Intercontinental Geoinformation Days

http://igd.mersin.edu.tr/2020/

Minimum spanning tree detection on raster data

Murat Çalışkan[1], Berk Anbaroğlu*[1]

[1]*Hacettepe University, Faculty of Engineering, Geomatics Engineering, Ankara, Turkey*

Keywords
Minimum Spanning Tree
Kruskal's
Raster

ABSTRACT

Various real-life systems including transportation, infrastructure and social networks require the use of a graph data structure. These graphs usually consist of weighted edges, such as the distance between two intersections on a highway or the cost to establish a power line between two electricity distribution centers. Therefore, having correct weights (costs) assigned to the edges of a graph is an important issue. Assigning these weight values manually is a time-consuming task in a real-world application, since graphs may consist of thousands of edges. On the other hand, assuming distance to represent weight would over-simplify the problem, especially when a graph is representing a real-life spatial phenomenon such as modeling of infrastructure lines. Specifically, raster data should be utilized to model such real-life entities. This paper investigates the effectiveness of three different methods to determine the costs of a weighted graph when the purpose is to detect the minimum spanning tree, which is a tree structure that connects all nodes with minimum cost.

1. INTRODUCTION

Minimum Spanning Tree (MST) is subset of a graph, which is defined as the tree that connects all of the nodes in a graph with the lowest cost (weight). MST can be determined using different algorithms. Kruskal's Algorithm and Prim's Algorithms are two of them. Kruskal's algorithm sorts all edges based on ascending costs, and include them in order to the MST as long as they do not form a cycle and all nodes are included(Kruskal, 1956). Prim's algorithm starts with a random node and grows the MST by including the node that is reachable by the lowest available cost within the nodes that are included in the set(Prim, 1957).

Detecting a MST of a weighted graph is used extensively in network optimization studies such as creating computer circuits, telecommunication network design, delivery routing etc(Ahuja, Magnanti, Orlin, & Reddy, 1995; Dippon & Train, 2000; Mareš, 2008; Pettie & Ramachandran, 2000; Rothfarb, Frank, Kleitman, Rosenbaum, & Steiglitz, 1970). Some of these usage areas are related to Geographic Information Systems. For example; in a study, the length of the line was reduced from 66 km to 49.9 km in the Amoco East Cross oil pipeline project in Alberta-Canada(Dott, 1998). Another study was carried out for the optimization of the South Gabon oil pipeline and with this study, the total line length was reduced from 188.2 miles to 156.2 miles using the Prim's algorithm(Brimberg, Hansen, Lih, Mladenović, & Breton, 2003).

One problem while working with graph data is assigning values to the edges as weight/cost. If the number of edges is few, it can be done manually. But there are more edges to consider, the problem gets more difficult. Suppose that you have a pipeline network that consists of ten thousands of elements that is used within an urban environment. It would take a lot of time to assign a cost value to each edge. The +simple solution is to use length of the edge as cost. But this solution may not be true in every condition. As an instance, while creating an electricity transmission line; land value, slope, land use, height from sea level, type of soil etc. may also be important beside the length of the electricity line. In these kind of situations, a combination of these factors should be taken into account. Specifically, instead of discrete vector data, continuous raster data should be used to estimate the costs of edges of a graph.

* **Corresponding Author**

*(banbar@hacettepe.edu.tr) ORCID ID 0000-0003-2331-6190
(muratcaliskan@hacettepe.edu.tr) ORCID ID 0000-0003-1863-9032

Cite this study

Çalışkan M & Anbaroğlu B (2020). Minimum spanning tree detection on raster data. Intercontinental Geoinformation Days (IGD), 48-51, Mersin, Turkey

2. METHOD

This section first describes the MST and Kruskal's algorithm, and then describes the three main cost methods to estimate the costs of edges when relying on raster data.

