Internship Report
Submitted for the degree of
Data Engineer

Universidad Politécnica de Yucatán

Analysis of the road network of the city of Merida, Yucatan, Mexico.

Alejandro De Jesus Puerto Castro

1. External Supervisor Gonzalo G. Peraza Mues
Data Engineering
Universidad Politécnica de Yucatán

2. Internal Supervisor Didier O. Gamboa Angulo
Data Engineering
Universidad Politécnica de Yucatán

February 22, 2021

Contents

List of Figures

List of Tables

Acknowledgments

Abstract

1 Introduction

1.1 Background

1.2 Problem Statement

1.3 Justification

1.4 Scope and limitations

1.5 Objectives

1.5.1 General

1.5.2 Specific

1. a)
   b)
   c)

2 Theoretical framework

On this section we present concepts to support and further explain the technologies and process found on the Methodology and Development section. We present concepts regarding spatial networks and its analysis measures based on network theory; spatial data; as well as some tools and services that we use as a base for this work.

2.1 Spatial Networks

A network (or *graph* in mathematics) is composed of nodes N connected by links, or edges, E. Nodes represent entities in a network, such as cities, people, airports, and street intersections. Edges represent relationship between nodes, such as friendships among people, flights between airports, street roads, and so on.

Graphs can be arranged nonspatially or spatially. Spatial graphs have nodes that are georeferenced, i.e. they are defined by their location in geographic space using a pair of coordinates; usually embedded in a two- or three-dimensional space [3]. Both nonspatial and spatial graphs can contain undirected, directed, unweighted, or weighted components [4].

An undirected graph has edges that can be used to represent flows of traffic on two-way streets, while edges in a directed graph represent one-way streets. A self-loop in a graph is an edge that connects a single node to itself. Two nodes can also be connected by parallel edges, in such case, graphs are called multigraphs, or multidigraphs if they are directed. The weight attribute in a weighted graph is used to quantify some value between connected nodes [5].

Network science was founded on the based on the findings by Euler [6], and gave way to the notion that graphs have different structural properties that can be discovered and cataloged using graph theory [7]. With this, early spatial network analysis focused on using graph theory and its measures to describe and catalog the properties of real systems represented as spatial networks [4].

Real spatial networks are complicated physical entities, with numerous, often complex, elements with a nontrivial configuration and structure that are neither purely fully regular nor fully random. A street network is an example of a complex spatial network with both nodes and edges embedded in space.

A spatial network is *planar* if it can be represented in two dimensions with its edges intersecting only nodes. A street network may be planar at certain small scales, but most street networks are non-planar due to overpasses, bridges, tunnels, etc. For tractability, these networks are studied as approximately planar. However, it can cause analytical problems due to the over-simplification.

Street networks can be based on two graph representations: primal graph or dual graph. In a primal graph representation, intersections are turned into nodes and street segments into edges. On the other hand, dual graphs invert the representation: streets as nodes and intersections as edges [8]. Primal graphs retain all the geographic, spatial information that are lost in a dual graph. For that reason, primal representation is the better approach

for analyzing spatial networks as it faithfully represents all the spatial characteristics of a street [9].

2.2 Spatial Networks Analysis Measures

The structure and behavior of networks can be described using a variety of graph theory measures. These measures can be found in different level of detail in [3, 7, 10–13].

Each network is characterized by the **total number of nodes** N and the **total number of edges** E. We call N the **size of the network**.

The ***degree*** k of a node is its number of edges, or neighbors, and it is a local measure. We use k_i to denote the degree of node i. A node with no neighbors has degree zero ($k = 0$) and is called a ***singleton***.

The **average degree** $< k >$ is a global measure for the average degree k across all nodes N in a graph. This measure is simplified by dividing twice the number of edges E by the number of nodes N, as follows:

$$< k >= \frac{2E}{N} \tag{2.1}$$

The average streets per node measures the mean number of streets (edges in an undirected graph) that come out from each intersection or dead-end.

The **degree distribution** $P(k)$ represents the fraction of nodes in a graph with degree k, calculated by dividing the number of nodes with degree k by the total number of nodes N in the graph G. The degree distribution $P(k)$ is often plotted on a histogram and is useful for providing an overall snapshot of graph G.

The **clustering coefficient** C measures the ability of an individual node i to associate with other nodes (cliquishness). It is commonly described as the probability that "friends" of i (i.e., nodes connected to node i) are also friends of each other: the chance that a friend of my friend is also my friend [14]. For node i of degree k_i, the clustering coefficient $C(i)$ is defined as:

$$C(i) = \frac{E_i}{k_i(k_i - 1)/2} \tag{2.2}$$

where E_i is the number of edges existing between the neighbors of i. When the local measure $C = 1$, the node v_i and its neighboring nodes are all perfectly connected. In contrast, when $C = 0$, neighbors of node i are not connected at all.

