Understanding Structure of Urban Built Environment and its Implications on Movement Affinities of Space

Thesis submitted in partial fulfillment of
the requirements for the degree of

Master of Science by Research
in
COMPUTER SCIENCE AND ENGINEERING

by

RAJESH CHATURVEDI

200902049

rajesh.chaturvedi@research.iiit.ac.in

International Institute of Information Technology, Hyderabad
(Deemed to be University)
Hyderabad - 500 032, INDIA
December 2017

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled “Understanding Structure of Urban Built Environment and its Implications on Movement Affinities of Space” by RAJESH CHATURVEDI, has been carried out under my supervision and is not submitted elsewhere for a degree.

30/12/2017
Date

Adviser: Prof. K. S. Rajan

To
Family & Friends

Acknowledgments

I had a wonderful time at Lab for Spatial Informatics (LSI), IIIT-H. First, I would like to thank my advisor Dr. K. S. Rajan, for giving me an opportunity to be associated with LSI and guiding me. He has been constant source of motivation and perseverance. His guidance has been invaluable on both academic and personal levels, for which I am extremely grateful. I would like to extend my heartful gratitude to all the professors who instructed various courses during my graduation.

I am deeply thankful to Swati for her support, motivation, and patience during the times I felt unconfident. A special mention to my friends Data, Kunal, Karthik, Kshitij, Ankush, Rahul, Deepank, Rituraj, Rai, Binay, Raghvendra, Sonil, Akhil, Ankit with whom the stay at IIIT-H was more than pleasant.

Finally, I would like to thank my family for standing by me through times good and bad, it’s because of their blessings and belief in me that I have been able to complete the thesis.

Abstract

The structure of built environment is defined by the collection of fundamental units contained in it. Two of the most basic units in an urban setting are Roads and Buildings. The buildings form most of the origin and destination pairs for intra-city movements and the road-networks aid these movements. The basic aim of this thesis is to identify patterns of arrangements of these elementary units in any urban agglomeration along with deriving a syntactic meaning out of the arrangement geometry and secondly, how the movements patterns and dynamic occupancy of spaces are often affected by the synergy of roads and buildings. We try to identify combinations of arrangement of these two that increase the odds in favor of increasing attractor values of spaces.

In past, several studies were carried out to extract patterns in urban environment which were often driven by considering networks only. A few studies paid attention to buildings and their characteristics such as their volumetric capacities and their locations along the network. The approach we have chosen to define symmetries and regularities in structures is hybrid in nature as it considers both buildings and road-networks as well as their intrinsic interactions such that the buildings and roads in closer vicinity inherit the properties of each other.

The spatial indices we used for analyzing built environments are based on navigational properties of any location which suggest accessibility of the location depending on its connectivity with neighborhoods. These indices are motivated by a combination of Space Syntax configurational theory that uses topological (Angular & Segment) distances along network and Urban Network Analysis which has metric distances as basis for accessibility computations. The accessibility based on topological as well as metric computations is defined as function of how often a spatial unit lies on the shortest paths while commuting all origin-destination pairs in its neighborhood, how many of surrounding places can be reached from this place, how compactly packed is the neighborhood for a given set of fixed radii, and do the shortest paths between origin-destination pairs resemble straightest distances. These characteristics can be discussed for both buildings and roads. We use spatial autocorrelation for our chosen accessibility indices to analyze significance of occurrence of specific patterns of neighborhoods for city of Hyderabad, India, and thus consider not only spatial units but areas as whole.

We discuss the inconsistencies in methodologies used, and propose a weight metrics for urban road network, weights are driven by dimensions of buildings. This enables us to reduce computational complexity by a significant margin using road networks as the only analysis layer for making configurational computations. We verify that it is intrinsic property of spatial organization that more similar areas tend to occur in closer vicinity of each other by using K Means clustering. In the process, we also propose that the clusters formed with varying radii of analysis are suggestive of spatial properties of urban environment.

Lastly, we present a data driven methodology to classify regions in terms of amount vehicular traffic. In this technique, Average Annual Daily Counts (AADT) of vehicular traffic are used to represent roads into three categories of high, medium and low which are then superimposed on buildings. The chosen spatial indices form feature vectors for training the model while the vehicular traffic categories superimposed on buildings act as prediction target labels. To get the best accuracy model, we compare performances of Logistic Regression Classification model, Decision Tree Classification model and ensemble based Gradient Boosted Classification model. The ensemble based model using gradient boosting algorithm outperforms all other models in this classification task.

While discussing all methodologies to denominate areas in terms of their symmetry and constrained by relative carriage of vehicular traffic, we go from a local measure of trip distances towards global trip distances that implies we go from inspection of nearby regions in close proximity of area under observation towards larger distances. This process enables us to draw comparisons between areas corresponding to their obvious choices of being preferred for various trip distances. For example, an area being preferred for short trip distances might not be a good fit for longer trips and vice versa.

The approaches presented in this can draw conclusive reports from the perspective of architectural and infrastructural planning for new city sites and modifying the existing ones.

Contents

Chapter Page

List of Figures

Figure Page

List of Tables

Table Page

Chapter 1

Introduction

The cities all over the world have grown tremendously. The small settlements have progressively become large agglomerations over a period. Researchers have been busy understanding the logic of morphological expansion of the cities. The focus of these studies has mostly been based on directionality and centrality [1] of fragments of urban patches that have collectively formed heterogenous system of spaces. We call it a heterogenous system because the urban system is often characterized by significant differences between its constituent spatial units, majorly roads and buildings. Urbanism inherits a high amount of dynamicity, an unimaginable number of movements and constant structurally changing spaces are true representation of this dynamicity.

Any spatial unit, be it a building or a road segment, is characterized by a mix of spatial and non-spatial attributes. Few studies in past have attempted to decode the relationship between these spatial and non-spatial attributes [2]. The spatial attributes are governed by topology of space whereas the non-spatial attributes are usually attractor value, movement affinity and volumetric capacities of the space.

The extreme growth of cities and consistent increase in habitants has started to pose sustainability challenges and this caused the approach for cities development to become very procedural and planned. The irregularities in symmetries of structural arrangements are being identified and more focus is being given on developing opportunist urban environments. We are at a juncture where fundamental questions about the future of our cities are raised very frequently. Should settlements be dense or sparse, nucleated or dispersed, monocentric or polycentric, or a mix of all types? [3]

Researchers have come with approaches to deduce logics from existing settlements. The urban grid is constituted of structures of different kinds, the spatial organization is very complex. It can hardly be compared against simple topological and geometrical arrangement like a perfect grid or a tree. There is a presence of a clear hierarchy as a tree with appearances of grids in patches. As can be seen in figure 1.1, from left to right, the aggregation degree increases, creating an “irregular – regular” scale. From top to bottom, the grid coefficient increases, creating a “tree – grid” scale [4].

With the advent of infrastructural growth and its study at such a huge level, there is need for administrative bodies to strive to provide efficient, affordable, occupants-focused, environmentally sustainable integrated solutions. There must be connections of the cities to villages, towns, and centers of industry, commerce, tourism and pilgrimage throughout. To cater these, the decisions on design and layout of urban infrastructure must be evidence based and best practices must be listed down. Because of such regularities in planning, in the last few decades, we have witnessed settlements that are arranged in grid like symmetries and follow repeatable trends across the city in terms of geometries.

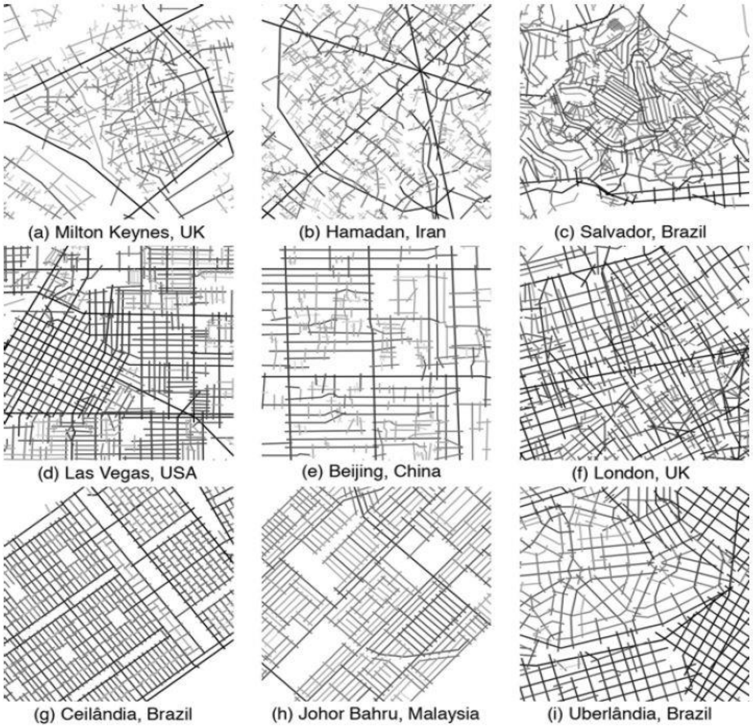


Figure 1.1 Urban grids worldwide

(The maps are courtesy of Space Syntax Laboratory (a, b – Kayvan Karimi, d – Mark David Major, f); Valério A. S. de Medeiros and DIMPU UnB (c, g, i); Tao Yang (e); and Loon Wai Chau (h)[4])

1.1 Implications of structure of built environment

There are several direct and indirect implication of structure of built environment. Studies [5,6,7,8,9,10] in past have tried to find positive and negative correlations between various configurational spatial indices and the movement patterns of pedestrians as well as vehicles. The basis of these spatial indices has majorly been confined to topology, there have been claims that characteristics of a space are significantly impacted by neighborhoods. Thus, movement pattern has been a frequently discussed implication of structure of built environment and advances are regularly made in alterations of built environment to deviate movement frequencies.

The movement patterns have also been related to land use patterns across the city. However, studies have suggested that land use does not drive movement patterns rather it is spatial configuration that defines appropriate land usage for activities and hence the movement patterns. As stated in “Space is the machine” [3], almost each movement in the urban grid is a function of the grid configuration. Several aspects of urban form are underlined by the intrinsic relation between the grid and movements: the distribution of land uses [5], retail and residence [6, 7], the spatial patterning of crime [8,9, 10], the evolution of different densities and even the part-whole structure of cities. The Study also claims that structuring of movement by the grid leads, through multiplier effects, to dense patterns of mixed use encounter that characterize the spatially successful city. The study mentions movement as a fundamental correlate of the spatial configuration. The idea of treating spatial configuration as the most powerful single determinant of urban movement, both pedestrian and vehicular has been used up in subsequent studies by several other studies. There have been claims that, with obvious biases toward higher density areas and major traffic interchanges, the structure of the grid itself accounts for much of the variation in movement densities.

There have been constant improvements in design methodologies for cities development based on spatial configurations. The spatial indices have undergone several transformations with time. New insights have been constantly added into ranking spaces for predicting human movement [11].

1.2 Problem statements and analysis

The important concerns addressed in this thesis are:

- **Does the urban built environment reflect repetitive configurational patterns?**

As addressed earlier, the city grids are very complex and have arrangements that have significant resemblances as well as variations. What approaches are suitable to capture such resemblances and variations at the same time. Also, is there any likelihood of formation of spatial clusters based on configurations. And, what is the confidence level of these configurational characteristics that enables us to relate them with movement behaviors in the city?

- **Can the configuration inherit variability in the structure of the city?**

As mentioned in above sections, there can be scale for tree-grid for network arrangement and regular-irregular scale for both networks and buildings. Can a framework be proposed that is good enough in capturing this variability and suggest the attractor values of spaces on common grounds?

- **Are spatial clusters based on configurational indices good estimators of urban attractor values?**

What difference does it make to consider spatial clusters than just simply clustering based on configurational similarity? This question suggests the locational importance of any spatial location and relative impression of all its neighborhood on its behavioral properties. These is a definite importance of neighborhoods in impacting the movement patterns along any spatial unit, how continuity of patterns of interaction with neighborhoods can be captured is an important question to answer?

- **How data-driven models can help in estimating movement affinities of space?**

Urban traffic has often been related to configuration of the urban grid. It has only been confined to defining likelihood of a spatial unit observing high amount of traffic based on positive correlations with the certain configurational features. But, bigger picture of considering an area has never been showcased. Also, classification of spatial units based on movement patterns has never been tried. We attempted this semi-supervised classification of spatial units based on configurational features and observed high accuracy. With availability of very limited spatial information, precisely only geometric information, the study aims to classify regions based on their carriage capacities.

1.3 A brief description of proposed approaches

This work is divided into two case studies. The first case study is conducted for city of Hyderabad, India. This case study demonstrates that though there may be several possible arrangements of urban grid, few may be regular, few other may be distorted and rest may be tree like settings, but the constituents of these grids viz. buildings and road networks predominantly have characteristics often represented by patches of clusters they lie in. The spatial clusters formed by buildings and roads, in terms of spatial configuration, are very significant and can often be segregated. This segregation depending on the association of configuration we chose is representative of certain values of place such as economic value, residential value, retail value etc.

The second study proposes a better version of configurational parameters which factors in information of network as well as buildings. The proposed methodology is implied on New York city, where we discuss advantages of our approach computationally as well as in terms of identifying conclusive evidence to denominate areas using clustering and classification approaches.

1.4 Organization of Thesis

Chapter 2 presents a summary of configurational theory of space syntax and Urban Network Analysis. This chapter presents fundamental definitions of topological [12] and metric [13] distances based spatial attributes used to understand logic of built environment in urban system. This chapter also discusses the potential inconsistencies in configurational theories.

Chapter 3 is case study of Hyderabad. It discusses the steps involved in preparation of data for the study. It addresses potential challenges involved and different GIS (Geographic Information System) approaches used in data preparation. Further, this chapter gives the definition of various urban attractor values and proposes a framework to extract urban values using spatial configurations based on set of semantic rules.

Chapter 4 presents proposes a new building-weight metrics on road elements and carries out a weighted configurational analysis for city of New York. Moving further, this chapter demonstrates that K means cluster analysis on feature set of proposed weighted spatial configurational attributes for roads and buildings can be a good predictor of urban attractor values and segregating areas based on their spatial variability. This chapter also presents a set of machine learning based classification approaches that can predict vehicular traffic dependency on structure of built environment with high precision.

Chapter 5 concludes the thesis and discusses possible future works.

Chapter 2

Background and Literature Review

This chapter discusses the spatial configurational theories. It also discusses potential challenges involved in bringing these theories in practice and how each theory contrasts other.

As the configurational theories are all driven by different notions of shortest paths, the following section graphically discusses the significance of different notions of shortest paths.

The different shortest distances in a graph can be stated as follows [14]:

- **Angular** = the shortest path is the one that minimizes the angle between you and your destination
- **Topological** = the shortest path is the one that uses the fewest number of turns (note that topological is the analysis as axial, but with a finer resolution)
- **Metric** = the shortest path is the one that is physically shortest
- **Segment** = the shortest path is the one that uses the least number of streets (the least number of "interjunction" stretches of street) to get to your destination.

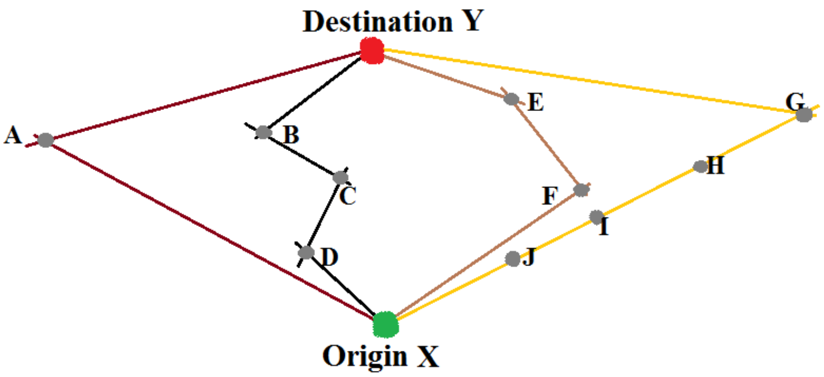


Figure 2.1 Graphical understanding of shortest distance notions

In figure 2.1, [X, Y] represent origin destination pair which can be alternatively traversed via paths [X A Y], [X D C B Y], [X F E Y], and [X J I H G Y]. Note that, the last referred path has multiple vertices falling on the straight-line path between X and G. This causes the path to traverse more segments as a segment is an edge between two junctions as represented by vertices in graph above. The path [X A Y] is shortest angular path as the sum of angular deviations made on this path is the minimum. The paths [X A Y] and [X J I H G Y] are shortest topologically as they both encounter only two turns in journey. The path [X D C B Y] is shortest metric path as the actual metric trip distance is minimum on traversing this path and lastly [X A Y] is shortest segment path as only 2 segments are traversed along this path.

Typically, a search radius for any configurational analysis is used such that only a fraction of streets in the surrounding neighborhood of the street under observation are traversed. These radii too can be angular, topological, metric and segment. An angular deviation 90± degrees is given weight of 1. So, specifying an angular radius of 5 would mean angular deviation for up to 450± degrees. The topological radius directly measure the number of turns, if a topological radius of 5 is mentioned, it would mean, no street farther than distance of 5 directional turns would be considered for the analysis. The metric distance is the physical traversal distance along the path which means a metric radius is often specified in unit meters. And, the segment radius is specified in terms of the number of junctions encountered during traversal. Using angular or topological radius will give similar results. Angular and Topological analysis approximate each other (especially at high radius). And, using metric or segment radius will give similar results. Metric and segment analysis approximate each other (especially at high radius) [14]. We have used topological, angular radius for direct distance based computations on streets and metric radii for computations on buildings in our study. The figure 2.2 below displays how the selected segments (marked in red) vary corresponding to a topological, metric or angular search radius chosen for analysis on a street (marked in yellow).