2.1. Detecting MST by Kruskal's Algorithm

Graph is a data structure that consists of nodes and edges between these nodes. This structure is used for modelling the transportation systems, internet network, relationship between people in social media etc. An exemplar network consisting of 9 nodes and 13 edges is illustrated in Figure 1.

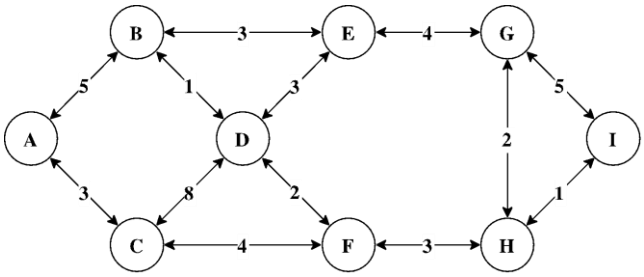

Figure 1. Graph Data Structure

One of the common algorithms that operate on graphs is determining the "Minimum Spanning Tree (MST)", which is defined as the tree connecting all of the nodes in a graph with the lowest cost(weight)(Degenne & Lo Seen, 2016; Graham & Hell, 1985).

In Figure 2 there is a graph consists of nodes A, B, C, D, E and F. Several spanning trees (red dotted lines), that connects all of these nodes, are illustrated. There are more than one options to connect all the nodes as seen. Among them, the one with minimum cost is called Minimum Spanning Tree Figure 2d.

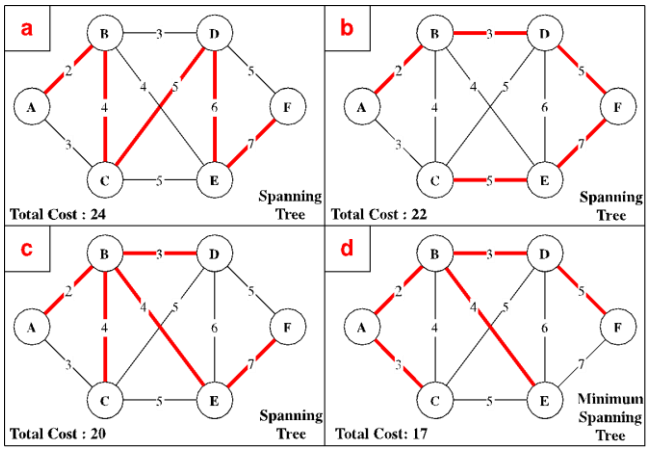

Figure 2. Spanning Tree and Min. Span. Tree

2.2. Generating MST Based on Raster

As mentioned in previous section; in some conditions it may be impossible to assign cost value to the graph manually due to large number of edges and using length of the edge as cost value may be insufficient.

In this study, we proposed a solution for this kind of situations. Our solution is; "Determining the MST, based on a cost raster" as can be seen in Figure 3. For this purpose, we created a plugin that runs on QGIS software and it is published in QGIS Plugin Repository. Using raster data to estimate cost value is also handled and explained thoroughly in Çalışkan and Anbaroğlu, 2020.

