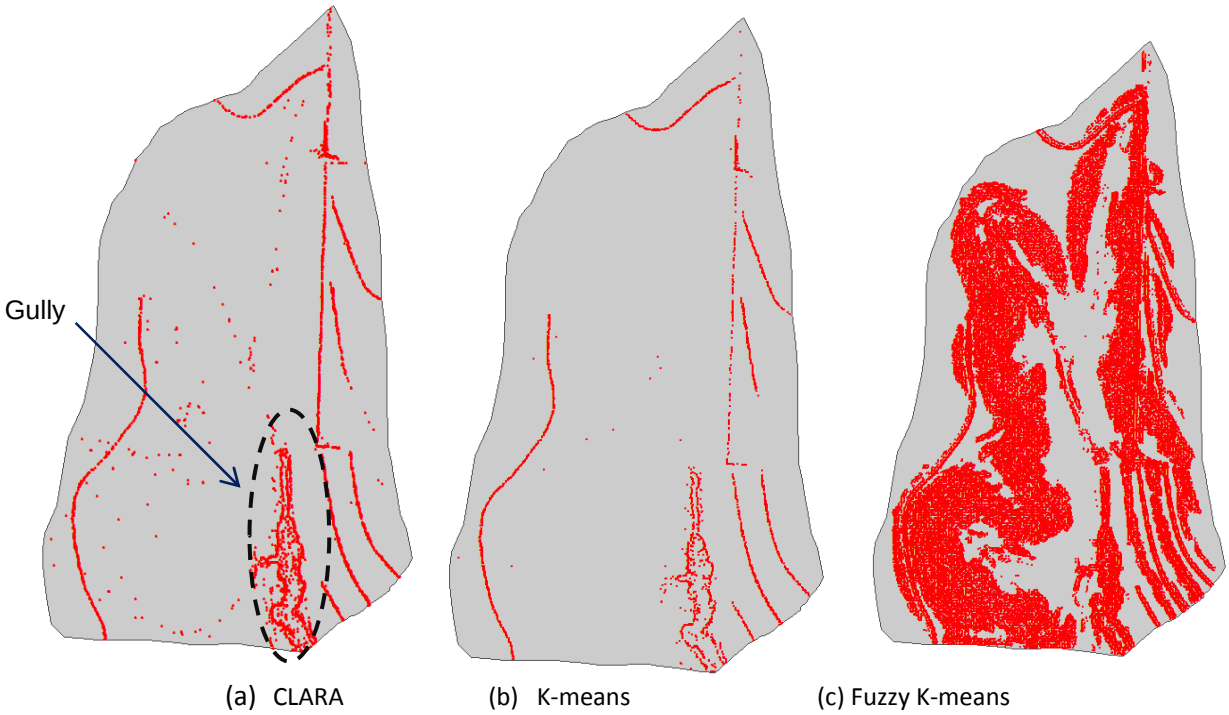

Fig. 3 - Cluster with gully characteristics grouped by (a) CLARA, (b) K-means and, (c) Fuzzy K-means techniques.

Cluster 1 computed by CLARA algorithm describes accordingly the gully perimeter. It is characterized by local high concaves topographies (mean: 21.51 rad/meter), and has negative profile curvature (mean: - 19.4 rad/m) which means convex flow, typically on gully edges and contour lines. Furthermore, the positive mean plan curvature (2.11 rad/m) suggests converging flows in the bottom-valley gully. Furthermore, we found that the range of values of Topographic wetness Index from 0.15 to 0.5 lies only in gully areas.

Conclusions

Accurate information about location of erosion features is of essential importance for landowners, stakeholders and researchers. In this study we compare three unsupervised methodologies to identify gully in a field-scale watershed using high-resolution LiDAR products such as topographic index and curvature. The approach using medoid approach (CLARA) achieves best classification rate and it seems promising because is not affected by outlier points. Furthermore, the median approach suggests that areas with similar hydrologic and topographic characteristics should have spatial proximity. So far, the methodology employed in this research presents certain degree of misclassification including contour lines as zones prone to erosion. Thus, improvements that isolate linear erosion features like classic gullies are our next steps.

As a future research, we will continue making advances in medoid approach to account for features of conservation practices such as contours. Also, we are planning to compare gullies location in areas with different topographies and hydrologic features incorporating climate and soil aspects.