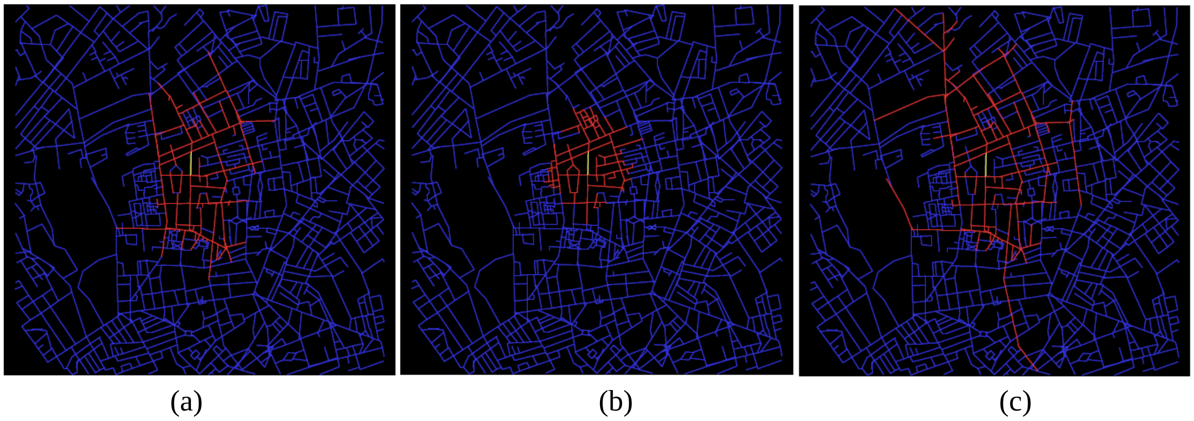


Figure 2.2 Topological Radius (a), Metric Radius (b) and Angular Radius(c)
(The maps are courtesy of Space Syntax Laboratory (Alasdair Turner))

2.1 Space syntax – Axial Analysis

Space syntax methodologies for urban analysis were proposed by Hillier and Hanson [12]. The approach used is science-based and human focused. The relationships between spatial layout and a range of environmental, economic, and social phenomena is investigated through the approach. The studied observable phenomena are usually confined to movement patterns, land use and land value, urban growth and societal differentiation and several distributions such as crime patterns and safety analysis. Space syntax entirely is encompassed of a set of theories as well as techniques for analyzing spatial configurations of indoor as well as outdoor environments. The idea used in practice is that spaces are composed of many components, by breaking down the spaces into these components, it is then analyzed as networks of choices, further represented as maps and graphs that describe the relative connectivity and integration of those spaces. The three basic conceptions of space are:

- **Isovist** – The set of all points in space that are visible from a specific point in space and with respect an environment is called as an isovist. The shape and size of an isovist is liable to change with position. Figure 2.3 (a) shows geometric view-shed representation of isovist from a point of inspection with blocks causing visual hindrance. Figure 2.3 (b) shows application of isovist in an urban setting where visual hindrance is caused by buildings. Isovists are used for construction of axial lines.

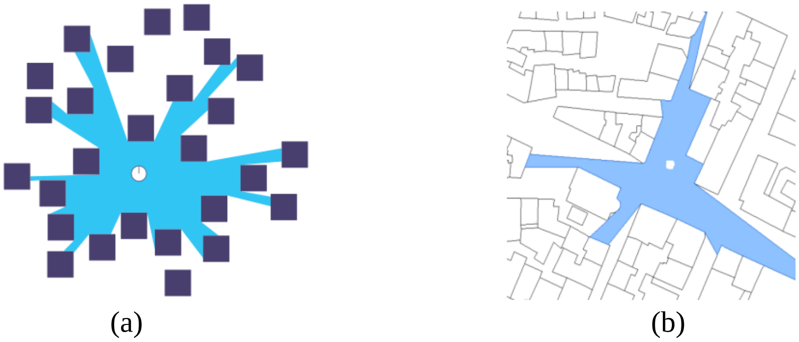

(a) (b)

Figure 2.3 Isovist Geometric representation (a) and urban environment representation (b)
(The representations are courtesy of Wikipedia [15] and blog by Anders Holden Deleuran [16])

- **Axial Line -** The longest straight-line segment representing the maximum extension of a point in space is called as the axial line. It can be objectively created. Figure 2.4 (a) shows organization of axial lines in urban space and their corresponding edge – vertex based graphical representation in figure 2.4 (b).

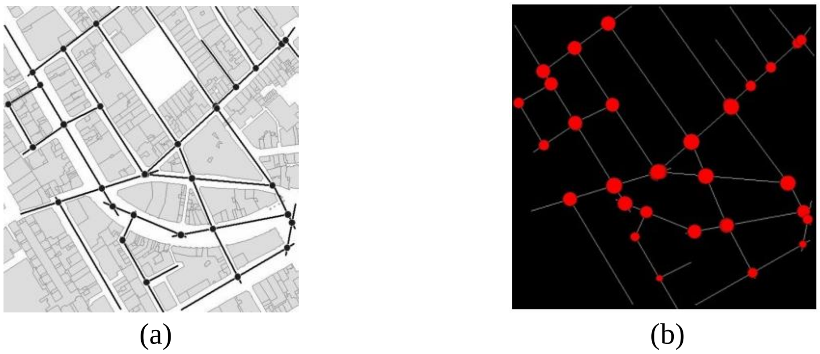

(a) (b)

Figure 2.4 Axial Lines and Junctions in the Regent Street Area of Central London, urban perspective (a), graph perspective (b)
(The representations are courtesy of Michael Batty [17])

- **Convex Space** – The space is one in which no straight-line segment drawn between any two chosen points goes outside the boundaries of the space is called as convex space. Convex space analysis is usually utilized in the accessibility analysis of interiors of buildings, it will not be discussed further.

The configurational analysis at urban environment level rests on isovists and axial line maps. The collection of axial lines in a space is called as axial map. The procedure to construct axial map include taking an accurate map and drawing intersecting line segments through all available open spaces of grid such that entire grid is covered. In the process, buildings are treated as closed spaces and roads are referred to as open spaces. The open spaces are used for movements, all spaces which favor city-wide movements are also included in the set of open spaces. The axial maps thus formed is used for space syntax based configurational analysis. Figure 2.5 shows segregation and representation of closed and open spaces from space syntax perspective.

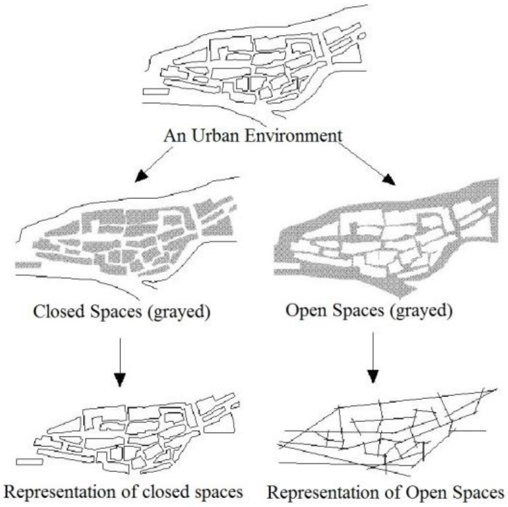


Figure 2.5 Segregation and representation of closed and open spaces in urban environment (The representations are courtesy of Bin Jiang [18])

2.1.1 Theoretical understanding of configuration parameters for axial analysis

The three most popular ways of analyzing a street network are Integration, Choice and Depth Distance.

2.1.1.1 Integration

To quantify the number of turns to be made from a road segment to reach all other segments in the defined system of network using shortest topological paths, integration measure is used. The radius defines the extent of the analysis in the system. If analysis is done such that for each segment number of turns required is calculated for reaching all other segments in the entire network, the analysis is said to be conducted at radius ‘n’. With each intersecting segment, the number of turns increase by one, i.e., the first intersecting segment is said to be one turn away, the second intersecting segment is two turns away and so on. The most integrated road segments are those which require the fewest turns to reach all others. In other words, the integration shows the cognitive complexity for reaching any other segment in the system from the specific segment.

2.1.1.2 Choice

The number of times a road segment in an urban system is passed during shortest topological distance traversals between any two other segments of the system is measured by choice attribute. For calculation purposes, initial value of 1 is set while starting the traversal and this value keeps spitting in equal portions at junction subject to number of segments meeting at the junction. This procedure is carried out until all the segments in the defined system are traversed. Also, the entire process is iterated with all segments as starting points. The final accumulated values on the segments is termed as the choice of individual segment. The streets with the highest total values of accumulated values are said to have the highest choice values.

2.1.1.3 Depth Distance

The attribute explaining the linear distance from center point of each segment in the defined spatial system to center points of all the other segments is called as depth distance. In the procedure of successively choosing every segment as the starting point, the graph with accumulative final values is reached. The road segments with minimum depth distance are said to be most near to all other segments. Like choice and integration calculations, the depth distance calculation can be limited to a given search radius to define the part of system for analysis.

It must be noted that, for calculations of all the three discussed parameters, the distance is not the actual metric distance. Rather, it is topological distance measured by the number of axial lines traversed or can be simply stated as the no. of turns encountered while traversing. All radial distances stated in context of axial lines are axial distances.

2.1.2 Mathematical understanding of configuration parameters for axial analysis

The accessibility notion can be detailed out through a depth wise connection analysis using mean-depth analysis, which is the average number of units required to cross from one unit to the other. For reaching all the units of a depth, in a justified graph, the value of a specific depth denotes the number of units the trip maker will need to cross.

$$D = (\sum d.n)/(k-1) \qquad (2.1)$$

Where,

D = mean depth

d = depth

n = number of unit spaces at a specific depth

k = total unit spaces that comprise the system

The measure of mean depth is a relative as to how a unit is in the system, hence, to bring all calculation to same scale so that units can be compared with each other, scale of symmetricity was introduced [12]. The lowest and highest measures of mean depth are 1 and *k*/2 respectively, the scale of symmetricity is defined with these values as reference. ‘k is the total number of unit spaces of the system. The relative asymmetry is the relative measure of a mean depth, and it is calculated using equation below.

$$RA = 2(D - 1)/(k - 2) \tag{2.2}$$

Where,

RA = Relative Asymmetry
D = mean depth
k = total unit spaces that comprise the system

It is rather difficult to compare relative asymmetries of two different compositions of spatial systems as they are constituted of unequal number of units. To generalize the concept of relative asymmetry, a measure called real relative asymmetry was introduced. For a unit space, it is the ratio between its relative asymmetry and a factor, expressed as D_k factor that distinguishes various systems with their sizes as the basis.

$$D_k = 2[k\{log_2(((k + 2))/3) - 1\} + 1]/(k - 1)(k - 2) \tag{2.3}$$

And,

$$RRA = RA/D_k \tag{2.4}$$

Where,
RA = Relative Asymmetry
RRA = Real Relative Asymmetry

Mathematically, the integration of a unit space is reciprocal of *RRA*, and it describes how closely or distantly the unit is accessible from all other units within a given system.
The dynamic measure of the flow through a space is given by choice. A space has a high choice value when many of the shortest topological paths, connecting all spaces to all other spaces of a system, passes through it. Choice measures how likely a road segment is to be passed through on all shortest routes from all spaces to all other spaces in the system within a predetermined distance (radius) from each segment. Mathematically, it is simply a measure of number of times shortest paths between any two units in the chosen sub system of axial lines that pass through it.

2.2 Inconsistencies and Limitations of Space Syntax

Axial analysis suffers few inconsistencies as explained by Ratti [19, 20, 21]. Space syntax is entirely dependent on topological representation and discards all metric information. Criticism of this approach from a scientific point of view is usually because all paths/axes are weighted equally in the analysis. As seen in the figure 2.6, configurational attributes of choice and integration will have same values for horizontal lines in both and (a) and (b) as the analysis do not take in account the lengths of these horizontal lines but only topological distances which are measured in terms of number of turns on the route between any origin-destination pairs. Hillier [22] claims that the existence of pervasive regularities in urban systems ensures that the axial map does not ignore the geometric properties of space but internalizes them, this argument is not very convincing to avoid metric information. Another inconsistency stated by [19, 20, 21] is buildings don't come into picture at all in space syntax analysis. The method does not account for the three-dimensional geometry of the built environment. A street that has no buildings on it is weighted equally with a street that has several tall buildings; an area covered with residential land uses is weighted equally with an area full of commercial land uses. The addition of three-dimensional built-form indicators as well as land use characteristics would allow graph measures to capture a more realistic description of the built environment and address some of the criticisms. As per space syntax analysis, structurally similar areas (without considering information like volumetric capacities of buildings) will have similar natural movement patterns [23], however, our study assumes that two similar arrangements may have different movement patterns based on the shapes and sizes of the buildings in the area, do not assume that that urban attractors are a mere consequence of configuration.

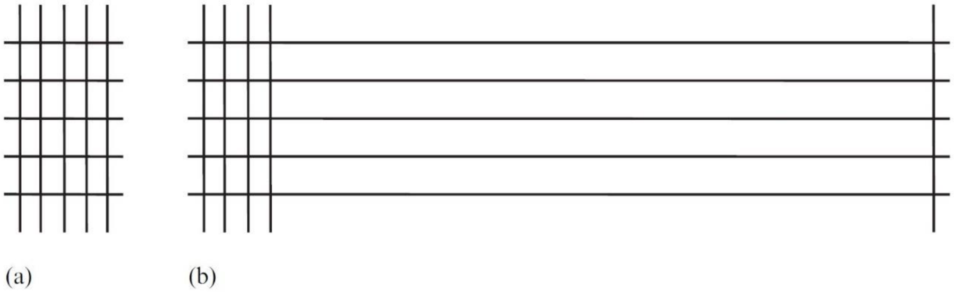


Figure 2.6 A five-by-five street portion of New York City (a) has been deformed by inserting a large unbuilt area between two streets (b). Topology is unchanged, but geometry is radically different. (The representations are courtesy of Carlo Ratti [21])

2.3 Space Syntax - Angular Segment Analysis by Metric Distances

As understood by space syntax researchers' as well, when trying to detect semi-continuous lines in the system, the use of axial lines appeared to be less helpful. This was observed when cities with uniform structures with very little disruptions were analysed. In all such cases smooth, linear streets were found to cross regular streets diagonally. Therefore, it was understood that a new angular representation was required in place of existing topological representation that could help in detecting linear or semi-linear connections. Eventually, a finer grained representation was introduced for space syntax modelling. In the underlined representation, each road segment was considered the elementary unit of the system and is defined by the interjunction between two intersection points. The cumulative angle between two intersecting segments is the geometric property used for configurational analysis. A broken representation of axial map where each axial line is split at the intersecting junctions is called as a segment map. Each edge is weighted by the angle of connection with other edges and angular segment depth is calculated by summing up these weighted values of edges. To make it clearer, a 90± degrees is given weight of 1. With respect to this all angular weights are calculated. An angle of 48± degrees may be approximated as 45± degrees and is given weight of 0.5. If one of these segments is intersecting with a different segment at 107± degrees, then it will have a weight of 1. If these three segment elements are connected in the same direction then the depth between these street elements is the sum of their weighted angular intersections, that is; 0.5 +1=1.5. This weighted sum is cost of a journey through the graph system. The shortest path is the one that has least angular cost from one segment to all other segments in the defined network [24, 25]. The angular turn is always regarded as positive. To compensate for lack of metric information in axial analysis, the configuration features may be weighed in segment analysis. One of the most common and valuable weighting methods is to weigh by length of segments. Mainly because longer segments are likely to have more entrances and blocks adjacent to them on each side which is leading consequently to higher rates of movement activity in the vicinity. This weighing tends to solve the inconsistency of metric distance information stated by Ratti [19].

2.3.1 Mathematical understanding of configuration parameters for segment angular analysis

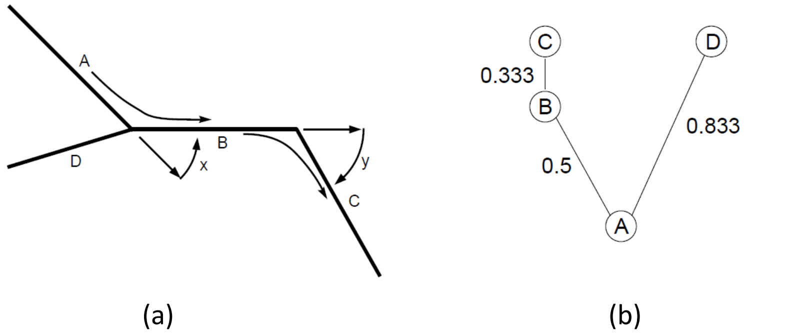


Figure 2.7 Paths through a network and associated graph with angular distances
(The network and associate graph are courtesy of Alasdair Turner [25])

The number of segments passed by on the route from the current segment to all others in the defined spatial system is called as the node count. In the case of figure 2.7 (a), the node count (NC) is 3 because the shortest angular path from A to C goes through three segments. Angular depth between adjacent segments for corresponding section of paths is displayed in figure 2.7 (b). Total angular depth is the cumulative total of the shortest angular paths to all segments. In the case of segment 'A' in figure, the angular total depth is:

$$\text{TD 'A'} = (\text{B})0.5+(\text{C})0.833+(\text{D})0.833= 2.166 \tag{2.5}$$

The angular mean depth value for a segment is the sum of the shortest angular paths divided by the sum of all angular intersections in the defined system rather than the number of lines in the system. Mean depth is indicative of how shallow or deep a node is with respect to the rest of the graph, a measure defined as centrality. In figure, angular mean depth for the segment ‘A’ is:

$$MD\ ‘A’= TD\ ‘A’/NC = ((B)0.5+(C)0.833+(D)0.833)/3 = 0.722 \qquad (2.6)$$

In angular segment analysis, integration is predictor of potentials for each segment to be a highly desired destination within defined boundaries i.e. given search radius. The measure forecasts to-movement possibilities for each segment while measuring the shortest angular patch in the defined system between all origin-destination pairs.