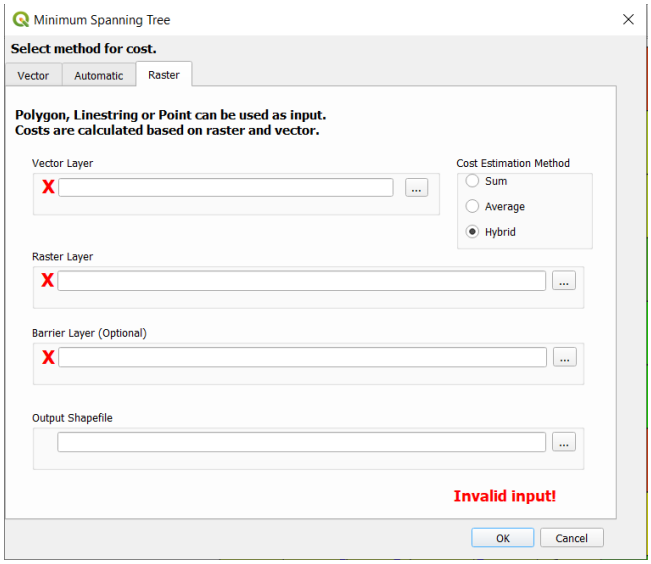

Figure 3. QGIS Plugin for Creating Raster Based MST

Raster, is a grid-like data type that consists of pixels. And in this method, every pixel represents the cost of its location. Therefore, a cost raster is required for edges to be able to gain the cost values. The edges get their cost/weight from corresponding pixels of the cost raster.

There are three ways for calculating the cost of the edge. First one is "summation of the pixels". In this method; the values of the pixels, that the edge intersects, are summed up and used as cost. The second one is "average of the pixels". In this method; the average of the values of the pixels, that the edge intersects, is used as cost. And the last method is "hybrid method". In this method; the average of the values of the pixels, that the edge intersects, is multiplied by the length. And the result is used as cost.

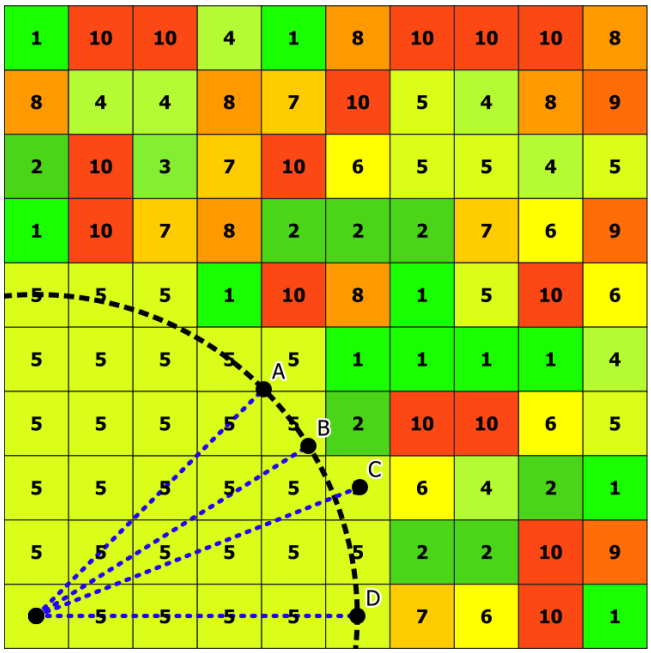

Figure 4. Illustration of Cost Estimation

Based on raster data illustrated in Figure 4, the associated cost values of the four edges are stated in Table 1.

Table 1. Cost Estimation Methods

Edge	Distance	Average	Sum	Hybrid (Dist*Avg)
A	5	5	25	25
B	5	5	40	25
C	5.5	5	40	27.5
D	5	5	30	25

All of these three methods can be used to gain the cost from rasters, but the hybrid method is advised for this process. Because, there are some problems about the first and the second methods. Specifically, in the first method, all the pixel values that an edge intersects are added up. In some conditions the results of this method can be wrong. For example, as seen in Figure 4, the length of edges A and B are the same and they are both in same homogenous surface. But their costs are different because of the size of the pixels.

In the second cost estimation method, average of the pixel values, that an edge intersects, are used as cost. This method isn't affected from the distance in homogenous surfaces, but this method also has some weakness in some situations. As illustrated in Table 1; edges B and C has the same cost in same homogenous surface, although their distances are different. In practical their costs are expected to be different.

In the last method, the cost is multiplication of the length of the edges and the average of the corresponding pixels. This method eliminates the errors, made in previous methods.