The **average clustering coefficient** $< C >$ is a global measure that determines the cliquishness of all nodes in a graph and is calculated as the average C over all individual nodes. When $< C >= 1$, the graph is perfectly connected. In contrast, when $< C >= 0$, the graph is not connected at all.

Path P is and ordered sequence, or collection, of edges that connects some ordered sequence of nodes. The collection of nodes N and edges E in a path can be defined as:

$$N_p = \{0, 1, 2, ...n\} \tag{2.3}$$

$$E_p = \{0, 1, 2, ...m\} \tag{2.4}$$

There may be many paths of varying lengths l between two nodes i and j. The **shortest path length** l_s is calculated by counting the total number of intermediate nodes or edges along the shortest path between two nodes i and j and is defined as:

$$l_s(i,j) = \min_{\text{paths}}(i \to j) \tag{2.5}$$

The **average shortest path length** $< l >$ is defined as the average shortest path length between all possible pairs of nodes in the network. The **diameter** d_G of a graph G is defined as the maximum shortest path length l_s found in the graph.

Average street length is the mean edge length measured in meters, an example of spatial units, and indicates how fine-grained (small block size) or coarse-grained (large block size) the networks is.

Density measures provided how fine-grained the network is. **Node density** is the number of nodes divided by the area covered by the network. **Intersection density** is the node density of the set of nodes with more than one street emanating from them, excluding dead-ends. The **edge density** is the sum of all edge lengths divided by the area. The physical **street density** is the sum of all edges (in the undirected graph) divided by the area.

The **average circuity** is the circle distances between the nodes of each edge, and it is defined by the sum of all edge lengths divided by the sum of the great-circle distances between the nodes incident to each edge [15].

Eccentricity is the largest distance (the maximum of the shortest-path weighted distances) between a node and other nodes i.e., how far the node is from the node that is furthest from it [16]. The **diameter** of a network is the maximum eccentricity of any node in the network and the **radius** is the minimum eccentricity [17]. The **center** if a network is the node or set of nodes with an eccentricity equals the radius, and the **periphery** of a network is the node or set of nodes with eccentricity equals the diameter. These distances serve as indicators for network size and shape if we use length as weight.

Connectivity measures the minimum number of nodes or edges that must be removed from a connected graph to disconnect the network [16]. In the case of street networks, we use **average node connectivity** as a resilience indicator, which is the mean number of internally node-disjoint paths between each pair of nodes. This measure is more useful to represent the expected number of nodes that must removed to disconnect a randomly selected pair of non-adjacent nodes [18,19] Networks with low connectivity may have multiple points of failure, this yield to a vulnerable system.

Centrality measures indicate the most important nodes in a network [20, 21]. **Betweenness centrality** g_i measures the total number of shortest paths between any two nodes in the graph that pass through node i [22, 23] and is defined as:

$$g_i = \sum_{u \neq v} \frac{\sigma_{uv}(i)}{\sigma_{uv}} \tag{2.6}$$

where σ_{uv} is the number of shortest paths going from node u to note v and $\sigma_{uv}(i)$ is the number of shortest paths going from node u to node v through node i. The importance of an edge j is also measured by betweenness centrality g_j that instead calculates the total number of shortest paths between any two nodes in a graph that include edge j [24] and is defined as:

$$g_j = \sum_{u \neq v} \frac{\sigma_{uv}(j)}{\sigma_{uv}} \tag{2.7}$$

where σ_{uv} is the number of shortest paths going from node u to node v and $\sigma_{uv}(j)$ is the number of shortest path going from node u to node v through edge j. In many graphs, betweenness centrality g_i and node degree k_i correlate, where the most central node can also have the most connections. The **average betweenness centrality** is the mean of betweenness centralities of all the nodes in the network [3]. The maximum betweenness centrality in a network specifies the proportion of shortest paths that pass through the most important node. If the maximum betweenness centrality is high, the network is more susceptible to failure or inefficiency.