Integration for angular segment analysis is:

$$\text{Integration} = NC/MD = NC * NC/\ TD \qquad (2.7)$$

For calculating the choice value for segments, when calculating every shortest angular path within the system, a value of ‘1’ is assigned to each segment on the route from any origin to any destination. If the shortest paths go through an element twice the angular choice records the value ‘2’ for that segment. This summing up continues until all shortest angular paths are identified and calculated in the defined system.

2.4 Limitation of segment angular analysis

Though segment angular analysis efficiently handled the issue in configurational analysis caused due to lack of utilization of metric distance information, it still does not consider buildings in the analysis. Thus, the areas with similar network infrastructure core with different symmetry of buildings will eventually be deemed similar.

Consider urban patch shown in figure 2.8, section A has smaller buildings arranged regularly, section B has high rise buildings with large volumetric capacities and section C has small buildings in patches. Road Network alone can never depict the attractor values inherited in buildings. Such regularities and irregularities may often arise where buildings at few places in an area under observation are of same dimensions and placed equidistant from each other whereas at other places they may be varying in sizes with high variance in building to building distances. This variability cannot be captured by segment angular analysis.

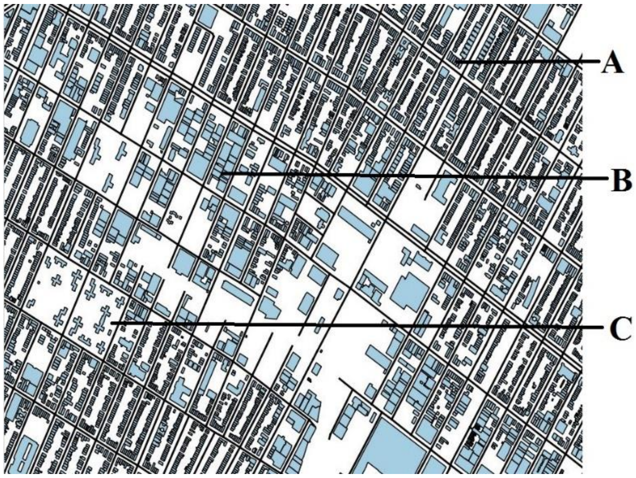


Figure 2.8 Arrangements of Buildings in a grid (New York city)

2.5 Urban Network Analysis

Unlike space syntax theories, urban network analysis [26] depends on an accurate consideration of distance and angularity between places. Also, the distance based configurational computations take place on buildings as origin – Destination pairs unlike streets in Space Syntax Analysis.

2.5.1 Reach

The reach measure [26] calculates the number of buildings in surroundings of each building reaches within a given search radius on the network such that the reached buildings are at a shortest path distance of at most the given search radius. It is defined as follows:

$$R^{r}[i] = \left|\left|\{j \in G - \{i\}: d[i,j] \leq r\}\right|\right| \tag{2.8}$$

Where,

$R^{r}[i]$ = reach of a building 'i' within search radius 'r'
$d\,[i,j]$ = Shortest network distance between nodes 'i' and 'j' in graph G

If the nodes are weighted, then reach is defined as follows:

$$R^{r}[i] = \sum_{j \in G-(i), d[i,j] \leq r} W[j] \tag{2.9}$$

Where,

$R^{r}[i]$ = reach of a building 'i' within search radius 'r'
$W[j]$ = weight of node destination that is reachable from 'i' within the radius threshold 'r'

2.5.2 Betweenness

The betweenness [26, 27] of a building is defined as the fraction of shortest paths between pairs of other buildings in the network that pass by building. The betweenness measure is defined as follows:

$$B^{r}[i] = \sum_{j,k \in G-(i), d[i,j] \leq r} \frac{n_{jk}[i]}{n_{jk}} W[j] \tag{2.10}$$

Where,

$B^{r}[i]$ = betweenness of a building 'i' within search radius 'r'
$n_{jk}\,[i]$ = number of shortest paths from 'j' to 'k' that pass by 'i'
n_{jk} = total number of shortest paths from 'j' to 'k'

2.6 Limitations of Urban Network Analysis

The Urban Network Analysis uses buildings as origin destination pairs and calculates an adjacency matrix before computing the configurational attributes. As the number of buildings in city is very large, the size of adjacency matrix utilizes large portion of memory. Also, the shortest distance between each pair of the buildings are computed which is computationally costly. This computational cost multiplies when urban network analysis is carried out for large search radius due to added overhead of calculation for buffer area of buildings equal to the size of search radius. Due to this reason, urban network analysis is often restricted to short search radius which are more suggestive of walking distances. Another imitation of urban network analysis is that it conveys no information about the roads which have no buildings adjacent to them as buildings layer is the base layer of the study. In the process, important roads such as highways, flyovers are not considered in the analysis.

Consider a portion of New York city shown in figure 2.9, the highways road centerlines do not have any buildings adjacent to them. If urban network analysis parameters are analyzed for buildings and movement patterns need to be analyzed, these important regions in terms of movement affinities would not participate.

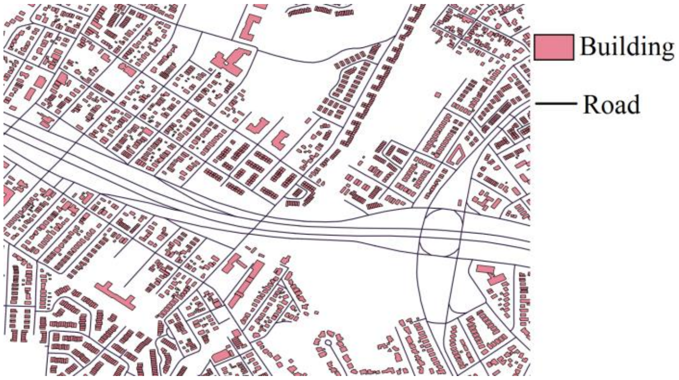


Figure 2.9 Arrangements of roads and buildings in western regions of New York city

2.7 Summary

This chapter discussed various configurational theories in practice. All configurational theories suffer one or more limitations. To overcome these limitations as identify characteristics of the areas, there need to be hybrid approaches in addition of methods that can inspect areas as whole and not individual spatial units. Proposals of such approaches and methods are made in the chapters to follow.

Chapter 3

Patterns Observed in Urban Built Environment and semantic rules to understand Urban Attractor Values – Case Study of Hyderabad

Previous chapters discussed various methods for understanding Urban syntactic logic. This case study uses implementation of Segment Angular Analysis for set of radii and Urban Network Analysis for a fixed short search radius, and tries to extract urban attractor values based on semantic understanding derived from features of the syntactic configuration. In the process, a framework is proposed that integrates the analysis on roads and buildings and further denominates the areas based on their attractor values. The framework and its flow will be discussed in further sections.

The space syntax based segment angular analysis for a set of metric radii, and urban network analysis for buildings for a fixed radius is conducted for city of Hyderabad, India. This study preferably aims at understanding significance of spatial clusters and knowledge based denomination of areas that are expected to have high attractor value as compared to others.

The study leverages the use of GIS (Geographic Information System). The analysis conducted uses GIS data layers of roads and buildings in form of shape files (.shp). Modules of GDAL (Geospatial Data Abstraction Library) are used to superimpose features from roads to buildings.

3.1 Defining Urban Attractor Values

3.1.1 Land use attraction

This corresponds to location, size and type of different land use activities [28]. The places can be differentiated as market places, residential places, community gathering places, corporate hubs and many others. How do spatial configuration help is identifying hot spots of spaces that could possibly favor one activity or the other?

3.1.2 Spatial layout, transport and movement attraction

This corresponds to geometry of the street, influencing more movement on more direct and connected streets or spaces [28]. Given a spatial layout, the movement affinities are defined directly based on proper connectivity with nearby regions. What spatial configuration helps in identifying such streets that rank higher on fair connectivity with neighborhoods? The case studies are predominantly aimed at capturing the transport attraction values of spaces. The synergy of roads and buildings is utilized to capture local and global behaviors of movements, global behaviors correspond to large trip distances whereas the local movement patterns correspond to short trips which also takes pedestrians in account.

3.2 Hyderabad GIS Data

The city of Hyderabad is divided in 18 Circles. Further, circles are subdivided 160 corporate wards. The Geographic data is provided by Greater Hyderabad Municipal Corporation (GHMC). The Geographic vector data layers included for this study are Corporate Ward Boundaries, Building Footprints, Road-Centerlines, and Road-Polygons. The data files are obtained in shape (.shp) file format that can be utilized by Geographic Information System (GIS) as-is for analysis, interpretation and visualization. As it is fairly impossible to conduct Axial as well as urban network analysis on a city scale as there are enormous number of streets and close to 11 lacs of buildings. Therefore, a localized analysis on ward scale is performed, and each ward results are analyzed separately. Figure 3.1 (a) shows the road centerlines vector data layer and figure 3.1 (b) represents building footprint polygons converted to their respective centroids.

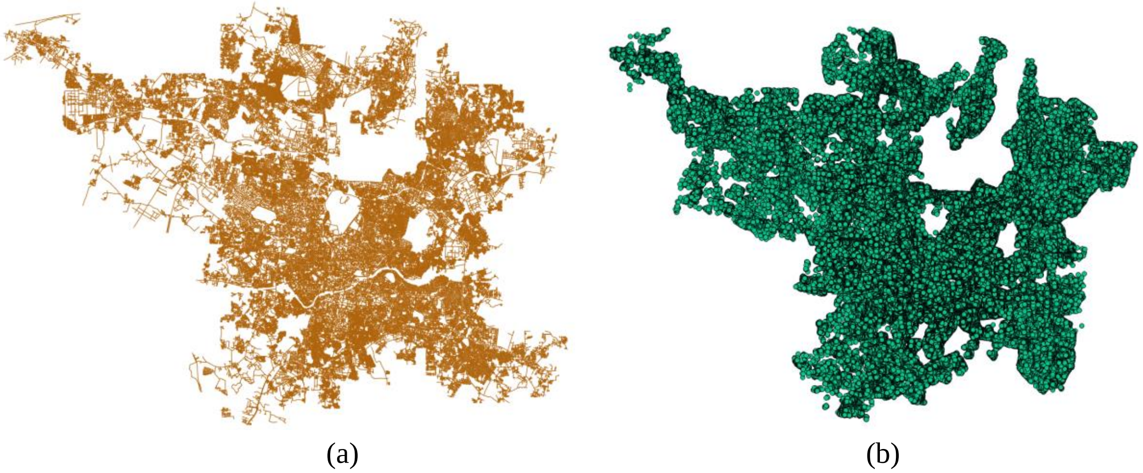


Figure 3.1 Hyderabad, Road polygons (a) and centroids of buildings (b)

Following sections discuss challenges and procedures involved in calculation of configurational features.

3.3 Edge Effects (Boundary Problem)

The spatial properties of individual units of a system that are dependent on neighbors are bound to suffer from boundary problem when neighbors are lost or not considered in the cumulative analysis. During the measurement of geographic phenomena, and analysis within a specific search radius, identical spatial data can appear dispersed or clustered depending on the boundaries around the region. In analysis with area data, interpretations should be based on the boundaries around the area under consideration. For example, in figure 3.2, the area under observation is spatially analyzed with search radius of ‘r’, the analysis results are assumed to be dependent on neighborhoods as it happens in Space Syntax and Urban Network Analysis calculations, the data within ‘r’ from the edge will suffer edge effect and give fluctuated observation results from original. To handle same, data-points outside the spatial boundaries of area under observation up to a maximum distance of ‘r’ must also be considered in analysis, the results corresponding to which must be discarded later.

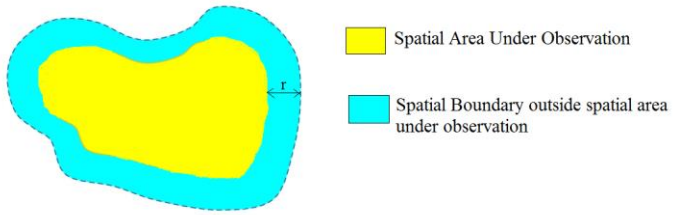

r
Spatial Area Under Observation
Spatial Boundary outside spatial area under observation

Figure 3.2 Spatial Boundary outside area under observation

In geographical research, there are two types of areas taken into consideration with respect to to the boundary: an area surrounded by fixed natural boundaries (e.g., coastlines or streams), outside of which no neighbors exist [29], or an area included in a region defined by arbitrary heuristics based artificial boundaries [30]. In all areas isolated by the natural boundaries, the spatial process does not continue at or beyond the boundaries. Whereas, if a study area is delineated by the defined artificial boundaries, the process also continues beyond the area. If a spatial process occurs beyond the study area or has an interaction with neighbors outside selected artificial boundaries, the most common approach is to neglect the influence of the boundaries and then assume that the process occurs at the internal area. However, such an approach leads to ambiguous results and a significant model misspecification problem [31].

3.3.1 Handling edge effects in study

By taking a radial buffer equivalent to that of search radius for all calculations of configurational parameters solves the problem of edge effect. For this study, as configurational parameters are studied at a local scale at ward level, we consider a radial distance of 1km, figure 3.3 (a), that convers as a sufficient buffer size for metric radius of 500m for UNA study as well and a radial buffer size of 10km for segment angular analysis because maximum metric search radius in set of radii is 10km, figure 3.3 (b).

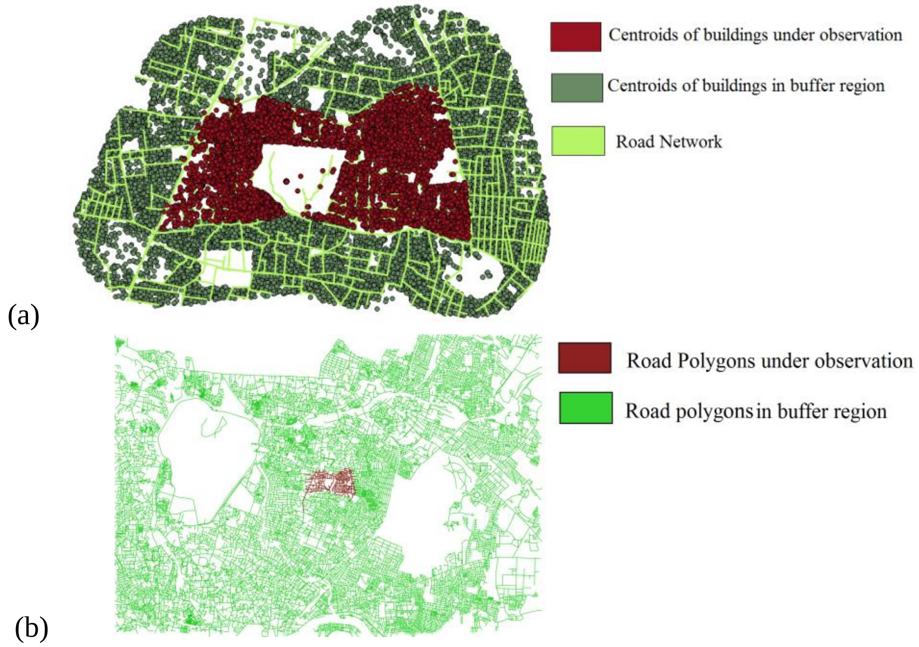
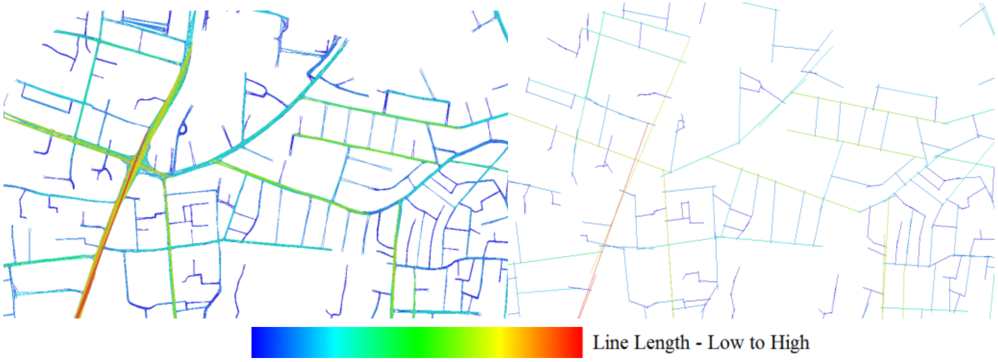

Centroids of buildings under observation
Centroids of buildings in buffer region
Road Network
(a)
Road Polygons under observation
Road polygons in buffer region
(b)

Figure 3.3 Buildings and axial representations of Road Networks of ward 44 of Hyderabad city and 4km buffer around them

3.4 Drawing Axial Map

The preprocessing step includes conversion of road polygon shapes, for Hyderabad streets, to polylines and further simplifying the polylines using modified Douglas-Peucker algorithm (built-in QGIS) with a tolerance limit of 0.02 for angular deviation to reduce the computation overhead and hence minimize the time taken to generate results. Polylines generated in preprocessing are used as visible spaces and thus utilized to generate axial line map for each ward [32] with generation of an all line map as an intermediate step. An all line map is representation where all pairs of points on the edges of road polygons that are visible to each other are joined (figure 3.4 (a)). The minimal set of line segments joining these pairs of points on edges of road polygons is picked such that entire visible space is covered (figure 3.4 (b)). This minimal set of line segments is the desired Axial Map.