For the process in this study, two obligatory and one optional parameter are required. The optional parameter is barrier data. It is a restriction in the process and must be in line format. The MST cannot cross the barrier line. And the obligatory ones are; graph(vector) data and cost(raster) data. The graph data can be in point, line or polygon format. If a point data is used as input, the point features are used as nodes and edges are generated by Delaunay Triangulation. If a polygon data is used as input; at first, centroids of the polygon features are created and the rest is the same as point data. If the input is in line format, the line features are used as edges of the graph.

Cost raster must be created based on needs before the process is started. While creating the cost raster; it must be taken into account that, the higher the pixel value is, the harder it would be to pass that pixel. Restriction, is also possible via the cost raster. If some places are desired to be avoid while generating the MST, the pixels that correspond to the desired places are set as null. In this way, the edges of the MST won't be able to pass those places.

3. RESULTS

As an exemplar scenario for the steps, explained in Figure 5, a case study was performed and the details are explained below.

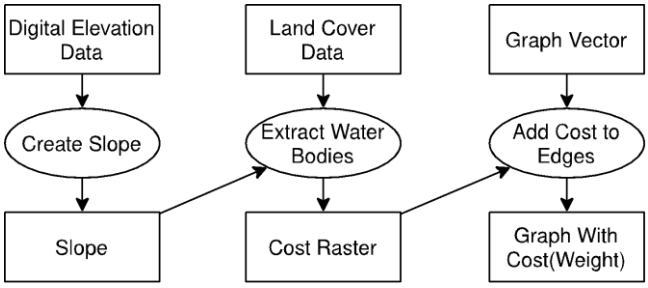

Figure 5. Illustrative Explanation of Gaining Cost Raster

In this example; a slope raster which is assumed to contain cost values was created. In order to create the cost raster data; first of all, Digital Elevation Model(3''≈100m) data for Turkey (from viewfinderpanoramas.org) and CORINE Land Cover data(from https://land.copernicus.eu/pan-european) were downloaded. Next, a slope raster was generated using the DEM and the water bodies are extracted from the DEM as they are thought to be restricted zones in our scenario. The resulting slope data was used as cost raster.

As graph data, the centroids of the cities, which are in point format, in Turkey was used. For barrier line, which is optional in this process, roughly drawn TANAP Natural Gas Pipeline is used(https://enerji.gov.tr/bilgi-merkezi-dogal-gaz-boru-hatlari-ve-projeleri). After all these preparations, only thing to do is to use these data as input for the plugin. In this process, it is required for all of the inputs to be in same coordinate system. The result is illustrated in Figure 6.

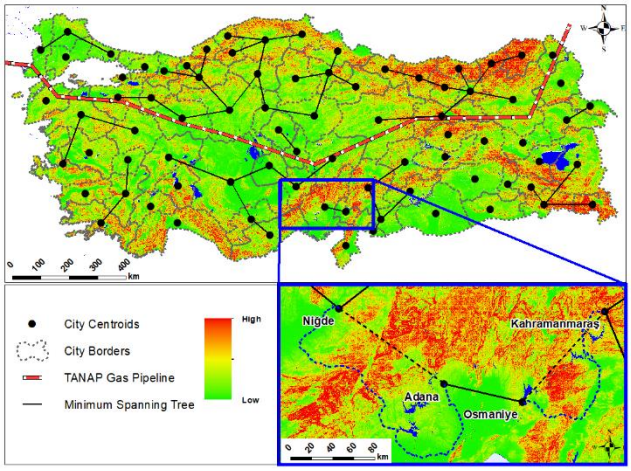

Figure 6. Cost Raster, City Centroids, Barrier Lines and MST in Turkey

In Figure 6; the cost raster, barrier lines, city centroids and the result MST is illustrated. It is obvious in the result that, none of the edges of the MST crosses either the barrier line or the null pixels of the raster.