Closeness centraltity is another way to measure the centrality of a node by determining how close a node is to the other nodes. This can be done by averaging the sum of the distances from the node to all others. This measure gives low values for more central nodes and high values for less central ones [12]. It is defined as the inverse of the sum of distances of a node from all others:

$$g_i = \frac{1}{\sum_{j \neq i} l_{ij}} \tag{2.8}$$

where l_{ij} is the distance from i to j and the sum runs over all the nodes of the network, except i itself. An alternative formulation to discount the graph size and make the measure comparable across different networks is obtained by multiplying equation 2.8 by the constant $N - 1$, which is just the number of terms in the sum at the denominator:

$$\tilde{g}_i = (N - 1)g_i \tag{2.9}$$

Finally, **PageRank** is an algorithm to compute a centrality measure that aims to capture the prestige or importance of each node and it is typically used in directed networks. It ranks nodes based on the structure of incoming links and the rank of the source node. This measure can also be applied to street networks [25–28]. It is worth to mention that multiple studies use centrality measures in combination to assess street networks (e.g., [29–34]).

A graph community is defined as a set of nodes that have more connections among themselves than other nodes in the graph [35]. This feature is important in spatial networks since dense connections tend to take place between nodes that are closer in proximity. Moreover, this implies that the majority of flows between nodes occur as a function of nodes belonging to the same geographical region [36].

A community is typically identified by calculating **modularity** Q [37] and is defined as:

$$Q = \sum_{s=1}^{n_M} \frac{l_s}{E} - (\frac{d_s}{2E})^2 \tag{2.10}$$

where n_M is the number of modules of the partition, l_s is the number of edges inside module s, E is the total number of edges in the network, and d_s is the total degree of the nodes in the module s.

The above measures do not account for the distance between linked node pairs, an important measure that can be used to quantify real spatial networks embedded in geographic space. Distance can be measured in a variety of ways, the most common being Euclidean distance $d_E(i, j)$ or as the direct distance between two points. In contrast, the route distance $d_R(i, j)$ is computed by summing the geographical length of edges, which make up the shortest path between node v_i and v_j [4].

2.3 Spatial Data

As our data is embedded in space, we need to understand its properties:

A datum is a model of the Earth's shape. Sometimes the Earth is assumed to have an spherical shape who is described by two coordinates, latitude (north) and longitude (east). However the Earth is not a sphere; its shape is more like an ellipsoid. There are many possible approximations to this shape, which define their own latitude-longitude coordinate system. A coordinate system (CS) is a sequence of coordinate axes with specified units of measure, and its types are: ellipsoidal, Cartesian, affine, gravity-related, linear, spherical, polar, and cylindrical. A coordinate reference system (CRS) associates a CS with an object by mean of a datum (see Figure 2.1) [1]. Some are more accurate than others for particular regions of the Earth's surface. If our data is notated in different datums then we will need to convert them into one standard format. The most common global datum is called WGS84 (World Geodetic System, 1984) [38].

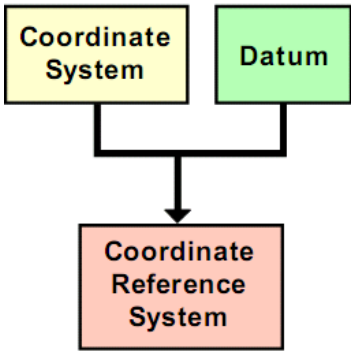

Figure 2.1: A coordinate reference system combines a coordinate system with a datum, which gives the relationship of the coordinate system to the surface and shape of the Earth. Retrieved from [1].

A projection is the change of the representation of locations from one coordinate system to another. Sometimes it is more convenient to work with a flattened 2D projection of a datum rather than its spherical coordinates. With this, we project the coordinates into Cartesian x and y meters. We take x = Easting and y = Northing, in the order (x, y), in meters from some origin. When we do a projection, we must make some compromise because it is not possible to make a perfect flat version of an ellipsoid surface. All projections of locations on the Earth into a two-dimensional plane are distortions as something always will be distorted [39]. A good projection to use is one of the Universal Transverse Mercator (UTM) family. UTM splits the Earth's surface into state-sized regions, and defines separate projection for each one, to minimize the distortion there (see Figure 2.2).

With geographic data it is common to work in only two of the three dimensions. Two dimensional space support three basic types of spatial entity [40]:

- points - having a location
- lines - comprising two or more locations in an ordered sequence
- polygons - areas defined by three or more vertex locations in an ordered sequence

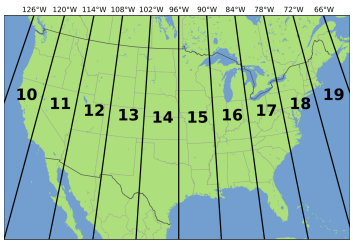

Figure 2.2: UTM zones across the continental United States. Source: [2].