(a) (b)

Figure 3.4 All line map (a) and fewest line maps (b) of a section in ward 44 of Hyderabad

3.5 Converting Axial Maps to segment Map

The axial lines are broken at points where 2 axial lines meet each other. The resultant lines broken at junctions are called as segments. The vertices of the segment represent actual road junctions. The segment maps acts as input layer for segment angular analysis computations. This is also the resultant line layer, configurational features from which are superimposed on buildings at a later stage in the framework.

3.6 Computation of configuration parameters from segment angular Analysis with metric radius

This analysis uses segment maps obtained from axial maps. The configuration parameters of Integration and Choice are computed for metric radius of 2km, 5km, 10km. These calculations are based on mathematical understanding of spatial attributes discussed in section 2.3.1. The idea of using a set of radii is to capture variability of urban attractor values at scales as we go from a local scale to a global scale of trip distances.

3.7 Computation of configuration parameters from Urban Network Analysis

3.7.1 Generating network dataset

For modelling transportation networks, network datasets are most suited. Network datasets can be generated from simple source features of lines and point geometries and turn in the network, the resultant datasets store the connectivity of the source features (figure 3.5). One such network dataset is required for traversing building to building trips for calculation of UNA parameters as discussed in section 2.5. In the generated network dataset. Streets are represented as edges and junctions are represented as nodes. Road centerlines of both the cities are used for generating the network data sets.

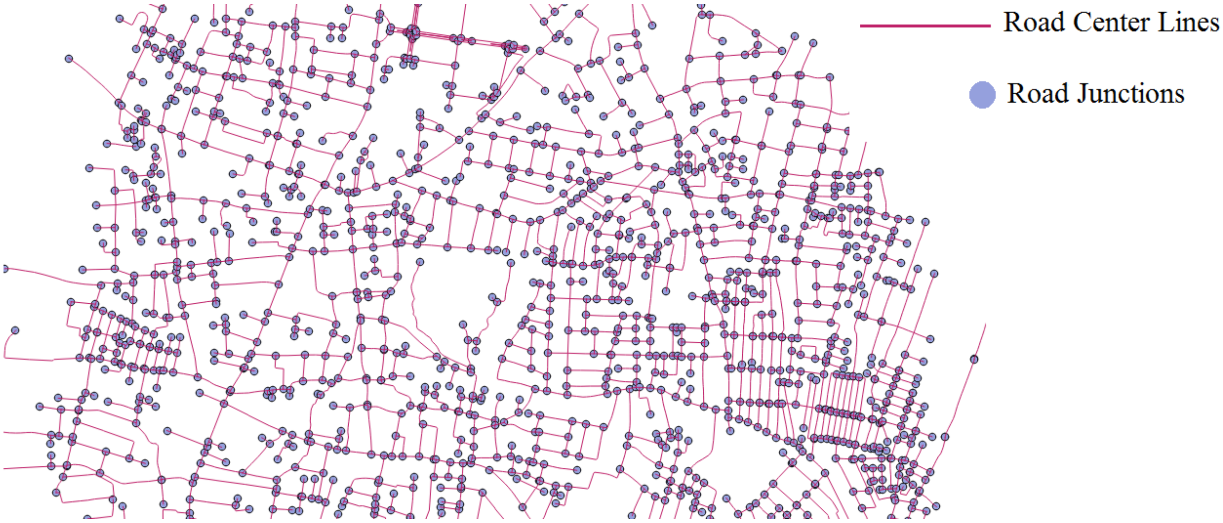


Figure 3.5 Network Data-set for road network of ward 44 of Hyderabad

3.7.2 Calculating UNA configurational parameters

The buildings polygons are represented by their respective centroid point features for all computations, ensuring that each centroid lies within the polygon boundary of the building it represents. The metric radius of 500m is considered, such short radii resonate well with very short distances movement that are usually pedestrians. However, UNA measures give a fair idea about urban infrastructural density at local scale. We do not consider large radius for UNA features because this may lead to enormous computation overhead as the number of buildings are way too many as compare to number of streets under consideration.

3.8 Spatial Join

Spatial join is a Geographic Information System (GIS) operation. Using this operation attribute from one feature layer's attribute table can be affixed to another layer's attribute table based on neighborhood. Each feature of a target layer is spatially compared to other feature layers. If the spatial reference of feature in target layer is same as that of a feature in other layer, it inherits the spatial attributes from that feature. Spatial join can thus be referred to as a method to add spatial properties from one feature layer to another. In practice, a new feature layer is created with combined data from one of the feature layers. Spatial join, thus, simply joins the spatial attributes of two layers based on location of spatial entities in the two layers. As in the figure 3.6, the features of line 'B' will be assigned to input feature as 'B' falls closest to it.

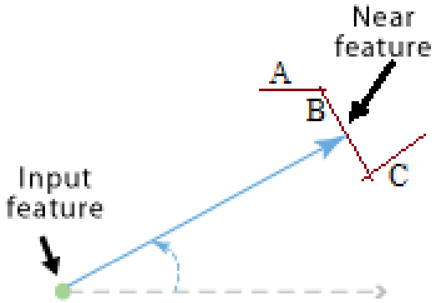


Figure 3.6 Concept of spatial join

3.8.1 Superimposing network configurational features on buildings

To study results generated from space syntax analysis and urban network analysis together, the two layers - lines from space syntax analysis and points from urban network analysis are spatially joined based on location such that a building point inherits integration and choice attributes of the axial line closest to it. This helps in characterizing buildings based on space syntax parameters. Thus, the final building layers that we have contains all computed spatial attributes from segment angular analysis as well as urban network analysis. These buildings data layers are the ones that will be used for further analysis in this study for understanding spatial geometries of city, denominating areas in city based on knowledge driven semantic rules. Following is the configurational feature set contained by buildings layer.

Feature set = {Choice_2km, Integration_2km, Choice_5km, Integration_5km, Choice_10km, Integration_10km, Reach_500m, Betweenness_500m}

3.9 Quantile spatial Visualizations of computed features

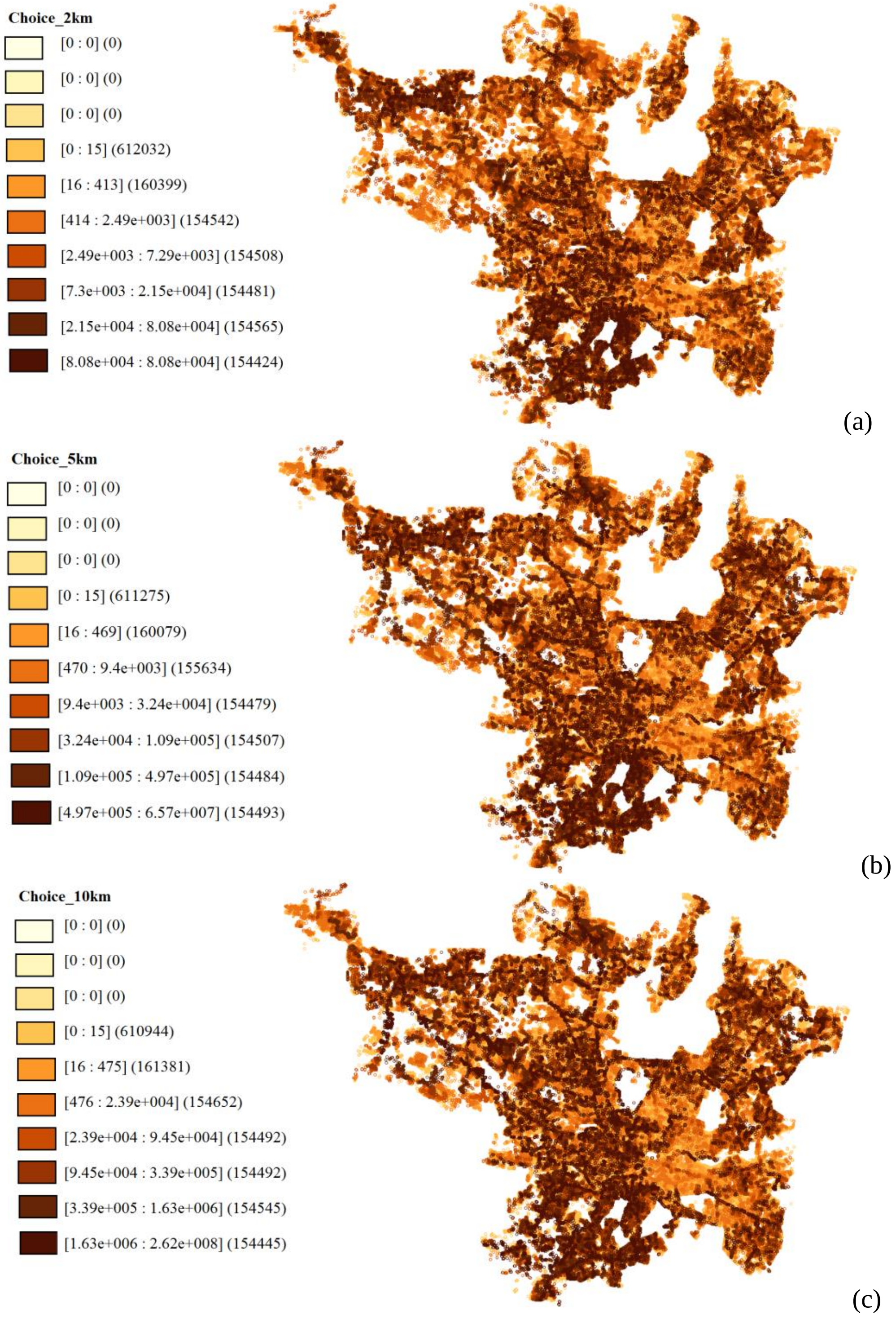

Choice_2km
[0 : 0] (0)
[0 : 0] (0)
[0 : 0] (0)
[0 : 15] (612032)
[16 : 413] (160399)
[414 : 2.49e+003] (154542)
[2.49e+003 : 7.29e+003] (154508)
[7.3e+003 : 2.15e+004] (154481)
[2.15e+004 : 8.08e+004] (154565)
[8.08e+004 : 8.08e+004] (154424)
(a)
Choice_5km
[0 : 0] (0)
[0 : 0] (0)
[0 : 0] (0)
[0 : 15] (611275)
[16 : 469] (160079)
[470 : 9.4e+003] (155634)
[9.4e+003 : 3.24e+004] (154479)
[3.24e+004 : 1.09e+005] (154507)
[1.09e+005 : 4.97e+005] (154484)
[4.97e+005 : 6.57e+007] (154493)
(b)
Choice_10km
[0 : 0] (0)
[0 : 0] (0)
[0 : 0] (0)
[0 : 15] (610944)
[16 : 475] (161381)
[476 : 2.39e+004] (154652)
[2.39e+004 : 9.45e+004] (154492)
[9.45e+004 : 3.39e+005] (154492)
[3.39e+005 : 1.63e+006] (154545)
[1.63e+006 : 2.62e+008] (154445)
(c)

Figure 3.7 Choice_2km (a), Choice_5km(b), Choice_10km (c)

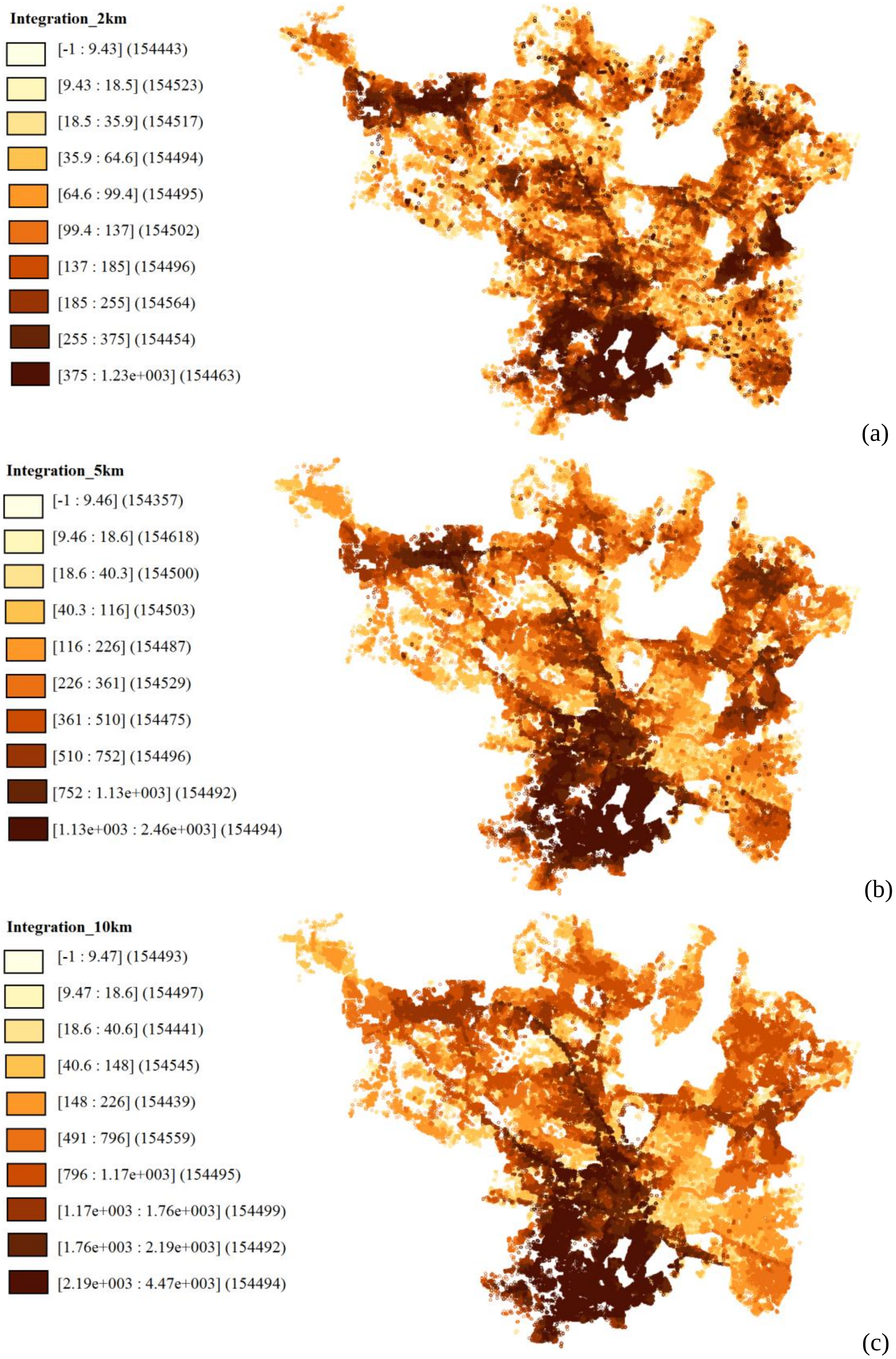


Figure 3.8 Integration _2km (a), Integration _5km(b), Integration _10km (c)

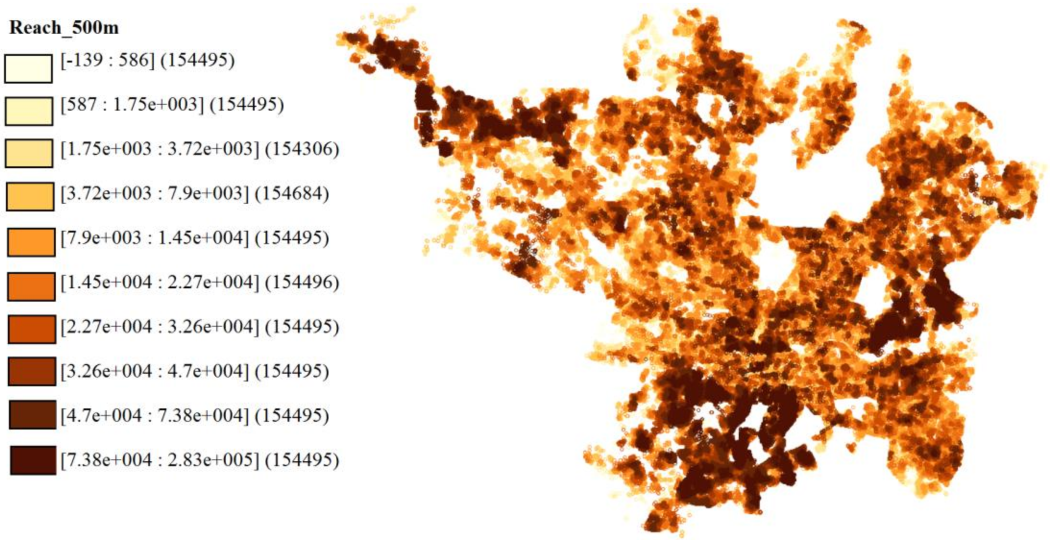

Reach_500m
[-139 : 586] (154495)
[587 : 1.75e+003] (154495)
[1.75e+003 : 3.72e+003] (154306)
[3.72e+003 : 7.9e+003] (154684)
[7.9e+003 : 1.45e+004] (154495)
[1.45e+004 : 2.27e+004] (154496)
[2.27e+004 : 3.26e+004] (154495)
[3.26e+004 : 4.7e+004] (154495)
[4.7e+004 : 7.38e+004] (154495)
[7.38e+004 : 2.83e+005] (154495)