We believe that this approach based on multidimensional clustering of topographic and hydrologic characteristics to locate classic gullies has potential to better inform decision makers relate to planning and implementation of soil conservation measures.

Acknowledgements

The authors acknowledge the valuable contribution Kevin Cole who provided the climate and hydrographic data for the Treynor site. This research was supported by Embrapa and Environmental Science Department of Iowa State University.

References

Begin, Z. B., and S. A. Schumm.1979. Instability of alluvial valley floors: a method for its assessment. *Trans. ASAE*. 22: 347-50.

Beven, K. J., Kirkby, M. J., 1979. A physically based, variable contributing area model of basion hydrology, *Hydrological Sciences Bulletin*. 24: 43-69

Bezdek, J.C., Ehrlich, R., Full, W., 1984. FCM: the fuzzy c-means clustering algorithm. *Computers Geosciences*. 10: 191–203

Burrough, P.A., McDonnell, R.A., 1998. Principles of geographical information systems. Principles of Geographical Information Systems. Oxford University Press, USA.

Burrough, P.A., Van Gaans, P.F.M., MacMillan, R.A. 2000. High resolution landform classification using fuzzy k-means. *Fuzzy sets and systems*. 113: 37-52.

Deng H., and Li X. 2002. Relationship of upslope contribution area and soil water content in TOPMODEL, *Progress in Geography*. 21(2): 103- 110.

De Gruijter, J.J., McBratney, A.B., 1988. A modified fuzzy k-means method for predictive classification. In: Bock, H.H. (Ed.), Classification and Related Methods of DataAnalysis. Elsevier Science Publishers, B.V, 97–104.

ESRI. Curvature. 2013. Available at : < http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=curvature>

Greenlee, D. D. 1987. Raster and Vector Processing for Scanned Linework. *Photogrammetric Engineering and Remote Sensing* 53 (10): 1383–1387.

Han, Jiawei,and Micheline Kamber. 2006. *Data Mining: Concepts and Techniques*, 2nd edition, Morgan Kaufmann.

Jenson, S. K., and J. O. Dominguez. 1988. Extracting Topographic Structure from Digital Elevation Data for Geographic Information System Analysis. *Photogrammetric Engineering and Remote Sensing* 54 (11): 1593–1600.

Kaufman, L. and Rousseeuw, P. J. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, Inc., New York, NY.

MacQueen, J. B. 1967. Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1:281-297

Momm, H. G., R. L. Bingner, R. R. Wells, J. R. Rigby, and S. M. Dabney. 2013. Effect of topographic characteristics on compound topographic index for identification of gully channel initiation locations. *Trans. ASABE* 56(2): 523-537.

Moore, I.D., A. Lewis, and Gallant, J. C. 1993. Terrain attributes: Estimation method and scale effects. 30-38. In A. K . Jakeman et al. (ed.) *Modeling change in environmental systems*. John Wiley & Sons, New York.

Pakhira, M.K., Bandyopadhyay, S., Maulik, U., 2005. A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification. *Fuzzy Sets and Syst*. 155, 191–214.

Poesen, J., J. Nachtergaele, Verstraeten, G. and Valentin, C. 2003. Gully erosion and environmental change: importance and research needs. *Catena*. 50: 91-133.

Poesen, J., D. Torri, and T. Vanwalleghem. 2011. Ch. 19 – Gully erosion: procedures to adopt when modelling soil erosion in landscapes affected by gullying. In Morgan, R.P.C., and M.A. Nearing (eds). *Handbook of Erosion Modelling*. Blackwell-Wiley: Oxford.

Vandaele, K., J. Poesen, G. Govers, and B. van Wesemael. 1996. Geomorphic threshold conditions for ephemeral gully incision. *Geomorphology*. 16: 161-73.

Vandekerckhove, L., J. Poesen, D. Oostwoud Wijdenes, and T. de Figueiredo. 1998. Topographical thresholds for ephemeral gully initiation in intensively cultivated areas of the Mediterranean. *Catena*. 33: 271-92.

Zevenbergen, L. W., and C. R. Thorne. 1987. Quantitative analysis of land surface topography. *Earth Surf. Proc. Land*. 12: 47-56.