2.4 Geographic Information Systems (GIS)

A Geographic Information System (GIS) is any system that is specifically designed to work with spatial data. GIS systems are implementations of some standard tasks, which may be present in as programming language libraries or functions, and/or as graphical user interfaces [38]. Standard tasks include:

- converting datums and projections
- searching quickly for entities in particular regions
- searching quickly for entities with spatial relationships to other entities
- handling lots of data at large and small scales
- geographical data visualizations
- converting between standard spatial data file formats

2.5 Spatial files

Shapefiles are a storage format for the Open Geo-spatial Consortium (OGC) definition: points, lines, and polygons. They are small collections of related files, usually stored together in a directory. The main file has a *.shp* extension and stores the actual feature geometries. Other files that may appear along with it include *.dbf* (associated non-spatial properties data), *.shx* (indexing structure), and *.prj* (datum/projection information) [38].

GeoJSON and Well-Known Text (WKT) are alternatives for storing such data. Most GIS systems can convert between those formats.

2.6 OpenStreetMap

OpenStreetMap (OSM) is a collaborative worldwide mapping project that provides a free and publicly editable map of the world. In February 2021, there were over 7M registered contributors, as outlined on the OSM wiki [41].

In Mexico, OSM imported the 2015 INEGI's Marco Geoestadístico Nacional (MGN) and the Red Nacional de Caminos (RNC) road data during the years 2015 and 2016 as part of the two OSM projects: Mexico Main Road Network Import Project [42], and Mexico's Administrative Divisions Import Project [43]. Additionally, individual contributions have been made from OSM members to keep updated and accurate the data. Acquiring the

spatial data from OSM is made via an API, called Overpass, to retrieve any data in the database. However, its usage and syntax are somewhat difficult and there are other services available that can simplify the process.

2.7 Network Analysis Tools

2.7.1 NetworkX

NetworkX is a free, open-source Python language package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks [44,45]. It provides data structures for graphs, digraphs, and multigraph, and implementations of many of the algorithms used in network science. These algorithms are implemented for structure and analysis measures, such as shortest paths, betweenness centrality, clustering, degree distribution and many more.

In addition NetworkX can read and write various graphs formats (e.g., adjacency list, edge list, GEXF, GML, pickle, GraphML, JSON, LEDA, YAML, SparseGraph6, Pajek, and GIS Shapefile), and provides generators for classic graphs, random graphs, and synthetic networks.

2.7.2 GeoPandas

GeoPandas

2.7.3 OSMnx

2.7.4 Gephi

2.7.5 NetworKit

2.7.6 PySAL

3 Methodology and development

3.1 Methodology

3.2 Deployment

4 Results

5 Conclusions and recommendations

5.1 Conclusion

5.2 Recommendations

Bibliography

[1] Pal Nikolli and Bashkim Idrizi. Coordinate reference systems used in albania to date. FIG Working Week 2011, Morocco, 05 2011.

[2] Chrismurf. Utm-zones-usa.svg, 2009.

[3] Marc Barthélemy. Spatial networks. Physics Reports, 499(1):1–101, February 2011.

[4] Taylor Anderson and Suzana Dragićević. Complex spatial networks: Theory and geospatial applications. Geography Compass, 14, July 2020.

[5] Geoff Boeing. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. Computers, Environment and Urban Systems, 65:126–139, September 2017.

[6] Leonhard Euler. Solutio problematis ad geometriam situs pertinentis (the solution to a problem relating to the geometry position). Commentarii Academie Scientiarum Imperialis Petropolitanae, 8:128–140, 1741.

[7] Albert-László Barabasi. Network science. Cambridge University Press, Glasgow, UK, 2016.

[8] Sergio Porta, Paolo Crucitti, and Vito Latora. The network analysis of urban streets: A primal approach. Environment and Planning B, 33(5):705–725, 2006.

[9] Carlo Ratti. Space syntax: Some inconsistencies. Environment and Planning B, 31(4):487–499, 2004.

[10] Ted G. Lewis. Network science: Theory and applications. John Wiley & Sons, Hoboken, NJ, September 2011.

[11] Mark. E. J. Newman. Analysis of weighted networks. Physical Review E, 70(5), Nov 2004.

[12] Mark E. J. Newman. Networks: An introduction. Oxford University Press, Oxford, England, 2010.

[13] Filippo Menczer, Santo Fortunato, and Clayton A. Davis. A First Course in Network Science. Cambridge University Press, 2020.

[14] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. Nature, 393(6684):440–442, 1998.

[15] David J Giacomin and David M Levinson. Road network circuity in metropolitan areas. Environment and Planning B: Planning and Design, 42(6):1040–1053, 2015.

[16] Dean Urban and Timothy Keitt. Landscape connectivity: A graph-theoretic perspective. Ecology, 82:1205–1218, 05 2001.