Figure 3.9 Reach_500m

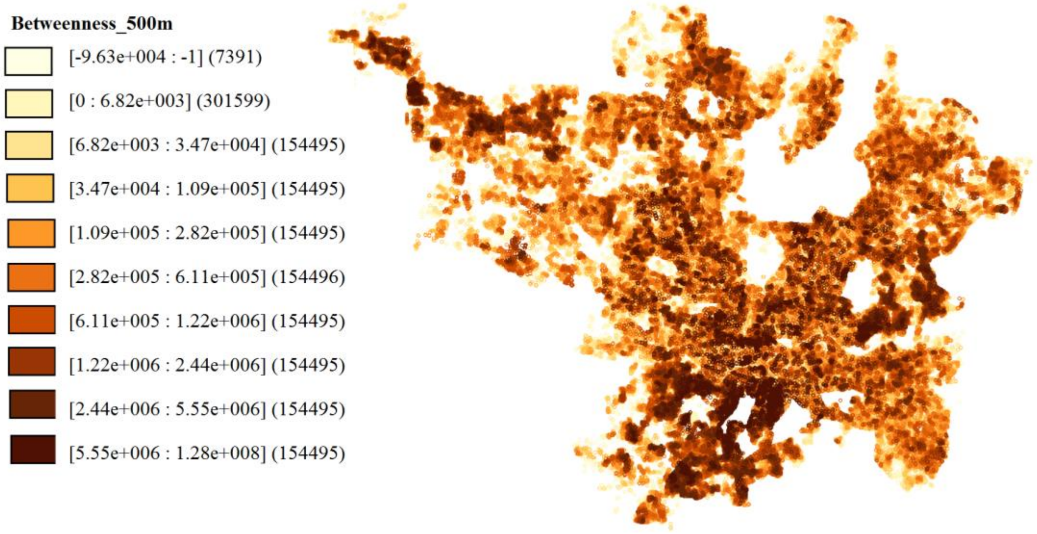

Betweenness_500m
[-9.63e+004 : -1] (7391)
[0 : 6.82e+003] (301599)
[6.82e+003 : 3.47e+004] (154495)
[3.47e+004 : 1.09e+005] (154495)
[1.09e+005 : 2.82e+005] (154495)
[2.82e+005 : 6.11e+005] (154496)
[6.11e+005 : 1.22e+006] (154495)
[1.22e+006 : 2.44e+006] (154495)
[2.44e+006 : 5.55e+006] (154495)
[5.55e+006 : 1.28e+008] (154495)

Figure 3.10 Betweenness_500m

The quantile distributions for all the configurational attributes are represented in figures above. The higher choice values, as can be seen in figure 3.7 (a), (b), and (c) appear to form a network core. As per space syntax theory, this network core is analogues to movement patterns in the city. Going from short search radius of 2km to large search radius of 10km, this urban core becomes more continuous in nature. The continuity corresponds to preferred connection links for all short and large trips made within the city. The distribution of integration which is subjective of better connections with neighborhood is radial in nature. The regions tend to be accessed with ease from neighborhood are present at the center and moving radially outward are regions with lesser connectivity with neighborhoods and have lesser integration values. Moving from short search radius towards larger radius, as seen in figure 3.8 (a), (b), and (c), the more integrated regions appear to shrink towards to a common global center. The reach and betweenness are calculated only for a single search radius of 500m, the high reach is more suggestive of urban cores having large building settlements or high compactness of urban environment as seen in figure 3.9. The high betweenness is suggestive of buildings traversed most in all short distance traversals in a system, betweenness resembles choice if building distributions throughout city is constant, but ideally this is not the case. As betweenness is also weighted by capacities of buildings, therefore the large buildings traversed most will have higher betweenness compared to smaller buildings traversed most. Therefore, the regions of high betweenness are suggestive of preferred routes for trips as well as density and compactness of the urban regions as reach. As seen in figure 3.10, the betweenness appear to form a continuous urban core as well as resemble the reach distribution.

The growth of roads and buildings in a city takes place in an opportunist fashion. The process can be deemed random; are these random processes bound to give rise to random environmental arrangements of urban infrastructure? Or the outward radial growth of city takes along with itself some repetitive patterns. If yes, what significance these patterns have on the dynamic activities taking place throughout the city. What is the semantic significance of the various combinations of configurations occurring in various parts of the urban environment and how can they be related to attractor values of the space? We try find answers to these questions through spatial clusters identification, outlier analysis, and knowledge based denomination of spatial units in terms of likelihood of them holding a high attractor values, by using Space Syntax and Urban Network Analysis findings in past.

3.10 Exploratory statistics for configurational parameters

Earlier section discussed conclusions drawn from visual inspections and interpretations of configuration. Move further, we would first inspect the distributions of Space syntax configurational attributes of integration and choice and UNA based attribute of reach and betweenness, and try to relate and justify what is conveyed in the fundamental semantic understandings of each of these.

3.10.1 Spatial variability in configuration with increasing search radius

The frequency distributions of choice and betweenness are heavily right skewed distributions as observed in figure 3.11 and figure 3.12. This justifies that in shortest distance traversals (angular distance based as well as metric distance based) between all pairs of origin destination pairs for a fixed radius relative to spatial unit under investigation, there are specific street segments that are repeatedly passed and have heavy movement attractor values as compared to other. Other than that, integration and reach are right skewed but not as heavily as choice and betweenness as seen in figure 3.13 and figure 3.14. The similarities in frequency raises 2 important questions, can integration serve as an informative proxy for reach and similarly can choice serve as a proxy for betweenness and vice versa?

Another important thing to note is that, choice frequencies for both axial analysis as well as in segment angular analysis get more and more skewed as the radius of analysis increases. This suggests, it is likely that the set of streets that are repeatedly passed by in short trip distances are also preferred in the long trip distances. The scatter plots of choice for 2km, 5km, and 10km for segment angular analysis partially confirms this as there are clear indications of linear dependency among these configurational attributes. Hence, along with question of integration and choice serving as informative proxies for reach and betweenness and vice versa, it also interesting to investigate whether the same set of streets have higher attractor value in terms of vehicular traffic for both short and long trip distances. A similar trend of linear dependency is observed in case of integration values for radii 2km, 5km, and 10km for segment angular analysis.

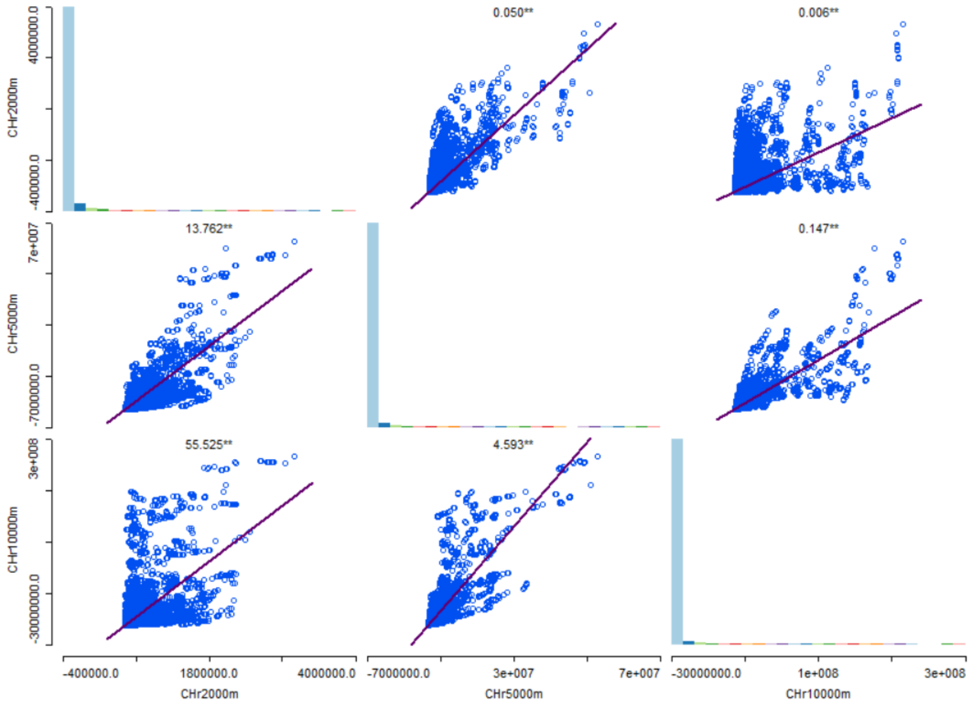


Figure 3.11 Frequency distributions for choice at metric radius of 2km, 5km and 10km

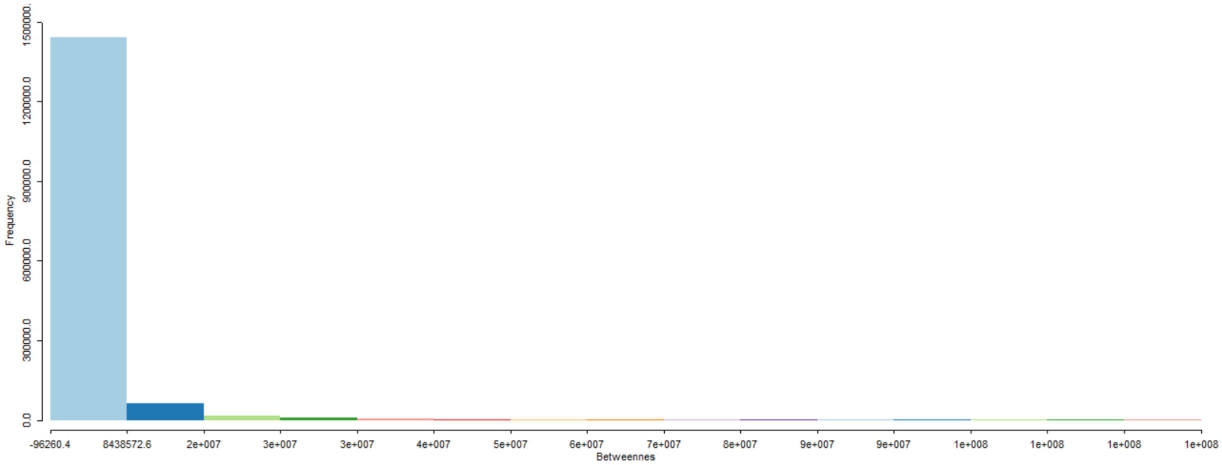


Figure 3.12 Frequency distributions for betweenness at metric radius of 500m

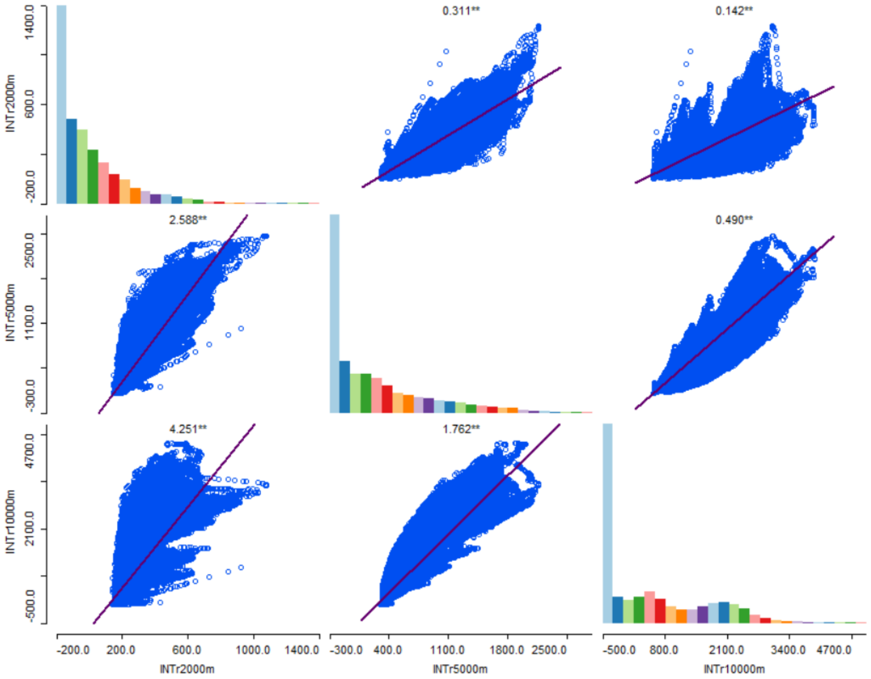


Figure 3.13 Frequency distributions for integration at metric radius of 2km, 5km and 10km

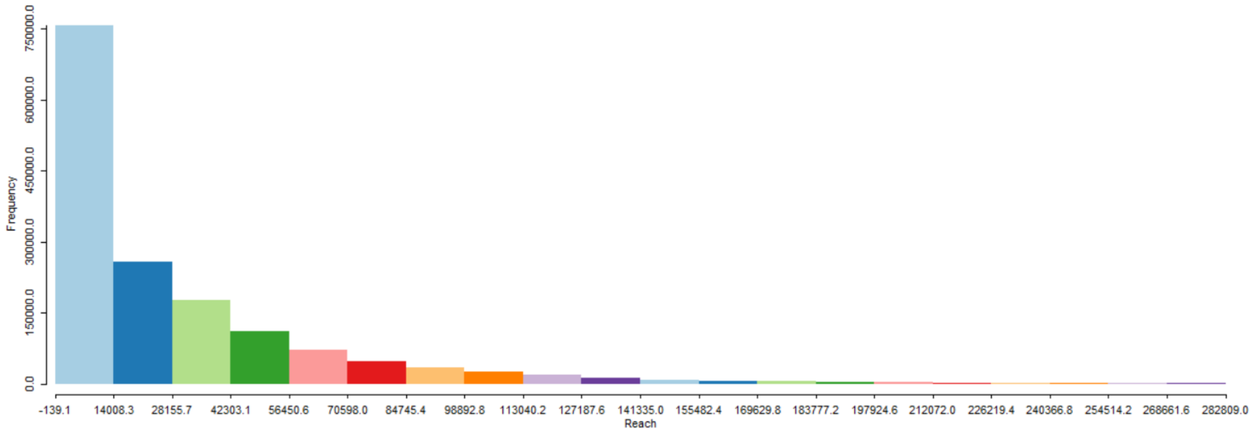
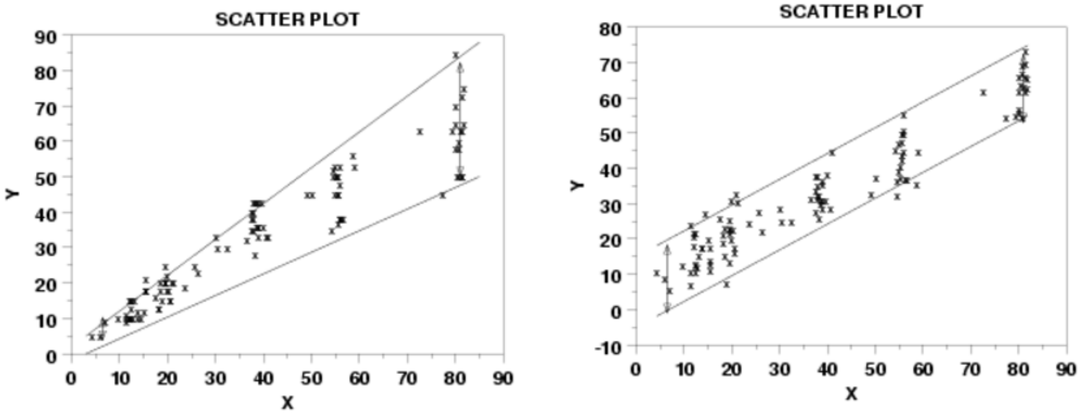


Figure 3.14 Frequency distributions for reach at metric radius of 500m

Though scatter plots for both choice and integration suggest of linear dependencies as we proceed from shorter search radius to larger search radius both for axial analysis and segment angular analysis but a significant difference to note in scatter plots is that the scatter plots for variability in choice values for different radius heteroscedastic in nature whereas for integration it is homoscedastic in nature.

The heteroscedastic scatter plot reveals an approximate linear relationship between X and Y, but more importantly, it reveals a statistical condition, that is, non-constant variation in Y over the values of X. For a heteroscedastic data set, the variation in Y differs depending on the value of X (figure 3.15 (a)).

The homoscedasticity scatter plot reveals a linear relationship between X and Y for a given value of X, the predicted value of Y will fall on a line. The plot further reveals that the variation in Y about the predicted value is about the same, regardless of the value of X (figure 3.15 (b)).