4. DISCUSSION

In theory, all of the points (city centroids) have to be connected by a continuous line. In Figure 6, it is seen that there are some isolated points. For example, in the blue bordered area, the connection between "Niğde-Adana" and "Osmaniye-Kahramanmaraş" is missing. These cities are expected to be connected by black dotted line. They are not connected, because, there are null values (water

bodies) in the raster under those edges (Niğde-Adana, Osmaniye-K.Maraş). Since the barrier line restricts the connection of the points, it also causes these kind of disconnections.

When it is obligatory to connect these points, some other solutions have to be taken into account. If the point is not in a fully restricted area, such as island, determining a line with multiple vertices that goes around the restricted area, rather than straight line, would be an option.

In Figure 6 an alternative route is illustrated as an example in blue dotted line between Niğde-Adana and Osmaniye-K.Maraş. These lines are created based on the cost raster to minimize the cost.

5. CONCLUSION

In Geographic Information Systems projects, most of the time costed edges are used when working with graph based data. If the costs of the edges aren't assigned while collecting the data, or the cost parameters are changed later, new cost values should be assigned. This process is time consuming. For this kind of situations an easy and practical solution is needed.

In this study, it is aimed to find an efficient solution for adding cost to graph datasets. For this purpose, a raster based method is suggested. Thanks to newly created QGIS plugin, performing this process is also very easy to implement.

Beside being such useful, there is a problem with generating MST using graph, that consists of edges whose costs are assigned from raster. The problem is that, if there are restricted areas in the raster (null values), some disconnected points may occur. As a future work we will be investigating ways in which to connect such disconnected nodes to the MST.

REFERENCES

Ahuja, R. K., Magnanti, T. L., Orlin, J. B., & Reddy, M. R. (1995). Applications of Network Optimization. *Network Models*, *7*(June), 1–83.

Brimberg, J., Hansen, P., Lih, K. W., Mladenović, N., & Breton, M. (2003). An oil pipeline design problem. *Operations Research*, *51*(2). https://doi.org/10.1287/opre.51.2.228.12786

Çalışkan, M., & Anbaroğlu, B. (2020). Geo-MST: A geographical minimum spanning tree plugin for QGIS. *SoftwareX*, *12*, 100553. https://doi.org/10.1016/j.softx.2020.100553

Degenne, P., & Lo Seen, D. (2016). Ocelet: Simulating processes of landscape changes using interaction graphs. *SoftwareX*. https://doi.org/10.1016/j.softx.2016.05.002

Dippon, C. M., & Train, K. E. (2000). The cost of the local telecommunication network: A comparison of minimum spanning trees and the HAI model. *Telecommunications Policy*, *24*(3), 253–262. https://doi.org/10.1016/S0308-5961(00)00009-4

Dott, D. R. (1998). Optimal network design for natural gas pipelines. *ProQuest Dissertations and Theses*, 117. https://doi.org/10.11575/PRISM/22946

Graham, R. L., & Hell, P. (1985). On the History of the Minimum Spanning Tree Problem. *Annals of the History of Computing*, *7*(1), 43–57. https://doi.org/10.1109/MAHC.1985.10011

Kruskal, J. B. (1956). On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. *Proceedings of the American Mathematical Society*, *7*(1), 48. https://doi.org/10.2307/2033241

Mareš, M. (2008). The saga of minimum spanning trees. *Computer Science Review*, *2*(3), 165–221. https://doi.org/10.1016/j.cosrev.2008.10.002

Pettie, S., & Ramachandran, V. (2000). An optimal minimum spanning tree algorithm. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *1853*(1), 49–60. https://doi.org/10.1007/3-540-45022-x_6

Prim, R. C. (1957). Shortest Connection Networks And Some Generalizations. *Bell System Technical Journal*. https://doi.org/10.1002/j.1538-7305.1957.tb01515.x

Rothfarb, B., Frank, H., Kleitman, D. J., Rosenbaum, D. M., & Steiglitz, K. (1970). Optimal Design of Offshore Natural-Gas Pipeline Systems. *Operations Research*. https://doi.org/10.1287/opre.18.6.992