[17] Per Hage and Frank Harary. Eccentricity and centrality in networks. Social Networks, 17(1):57–63, 1995.

[18] Lowell Beineke, Ortrud Oellermann, and Raymond Pippert. The average connectivity of a graph. Discrete Mathematics, 252:31–45, 05 2002.

[19] Peter Dankelmann and Ortrud R. Oellermann. Bounds on the average connectivity of a graph. Discrete Applied Mathematics, 129(2):305–318, 2003.

[20] Xiaoke Huang, Ye Zhao, Chao Ma, Jing Yang, Xinyue Ye, and Chong Zhang. Trajgraph: A graph-based visual analytics approach to studying urban network centralities using taxi trajectory data. IEEE Transactions on Visualization and Computer Graphics, 22(1):160–169, January 2016.

[21] Chen Zhong, Markus Schläpfer, Stefan Arisona, Michael Batty, Carlo Ratti, and Gerhard Schmitt. Revealing centrality in the spatial structure of cities from human activity patterns. Urban Studies, 54, 10 2015.

[22] L. C. Freeman. A set of measures of centrality based on betweenness. Sociometry, 40(1):35–41, 1997.

[23] Alireza Ermagun and David Levinson. An introduction to the network weight matrix: Introduction to the network weight matrix. Geographical Analysis, 50, 07 2017.

[24] Mark E. Newman. The structure and function of complex networks. SIAM Review, 45(2):167–256, 2003.

[25] Bin Jiang. Ranking space for predicting human movement in an urban environment. International Journal of Geographical Information Science, 23, 12 2006.

[26] Taras Agryzkov, José Oliver, Leandro Tortosa, and José-Francisco Vicent. An algorithm for ranking the nodes of an urban network based on the concept of pagerank vector. Applied Mathematics and Computation, 219:2186–2193, 11 2012.

[27] David Gleich. Pagerank beyond the web. SIAM Review, 57, 07 2014.

[28] Wei Chien Benny Chin and Tzai-Hung Wen. Geographically modified pagerank algorithms: Identifying the spatial concentration of human movement in a geospatial network. PLoS ONE, 10, 10 2015.

[29] Sergio Porta, Paolo Crucitti, and Vito Latora. The network analysis of urban streets: A dual approach. 11 2004.

[30] Sergio Porta, Paolo Crucitti, and Vito Latora. The network analysis of urban streets: A primal approach. Environment and Planning B: Planning and Design, 33:705–725, 02 2006.

[31] S. Porta, V. Latora, and E. Strano. Networks in urban design. six years of research in multiple centrality assessment, pages 107–130. Springer, September 2010.

[32] Paolo Crucitti, Vito Latora, and Sergio Porta. Centrality in network of urban streets. Chaos (Woodbury, N.Y.), 16:015113, 04 2006.

[33] Paolo Crucitti, Vito Latora, and Sergio Porta. Centrality measures in spatial networks of urban streets. Physical review. E, Statistical, nonlinear, and soft matter physics, 73:036125, 04 2006.

[34] Andres Sevtsuk and Michael Mekonnen. Urban network analysis. a new toolbox for arcgis. Revue internationale de géomatique, 22:287–305, 06 2012.

[35] Santo Fortunato. Community detection in graphs. Physics Reports, 486(3-5):75–174, 2010.

[36] Marc Bathélemy. Morphogenesis of spatial netwoks. Springer International Publishing, Cham, Switzerland, 2018.

[37] Mark. E. J. Newman and Michelle Girvan. Finding and evaluating community structure in networks. Physical Review E, 69(2), February 2004.

[38] Charles Fox. Spatial Data. In Data Science for Transport, pages 57–74. Springer International Publishing, Cham, 2018.

[39] Miljenko Lapaine and E. Lynn Usery, editors. Choosing a Map Projection. Lecture Notes in Geoinformation and Cartography. Springer International Publishing, Cham, 2017.

[40] Geographic information - simple feature access - part 1: Common architecture. Standard ISO 19125-1:2004, International Organization for Standardization, 2004.

[41] Openstreetmap wiki. Retrieved from `https://wiki.openstreetmap.org/wiki`.

[42] Mexico main road network import project - openstreetmap wiki. Retrieved from `https://wiki.openstreetmap.org/wiki/Mexico_Main_Road_Network_Import_Project`.

[43] Mexico's administrative divisions import project - openstreetmap wiki. Retrieved from `https://wiki.openstreetmap.org/wiki/Mexico%27s_Administrative_Divisions_Import_Project`.

[44] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, Proceedings of the 7th Python in Science Conference, pages 11 – 15, Pasadena, CA USA, 2008.

[45] NetworkX developers. Networkx.