(a) (b)

Figure 3.15 Heteroscedastic scatter plot (a) and Homoscedasticity scatter plot (b)

3.11 Semantics of the syntactic configuration

3.11.1 Semantic meanings of relative values of configurational features

The values of configurational attributes suggest behavioral characteristics of any spatial unit. The spatial unit under consideration are buildings in this study as urban network analysis is performed explicitly for buildings whereas configurational values of streets from space syntax analysis are superimposed on buildings too. A similar value of any configurational attribute is suggestive of multiple behavioral characteristics. This is a top-down approach of understanding a space, where configurational attributes are calculated first and based on their values, the spatial units are denominated after visual inspection. However, this approach is rather constrained on individual spatial units and not on areas as whole. Therefore, we discuss the semantic meanings of various configurational attributes here, and further go on to build a bottom-up approach of area denominations by virtue of these semantic meanings.

3.11.1.1 Choice

- *High choice* – The Building falls on a street that is favoured in shortest angular/topological paths between pair of streets for given search radius.

- *Low Choice*
    1. The Building falls on a street that is not favoured in shortest angular/topological paths between pair of streets in given search radius.
    2. The search radius is larger than the urban area under consideration which can happen in case of irregular shape of landmass with natural boundaries such as forests, oceans etc.

3.11.1.2 Integration

- *High Integration* – The building falls on a street with high topological/angular connectedness with surrounding streets.
- *Low integration* -
    1. The building falls on a street with low topological/angular connectedness with surrounding streets.
    2. The search radius is larger than the urban area under consideration which can happen in case of irregular shape of landmass with natural boundaries such as forests, oceans etc.

3.11.1.3 Betweenness

- *High Betweenness* -
    1. The building has high volumetric capacity falls on a street that is favoured in shortest metric distance based paths between pair of buildings in given search radius.
    2. The building has low volumetric capacity falls on a street that is favoured in shortest metric distance based paths between pair of buildings in given search radius.
    3. The building has very high volumetric capacity falls on a street that is not favoured in shortest metric distance based paths between pair of buildings in given search radius.

- *Low Betweenness* -
    1. The volumetric capacity is building is very low compared other buildings in given search radius and it still falls on a street that is favoured in shortest metric distance based paths between pair of buildings in given search radius.
    2. The volumetric capacity is building is very low compared other buildings in given search radius and it doesn’t fall on a street that is favoured in shortest metric distance based paths between pair of buildings in given search radius.
    3. The volumetric capacity is building is low compared other buildings in given search radius and it doesn’t fall on a street that is favoured in shortest metric distance based paths between pair of buildings in given search radius.

3.11.1.4 Reach

- *High Reach* -
    1. The building is surrounded by several high-rise buildings in given search radius.
    2. The building is densely surrounded by buildings in given search radius. The buildings may be large or small. This can also be inferred by closeness measure of UNA.

- *Low reach* -
  1. There are very few buildings in close proximity of the building under consideration.
  2. The search radius is larger than the area under consideration, this may happen due to occurrence of natural boundaries such as forests and ocean.

3.11.2 Semantic similarities of configurational features

We have earlier raised questions about pair of reach and integration and choice and betweenness serving as informative proxies for each other. To analyze same, the configurational attributes must be compared at similar search radius. Since the urban network analysis is conducted at 500m search radius, therefore we choose integration and choice values 2 km metric search radius.

For the analysis, we have chosen two wards from Hyderabad, one being regularly gridded (ward 43) with equally spaced buildings and other having non-gridded tree like road networks with buildings patchily arranged (ward 67).

We use Pearson correlation to understand this bivariate statistic. Pearson correlation coefficient [48] is a measure of the linear correlation between two variables *X* and *Y*. It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. Values close to 1 suggest strong lines relationship between the variables of interest whereas values close to 0 suggest weak linear relationships. The Pearson correlation coefficient is given by:

$$r = \frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2]\,[n\sum y^2-(\sum y)^2]}} \qquad (3.1)$$

3.11.2.1 Semantic similarity of integration and reach

Urban Environment	Pearson Correlation Coefficient
IRREGULAR ARRANGEMENT (Hyderabad ward 67) 	Integration (2 km), Reach (500m) r = 0.59
REGULAR ARRANGEMENT (Hyderabad ward 43) 	Integration (2 km), Reach (500m) r = 0.67

Table 3.1 Correlation between integration and reach

The integration values from segment angular analysis with metric search radius of 2km has strong relationship with reach values calculated at 500m search radius. The coefficients for linear dependency between integration for segment angular analysis at 2km metric radius and reach for 500m metric radius are 0.59 and 0.67 (table 3.1) respectively for irregular arrangement of building and roads against regular arrangement of buildings and roads.

This proves to an extent that integration from segment angular analysis with metric radius and reach can substitute each other if the geometry of roads and buildings is not heavily distorted and search radius is nearly the same for both analyses.

3.11.2.2 Semantic similarity of choice and betweenness

Urban Environment	**Pearson Correlation Coefficient**
IRREGULAR ARRANGEMENT (Hyderabad ward 67) 	Choice (2km), Betweenness (500m) r = 0.29
REGULAR ARRANGEMENT (Hyderabad ward 43) 	Choice (2km), Betweenness (500m) r = 0.61

Table 3.2 Correlation between choice and betweenness

The calculation of betweenness as well as choice measures depend on the number of trips which eventually depend on the number of spatial units in the area.

The spatial units considered for betweenness are buildings whereas for choice street elements are considered. Since, for calculation of betweenness we also weigh the respective buildings by their volumetric capacities, it is likely that high-rise buildings on a street that is not often passed in origin-destination journeys as compared to low rise buildings on a street or no buildings at all will have higher betweenness values. Cities are such, we do not often find buildings with same volumetric capacities all over the space. However, since the choice is solely dependent on networks, this will always give consistent results for similar grid-arrangements. The only cases of betweenness and choice serving as substitutes of each other is for strict regular grid arrangement where buildings too have similar volumetric capacities over entire space which is usually not the case in all urban environments. This is evident from Pearson correlation coefficient value for linear dependency between the 2 attributes for ward 43 at comparable short distance radius of 500m (urban network analysis) and 2km (segment angular analysis) (table 3.2) where correlation coefficient is higher compared to that for ward 67.

3.12 Semantic rule based approach for identifying high attractor value zones

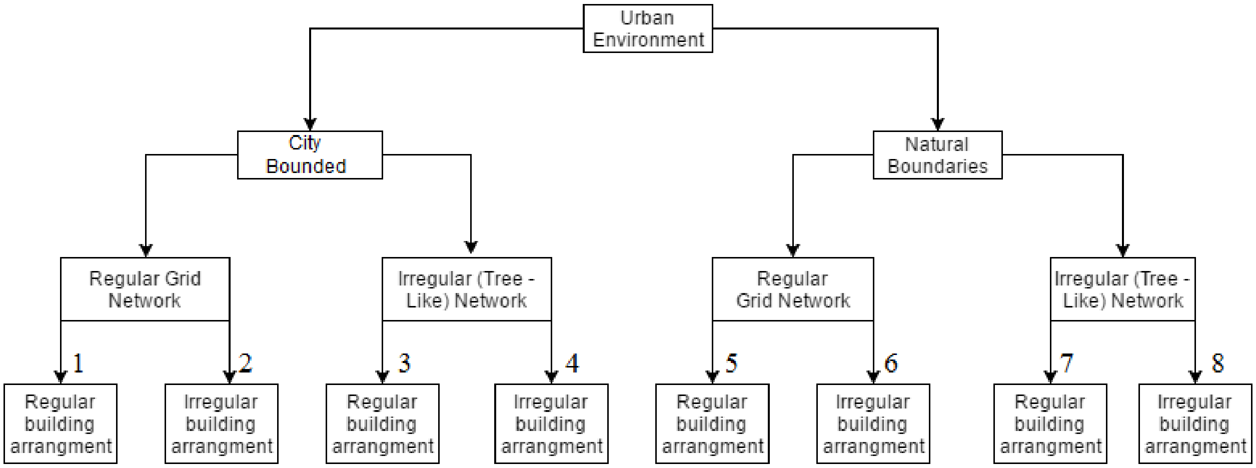

Urban Environment
City Bounded
Natural Boundaries
Regular Grid Network
Irregular (Tree - Like) Network
Regular Grid Network
Irregular (Tree - Like) Network
1
2
3
4
5
6
7
8
Regular building arrangment
Irregular building arrangment
Regular building arrangment
Irregular building arrangment
Regular building arrangment
Irregular building arrangment
Regular building arrangment
Irregular building arrangment

Figure 3.16 Possible conformations in an urban environment

Based on the semantic meanings of configurational attributes, this section proposes a bottom-up approach i.e., identifying nature of space first and then using values of configurational attributes for knowledge based denomination of attractor values, hence movement patterns in space. The various possible conformations of urban configuration are represented in figure 3.16. The network of roads in any city can be divided into two categories, either tree like network which are a result of opportunist development in city which are not procedurally planned and others are regular grids where urban core is planned first and then developed. Parts of city may be naturally bounded by oceans, forests etc. This causes naturally boundary problem; the configurational attributes suffer heavy inconsistency due to the same. Therefore, the conformations of city bound regions and those bounded by natural entities are analyzed separately. Further, regular and irregular arrangements of buildings in each of the above discusses configurational setting gives rise to new conformations. Regular arrangement of buildings corresponds to equal sized buildings arranged equidistant from each other. Earlier, Section 3.11.1 discussed relative values of configuration parameters with respect to individual space entities viz. building. However, attractor values and high movement patterns can only be discussed over areas as whole, this is where relative importance of neighborhood must be considered. We present a few set of rules that can be directly applied to conformations of urban space mentioned above.

RULE 1: The conformations 1 will have nearly the same attractor values for entire area under considerations. There won't be any significant differences between the configuration attributes, they may have identical values for all spatial units. The attractor values of spaces under such arrangements is likely to be deviated by non-spatial attributes such as aesthetics and economic value of space.

RULE 2: Like conformation 1, the network configuration in conformation 2 suggests similar attractor value of entire space under consideration. However, as the buildings are not symmetrically distributed, therefore the tendency of local movements is highly governed by regions with high building density, i.e. reach and street preference driven by betweenness and choice at short metric radius.

RULE 3: The general tendency of movement is: pedestrians take shortest possible paths, motorists take most convenient paths. In conformations 3 and 4, there may be multiple paths with varying angular and metric distances for given origin destination pair, but, a motorist would often take a path that is straight enough to avoid cognitive load if there is no significant difference in length from the shortest possible metric distance based path with multiple number of turns. The global movement in conformations 3 and 4 is driven by choice and integration for high search radius irrespective of considering the buildings in the region while the local movements are function of high reach, high choice and high integration for short metric radii in conformation 3 whereas for conformation 4 the movement affinity is higher for regions with high reach, high betweenness and high choice for short metric radius as in conformation 2.

RULE 4: It is suggested that for conformations 5, 6, 7, and 8, the search radius must be limited to their extents, such areas may or may not suffer natural edge effects due to natural boundaries viz. ocean, forests. It is also suggested that if such areas are analyzed at high search radius to capture global scale of movements, they must not at all be compared with landlocked regions as a relatively important regions might be heavily down-weighted against a relatively less important region in city locked area, this is due to the fact that the relative coverage for a city-locked area is significantly higher for high search radius which may cause its spatial units to have a high integration, reach, choice, and betweenness values against its naturally bounded counterpart. If such areas are analyzed at high metric search radius, choice remains the single largest attractor value identifier. All local movement patterns follow the configurational behavior as specified for conformations 1,2,3, and 4 for 5,6,7, and 8 respectively.

3.13 Realization of rule based approach

3.13.1 Spatial Autocorrelation for capturing global behaviors

The task of grouping a set of artifacts in such a way that similar objects are placed in a common group is called as clustering and formed groups are called as clusters. The objects in one group differ in properties from the objects in any other group. In exploratory statistics, clustering is a one of the important tasks that give evidences to reach a conclusion or aid it. Is it a good idea to understand the spatial data using traditional clustering approaches? It must be understood that we are trying to identify repetitive patterns of patches in urban environment based on configurational parameters. This suggests that other than the configuration attributes, location of entities plays a big part in process. This is where neighborhoods come in picture. Clustering might often furnish very good results pertaining to similarity of individual spatial units of roads and buildings but the significance of similarity reduces if these similar data points are farther apart geometrically. The understanding thus narrows down to a point where we need to understand spatial variability in context of neighborhoods instead of individual entities at this juncture. We will explore other Clustering methods in the further chapters when we talk of configurational variability in context of vehicular traffic variations. There may be dependency between spatial observations made at two different locations. For example, observations made for locations in closer vicinity may also be closer in value compared to that observed for two locations farther apart. This phenomenon is called spatial autocorrelation. Spatial autocorrelation measures the correlation of a variable with itself through space [33]. Spatial autocorrelation can be positive or negative. When similar values occur closer to one another, the spatial auto correlation is positive whereas when dissimilar values occur near one another, spatial autocorrelation is negative.

3.13.1.1 Weight Matrix for Spatial Correlation

A threshold distance or the number of neighbors needs to be defined first for choosing neighbors to assess spatial autocorrelation [33]. These threshold values are often presented as weight matrix, the relationships between locations where observations are made are defined by this weight matrix. If data are collected at 'n' locations, then the weight matrix will be $n \times n$ with zeroes on the diagonal.

The weight matrix can be specified in many ways:

- A fixed weight for all observations within a specified distance.
- A constant weight for any two different locations.
- Weight is proportional to inverse distance, inverse distance squared, or inverse distance up to a specified distance.
- A constant weight for K nearest neighbours, and all others are zero.

The weight matrix is row-standardized, i.e., all the weights in a row sum to one. We have chosen K nearest neighbor method with value of K being 20.

3.13.1.2 Measure of Spatial Autocorrelation – Moran's I

Moran's I [33] tests for global spatial autocorrelation for continuous data. It is based on cross-products of the deviations from the mean and is calculated for observations on a variable at locations I and j, as:

$$I = \frac{n}{S_0} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \tag{3.2}$$

Where,

$\bar{x}$ = mean of the variable
w_{ij} = elements of the weight matrix,
S_0 = sum of the elements of the weight matrix: $S_0 = \sum_i \sum_j w_{ij}$

3.13.1.3 Measure of Spatial correlation – Bivariate Moran's I and BiLISA

Bivariate Local Indicators of Spatial Autocorrelation (LISA) [33] is a correlation between two different variables in an area and in nearby areas. The global bivariate Moran's I statistic calculates spatial interdependency of two variables x_k and x_l in a same location. The equation for autocorrelation is given as follows:

$$I_{kl} = z_k w z_l / n \tag{3.3}$$

Where,

n = Number of observations

w = Row-standardized spatial weight matrix.

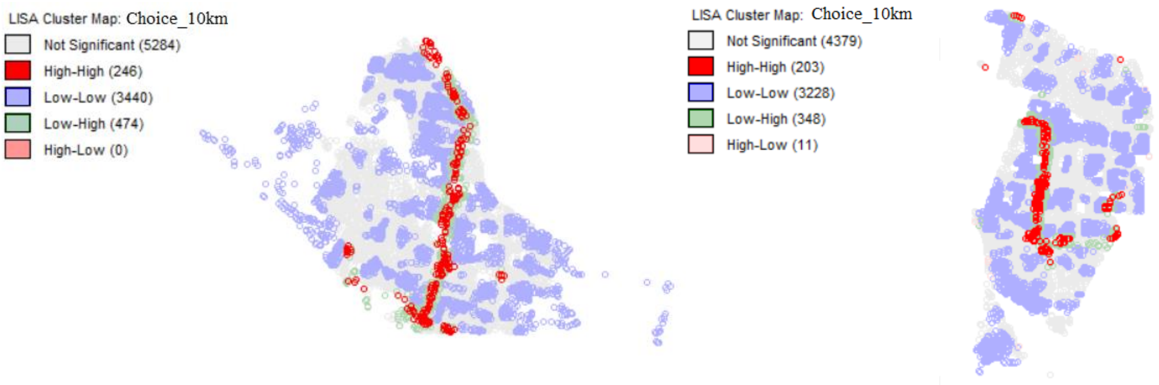
$z_k = [x_k - \overline{x_k}\,]/\sigma_k$ And $z_l = [x_l - \overline{x_l}]/\sigma_l$ have been standardized such that the mean is zero and standard deviation equals one. The weight matrix defines the neighbor set for each observation with non-zero elements for neighbor and zero for the others. The global Moran's I fail to give any concrete information for the existence of clusters i.e., occurrence of localized groups showing similar characteristic properties. Local Indicators of Spatial Association (LISA) helps to identify the type of spatial correlation and provides a measure of association for each spatial unit. The bivariate LISA can be defined as follows [34]:

$$I^i_{kl} = z^i_k \sum \widehat{w}_{ij} z^j_l \tag{3.4}$$

Where,

I^i_{kl} = Degree of linear association (positive or negative) between the values for variable x_k at a given location 'i' and the average of variable x_l at neighboring location such as 'j's (spatial lag).

3.13.1.4 Results and Discussions – Univariate Moran's I



(a) (b)

Figure 3.17 Univariate LISA for choice_10km on ward 43 (a) and ward 127 (b), Hyderabad

As per the semantic rules, if boundary areas are studied for global movement patterns, choice is the single largest estimator. The continuity of high choice regions is captured for wards 43 and 127 using univariate LISA for choice_10km (figure 3.17 (a) and (b)). The continuous lines represented by High-High correspond to significant regions that have high choice values and are also surrounded by neighbors having high choice values. All regions which do not have this continuity are not part of High-High clusters and thus are not the usual choice for movement.

3.13.1.5 Results and Discussions – Bivariate Moran's I

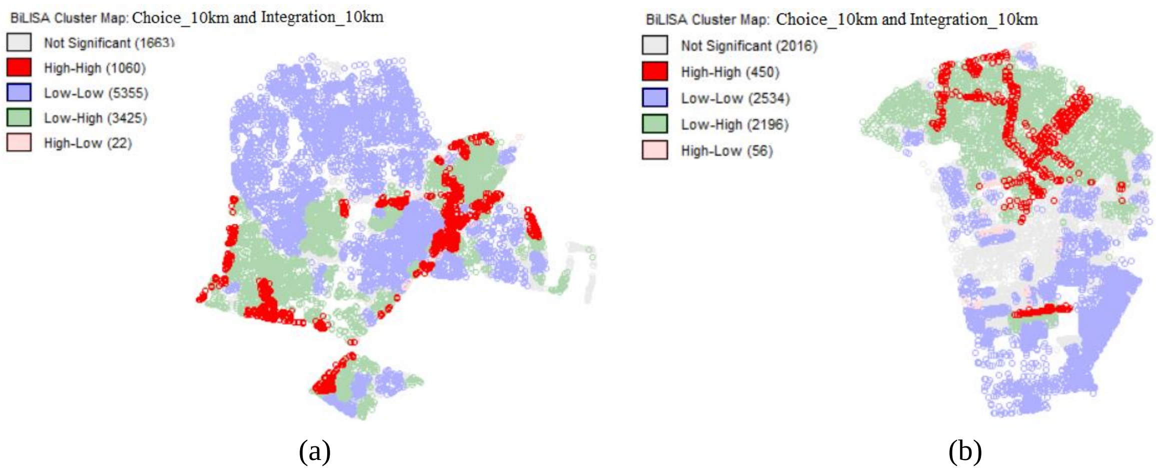


Figure 3.18 Bivariate LISA for choice_10km and integration_10km on ward 67 (a) and ward 82 (b), Hyderabad

As per the semantic rules, if the region is bounded from all sides by the city elements, and network is unsymmetrical with either consistent or inconsistent distribution of buildings, the global movements are driven by choice and integration both. For wards 67 and 82, High-High regions for bivariate LISA on choice_10km and integration_10km (figure 3.18(a) and (b)) are the regions of high occupancy for all global movements through the region.

3.13.2 Outlier Analysis to capture local behaviors

An outlier is generally considered to be a data point that is far outside the norm for a variable or population [35]. Hawkins [36] described an outlier as an observation that "deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism". Are outliers always that bad statistically?

It is a possibility that an outlier can come from the population being sampled legitimately through random chance. It is important to note that sample size plays a role in the probability of outlying values. Within a normally distributed population, it is more probable that a given data point will be drawn from the most densely concentrated area of the distribution, rather than one of the tails [37]. As the size of captured data increases, it is more likely for samples to resemble the population from which it was drawn, and thus the likelihood of outlying values becomes greater.

It is often confusing on how to deal with outliers when they occur as function of variability of data. What if data is not actually distributes normally and it is rather skewed with significantly heavier tails? Well, this is the case with spatial data-sets. What significance do these outliers have in context of spatial information? A few practical examples to mention are those of a very small fraction of city roads being heavily occupied, a very small fraction of city buildings being sky-scrappers, and a very small fraction of trips made across the city being very long distance trips and a very few fraction of places in city being retail outlets. The examples we quoted above are a mix of spatial as well as non-spatial characteristics of the city. These characteristics have been related to configurational parameters in past. The variability and patterns observed in configurational parameters give us a chance to identify occurrence of repetitive patterns in the city and establish the common trends of distributional statistics in the city.

3.13.2.1 Results and discussions – Outlier Analysis

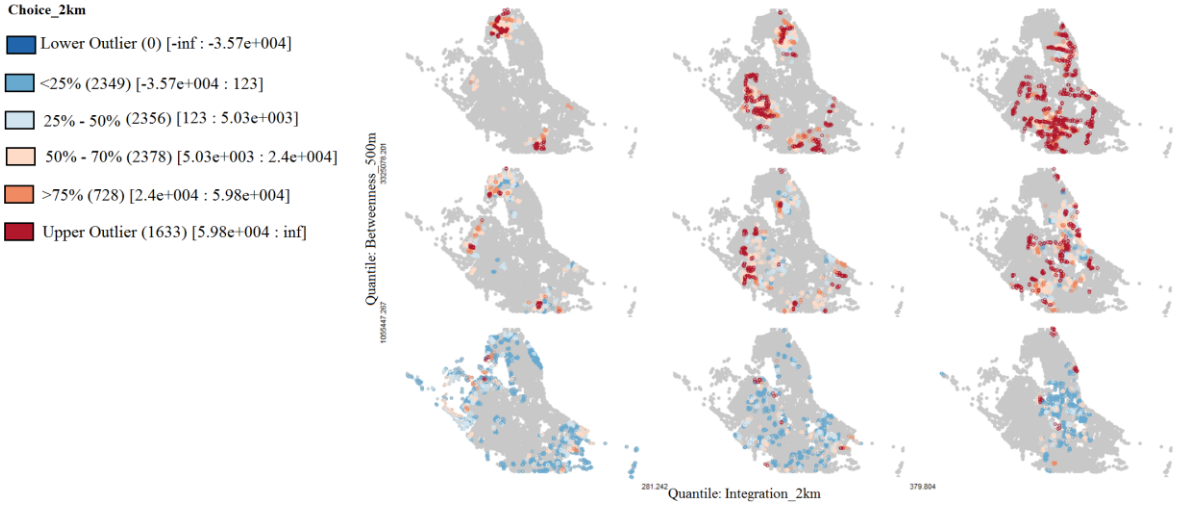

Choice_2km
Lower Outlier (0) [-inf : -3.57e+004]
<25% (2349) [-3.57e+004 : 123]
25% - 50% (2356) [123 : 5.03e+003]
50% - 70% (2378) [5.03e+003 : 2.4e+004]
>75% (728) [2.4e+004 : 5.98e+004]
Upper Outlier (1633) [5.98e+004 : inf]
Quantile: Betweenness_500m
3325078.201
1055447.267
281.242
Quantile: Integration_2km
379.804

Figure 3.19 Outlier analysis on choice under conditional setting with high integration and high betweenness – ward 43, Hyderabad

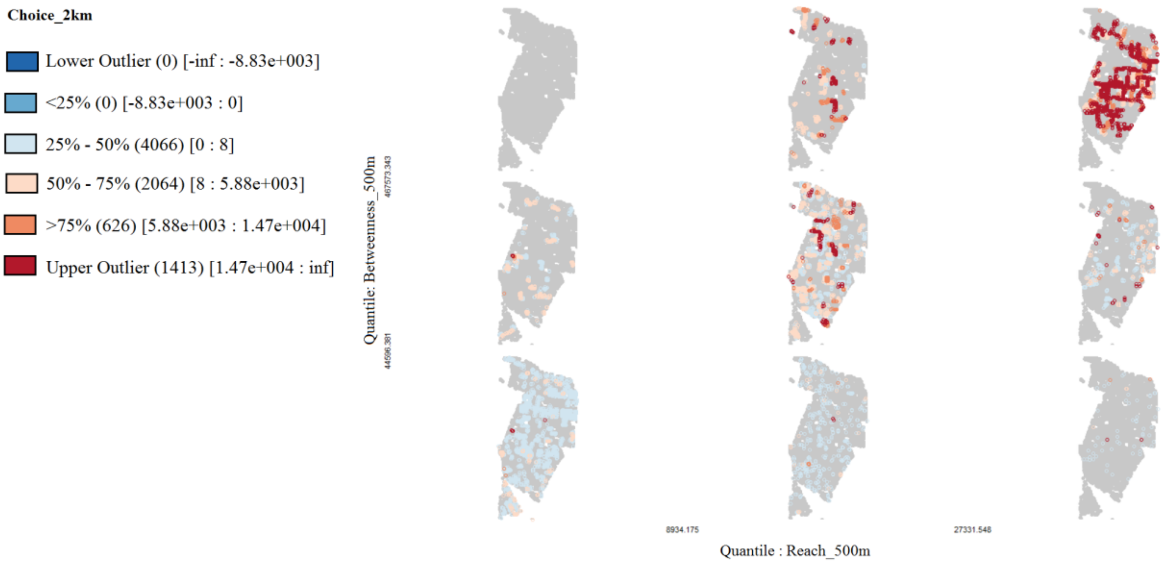

Choice_2km
Lower Outlier (0) [-inf : -8.83e+003]
<25% (0) [-8.83e+003 : 0]
25% - 50% (4066) [0 : 8]
50% - 75% (2064) [8 : 5.88e+003]
>75% (626) [5.88e+003 : 1.47e+004]
Upper Outlier (1413) [1.47e+004 : inf]
Quantile: Betweenness_500m
467573.343
44596.381
8934.175
27331.548
Quantile : Reach_500m

Figure 3.20 Outlier analysis on choice under conditional setting with high reach and high betweenness – ward 127, Hyderabad

For capturing the local movement patterns in regions, semantic rules have stated that for a tree like symmetry of networks, integration, betweenness and choice are major determinants whereas for a grid like network pattern, reach, betweenness and choice are defining parameters. In this case study, theme of map is chosen as choice_2km as it is largest estimator of movements, it is classified as an outlier when it lies more than a given multiple of the interquartile range (the difference in value between the 75% and 25% observation) above or below respectively the value for the 75th percentile and 25th percentile. The standard multiple used is 1.5 times the interquartile range. The upper outlier regions with high integration_2km and reach_500m for ward 43 (figure 3.19) are major attractor zones for local movements, and similarly the upper outlier regions with high reach_500m and reach_500m for ward 127 (figure 3.20) are major attractor zones for local movements.

3.14 Summary

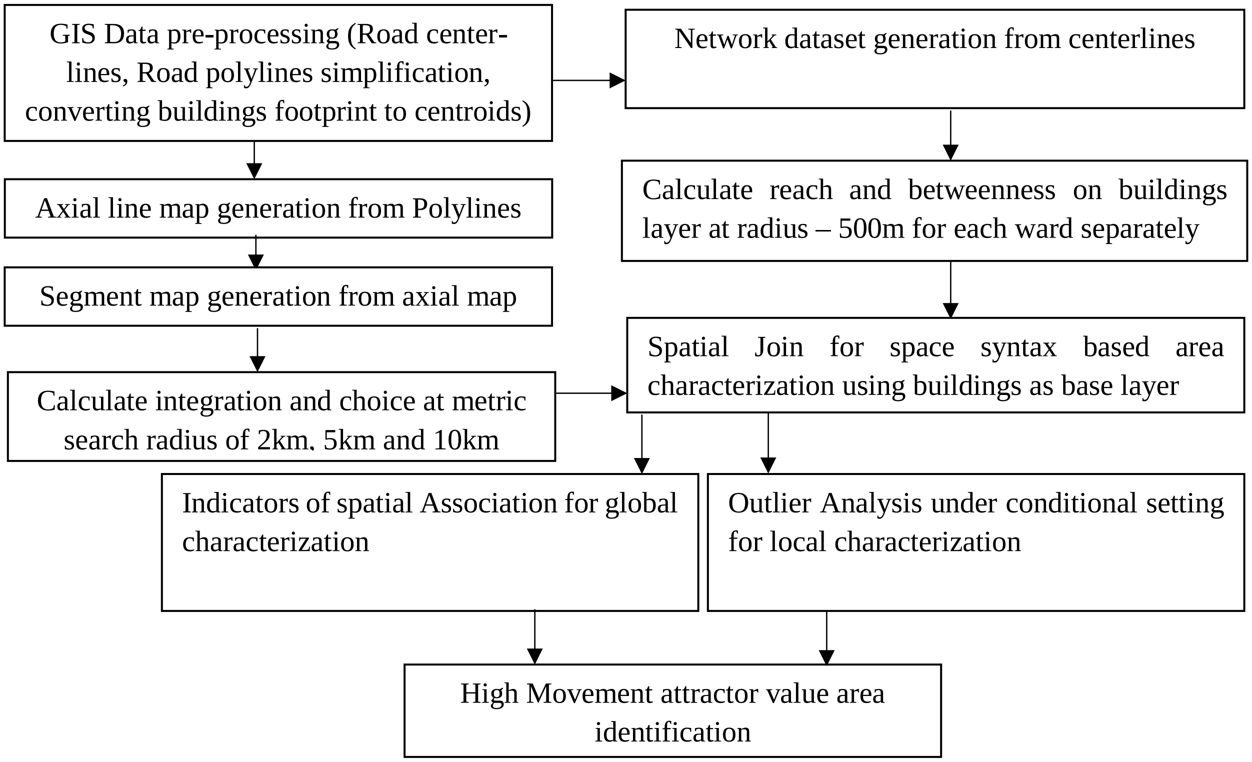

GIS Data pre-processing (Road center-lines, Road polylines simplification, converting buildings footprint to centroids)
Network dataset generation from centerlines
Axial line map generation from Polylines
Calculate reach and betweenness on buildings layer at radius – 500m for each ward separately
Segment map generation from axial map
Spatial Join for space syntax based area characterization using buildings as base layer
Calculate integration and choice at metric search radius of 2km, 5km and 10km
Indicators of spatial Association for global characterization
Outlier Analysis under conditional setting for local characterization
High Movement attractor value area identification

Figure 3.21 Framework flow diagram

Figure 3.21 shows the flow diagram for proposed framework for identification of high movement attractor value areas in city. The limitations of various urban analysis models have been overcome by using a hybrid approach. Further the indicators of spatial estimation showed that areas with similar configuration occur in form of spatial cluster that define the behavioral characteristics of the region. Using this property reflected by urban areas, the high attractor value zones are identified both at local and global scales.

Chapter 4

Proposing a spatial weight metrics for road networks and estimation of urban movement patterns – Case Study of New York city

Previous chapters discussed inconsistencies in various spatial configurational analysis in context of understanding attractor values of spaces and proposed a hybrid framework with buildings as the common layer of analysis to overcome the discussed inconsistencies. However, to bring everything down to common layer of buildings, Axial Analysis, Segment Angular Analysis, and Urban Network Analysis had to be computed separately. But, considering buildings as base layer, the framework approach discarded important roads which did not have any buildings adjacent to them. This is a major limitation and needs to addressed. As suggested in [38], segment angular analysis captures the attractor values better, but we strongly recommend that buildings information must not be discarded, therefore, we propose a spatial weight metrics on urban road network which is representative of buildings, and use this weight metrics to perform weighted segment angular analysis. Further we use clustering and classification approaches to capture movement attractor values specifically for city of New York.

Drawing an axial map and then converting it to a segment map is a hectic process, hence researchers have used road centerlines instead for segments angular analysis. Turner [25] has shown how road-Centre line maps and space syntax axial line maps may be analyzed in a comparable fashion by using angular segment analysis (ASA). Going by this suggestion, we too have used road centerlines for segment angular analysis in this case study.

4.1 Spatial Weight Metrics of Buildings over Road Network

Section 3.7 discussed concept of spatial join and in section 3.7.1, the concept is used for fusing network features on buildings. A similar operation is performed for weight metrics calculation. A building can be strictly adjacent to a road but unlike this, a road can be adjacent to multiple buildings as shown in figure 4.1(c). Therefore, the weight for a road element is arithmetic sum of floor areas of the buildings adjacent to it. As mentioned earlier, important roads such as highways and flyovers may not have any buildings adjacent to them, a Zero weight would cause them to be discarded from calculation. Hence, we give a unit weight to all such road elements. Apart from the buildings weight, we also maintain the information of the number of buildings adjacent to each road element.

$$k = \Sigma_{\{j:d(b_j,r_i)\leq d(b_j,r_l),\ l\neq i\}} 1 \tag{4.1}$$

Where,
k = number of buildings that have a l^{th} line closest to them.
b_j = j^{th} building
r_i = i^{th} road segment
r_l = l^{th} road segment
$i,l \in \{1, M\}$, M = number of segment in the city
$j \in \{1, N\}$, N = number of buildings in the city

And,

$$W_i = \sum_{n=1}^{k} (B_n) \tag{4.2}$$

If, k = 0, Wi = 1

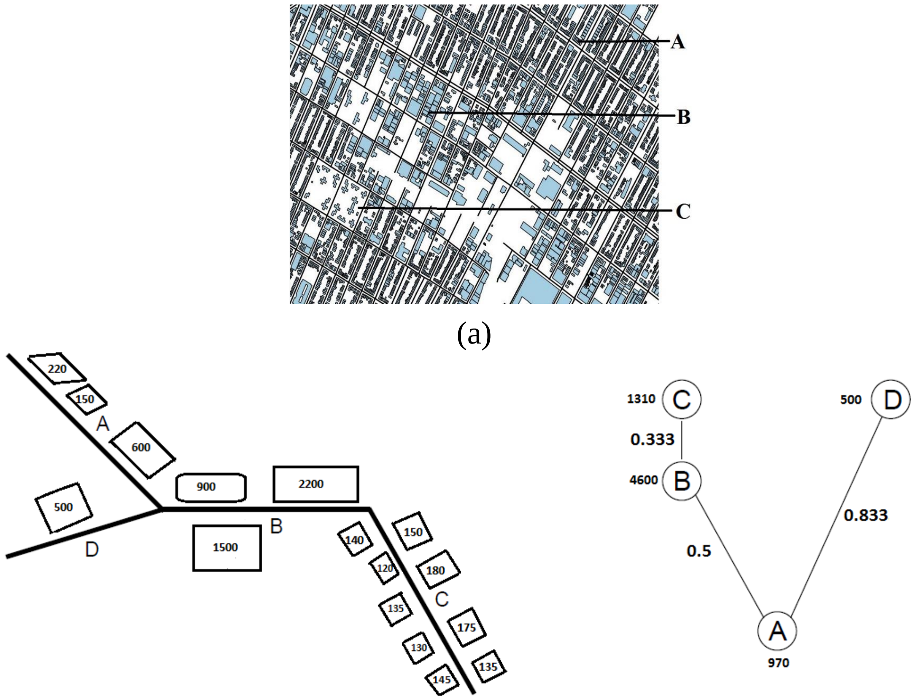
Where,
W_i = Weight of Road segment i.
B_n = Floor area of building n
$i \in \{1, M\}$, M = number of segment in the city

4.2 Segment Angular Analysis with Weight Metrics

Section 2.3.1 discussed mathematical understanding of segment angular analysis configurational parameters with, addition of building weights on network layers helps in factoring building information without considering building to building trips and saves the additional computation overhead. We use segment angular analysis specifically because previous studies suggested it to be a better estimator that actual metric distance based navigation [38]. Following section discusses the modifications in the configuration attributes when weighted calculation are performed for the network.

4.2.1 Mathematical understanding of segment angular analysis parameters with weight metrics

We have earlier discussed urban patch shown in figure 4.1 (a), section A has smaller buildings arranged regularly, section B has high rise buildings with large volumetric capacities and section C has small buildings in patches. Road Network alone can never depict the attractor values inherited in buildings. The buildings in a region may vary highly. Few of them might be large while other may be small. Few might form dense urban core while others may be patchily scattered over space. In context of limitations of various configurational theories, few have been addressed using the hybrid approach of superimposing network configurational features over buildings but limitation of urban network analysis still remains unattended. The building weight metric over network as a solution to this is analyzed in the sections that follow.



(a)

(a)

(c)

Figure 4.1 Arrangements of Buildings in a grid (New York city) (a), Paths through a network, buildings adjacent to them (b), and associated graph with angular distances (c)

In context of such variations in building arrangement in tree – grid networks, following section discusses calculations for segment angular analysis under weighted network condition:

The Node count is the number of segments encountered on the route from the current segment to all others [25]. In the case of figure 4.1 (b), the node count (NC) is 3 which is replaced weight of the streets falling on shortest paths. This weight count is denoted by WC. From A to C goes through three segments with weights 6410 in total. Angular depth between adjacent segments for corresponding section of paths is displayed in figure 4.3 (c). Total angular depth is the cumulative total of the shortest angular paths to all segments multiplied by the weight of the destination segment. In the case of segment ‘A’ in figure, the angular total depth is:

$$\text{TD `A'} = (4600)0.5+(1310)0.833+(500)0.833= 3423.63 \qquad (4.3)$$

The angular mean depth value for a line is the sum of the shortest angular paths divided by the sum of all angular intersections in the system rather than the number of lines in the system. Mean depth in general terms is indicative to how deep or shallow a node is in relation to the rest of the graph, a measure defined as centrality. In figure, angular mean depth for the segment ‘A’ is:

$$\text{MD `A'}= \text{TD `A'}/\text{WC} = (4600)0.5+(1310)0.833+(500)0.833/6410 = 0.534 \qquad (4.4)$$

In angular segment analysis, integration is predictor of potentials for each segment to be a highly desired destination within defined boundaries i.e. given search radius. The measure forecasts to-movement possibilities for each segment while measuring the shortest angular patch in the defined system between all origin-destination pairs.

Integration for angular segment analysis is:

$$\text{Integration} = \text{WC/MD} = 6410/0.534 = 12003.74 \qquad (4.5)$$

Weighted choice is given by:

$$C^l(x) = \sum_{i=1}^{n} \quad \sum_{j=1}^{n} \sigma^l(i, x, j) \text{such that } i \neq j \qquad (4.6)$$

Where,
$C^l(x)$ is weighted choice
The weighted sigma function σ^l used is defined as: if the shortest path from *i* to *j* passes through *x*, it is simply w(i)*w(j) (weight on segment *i* times weight on segment *j*); if *x* is the origin *i* then σ^l is *w(x)*w(j)/2* and if *x* is the destination *j*, it is *w(i)*(x)/2*; otherwise, if *x* is not on the shortest path between *i* and *j*, nor the origin or destination of the shortest path from *i* or *j*, σ^l is 0.

4.3 New York City GIS Data

The datasets of Road Centerlines Polyline vector files and Buildings Polygon vector files for New York City are obtained from Open Street Maps (OSM) extracts [39]. The Annual Average Daily Traffic (AADT) counts for year 2014 and 2015 polyline vector files are obtained from New York state official data catalogue [40]. To mitigate impact of edge effects on network analyses, road centerline of few areas of New Jersey in neighborhood (10km) of New York city are also added to dataset [41], these areas include Newark, Jersey City, Elizabeth, Linden, Carteret, Perth Amboy and Woodbridge Township. All data files gathered for this study are also in shape (.shp) format as that for Hyderabad geospatial data.

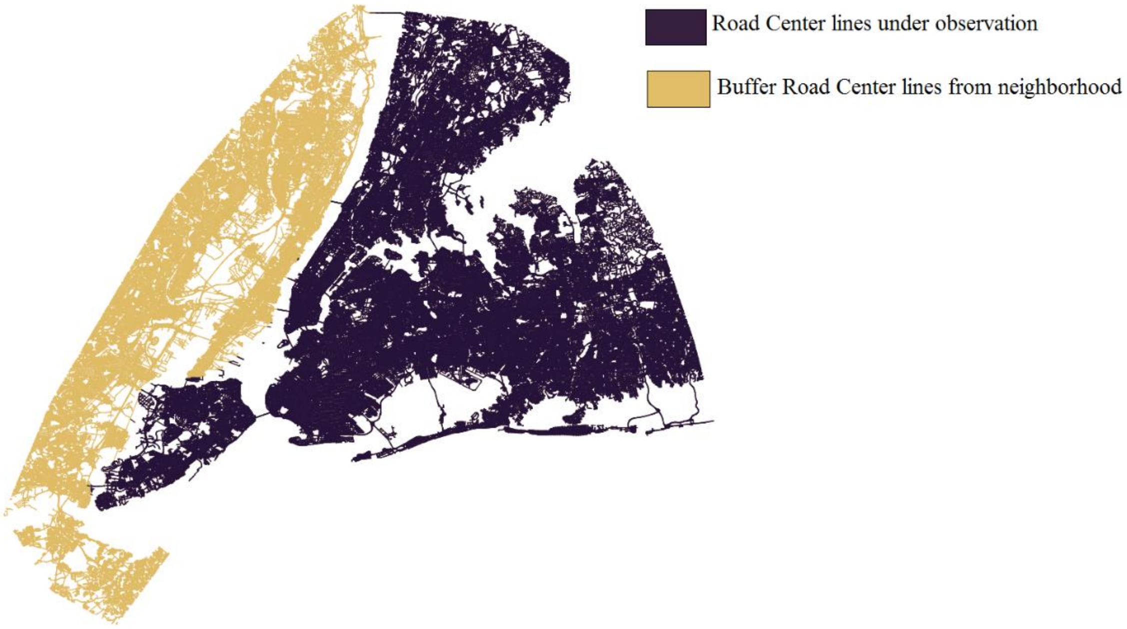

Road Center lines under observation
Buffer Road Center lines from neighborhood

Figure 4.2 Road center line Networks of New York and 10km buffer around them

4.4 Computation of weighted and unweighted segment angular analysis configurational attributes

The flow for calculation of configurational parameters for city of New York is almost same as that for Hyderabad case study. To avoid boundary problem, portions of New Jersey in 10km range of New York city are also added to the network layer (figure 4.2). The spatial weight metric of building on road network is calculated as discussed in section 4.1 above and further unweighted segment angular analysis attributes are calculated as per mathematical logic discussed in 2.3.1 and weighted segment angular analysis attributes are calculated as per mathematical logic stated in section 4.2.1.

4.5 Clustering to identify urban attractor values

Clustering in context of spatial data is discussed in section 3.13. Autocorrelation is good enough in understanding variability of feature in context of its own values in neighborhoods or values of a different feature in neighborhood. However, to analyze the meanings conveyed by clusters of configurational parameters and it is required to drop the neighborhood constraint from clustering. This gives flexibility to relate spatial statistics of similar areas far apart from each other. Unlike autocorrelation, clustering in not necessarily indicator of spatial association and is not driven by correlation of features rather is built on theory of independence of features.

4.5.1 K-Means Clustering

In clustering problems, there are n data points. ‘k’ is the number of clusters in which data points are to be partitioned. The approach used for K-means clustering method aims to find mean positions for ‘k’ clusters such that distance between these means and data points in the cluster is minimized. K-means clustering solves:

$$argmin_c \sum_{i=1}^{k} \sum_{x \in c_i} d(x, \mu_i) = argmin_c \sum_{i=1}^{k} \sum_{x \in c_i} \| x - \mu_i \|_2^2 \qquad (4.7)$$

Where,

ci = set of points that belong to cluster i.

The K-means clustering uses the square of the Euclidean distance:

$$d(x, \mu_i) = \| x - \mu_i \|_2^2 \qquad (4.8)$$

This problem is NP-hard, the K-means algorithm only hopes to find the global minimum, possibly getting stuck in a different solution of local minimum. Discussion of global and local minimum and concept of NP-hard problems is out of the scope of this study

4.5.2 K-Means Algorithm

The Lloyd's algorithm [42], mostly known as k-means algorithm, is used to solve the k-means clustering problem and works as follows. First, decide the number of clusters k. Then:

1. Initialize the center of the clusters: μi= some value, i=1,...,k.
   Attribute the closest cluster to each data point:

$$c_i = \{j : d(x_j, \mu_i) \leq d\,(x_j, \mu_l), l \neq i, j = 1, \ldots n\} \tag{4.9}$$

2. Set the position of each cluster to the mean of all data points belonging to that cluster:

$$\mu_i = \frac{1}{|c_i|} \sum_{j \in c_i} x_j, \quad \forall \mathrm{i} \tag{4.10}$$

3. Repeat steps 2-3 until convergence. |c|= number of elements in c.

4.5.3 K Means++ for initialization

The *k*-means problem is to find cluster centers that minimize the intra-class variance, i.e. the sum of squared distances from each data point being clustered to its cluster center (the center that is closest to it). The standard approach to finding an approximate solution (often called Lloyd's algorithm or the *k*-means algorithm) is used widely and frequently finds reasonable solutions quickly.

However, the *k*-means algorithm has at least two major theoretic shortcomings:

- First, it has been shown that the worst case running time of the algorithm is super-polynomial in the input size.
- Second, the approximation found can be arbitrarily bad with respect to the objective function compared to the optimal clustering.

The k-means++ algorithm [43] addresses the second of these obstacles by specifying a procedure to initialize the cluster centers before proceeding with the standard *k*-means optimization iterations. With the k-means++ initialization, the algorithm is guaranteed to find a solution that is O(log *k*) competitive to the optimal *k*-means solution.

The intuition behind this approach is that spreading out the *k* initial cluster centers is a good thing: the first cluster center is chosen uniformly at random from the data points that are being clustered, after which each subsequent cluster center is chosen from the remaining data points with probability proportional to its squared distance from the point's closest existing cluster center.

The exact algorithm is as follows:

1. Choose one center uniformly at random from among the data points.
2. For each data point *x*, compute D(*x*), the distance between *x* and the nearest center that has already been chosen.
3. Choose one new data point at random as a new center, using a weighted probability distribution where a point *x* is chosen with probability proportional to D(*x*)^2.
4. Repeat Steps 2 and 3 until *k* centers have been chosen.
5. Now that the initial centers have been chosen, proceed using standard *k*-means clustering.

This seeding method yields considerable improvement in the final error of *k*-means. Although the initial selection in the algorithm takes extra time, the *k*-means part itself converges very quickly after this seeding and thus the algorithm lowers the computation time.

4.5.4 Results and Discussions

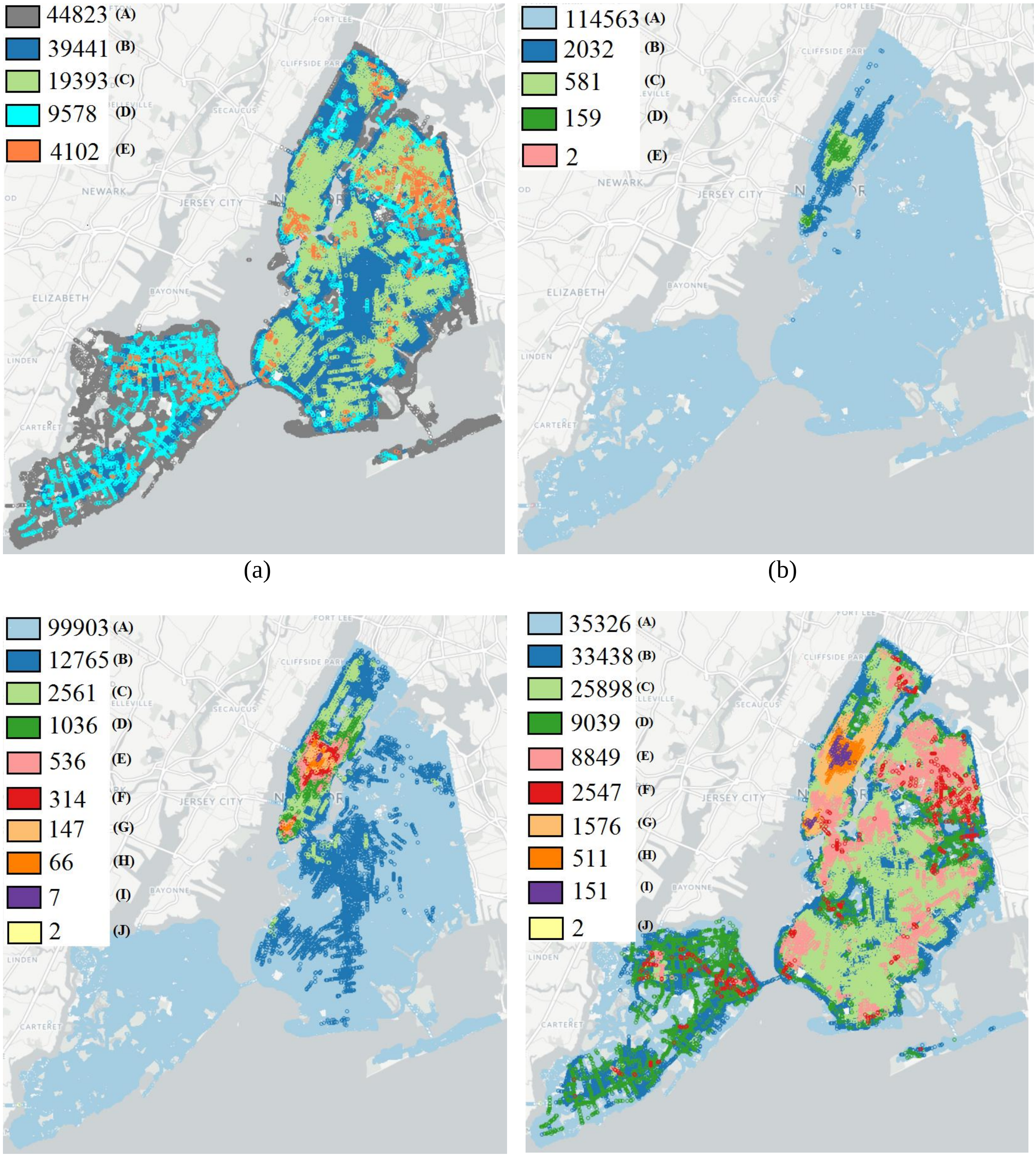
We choose to go from a local toward global radius in terms of space syntax based configurational attributes of choice and integration and apply K Means clustering after initializing the means using K means++ algorithm discussed above. To choose the number of clusters, we use empirical way by applying K-means clustering with different number of clusters and measure the resulting sum of squares of Euclidean distances of cluster points from respective centers. In the process, we draw a comparison between clustering applied on features calculated for unweighted road elements against that on weighted road elements. The conclusive clusters are obtained when clustering is performed on features obtained from both weighted and unweighted computations of segment angular analysis. The increasing search radii helps in capturing movement patterns better as the clusters of streets segments tend to go towards more continuous rather than patchy distributions. There are clear advantages of using the weighted road elements as we derive better meanings from the clusters, it is evident from the results for metric search radii of 2km, 5km and 10km for segment angular analysis as shown in figure 4.3, 4.4 and 4.5.

For unweighted features, 5 is the optimal number of clusters. However, under weighted condition 5 clusters are consumed in small region of Manhattan itself due the large variability in building sizes and street occupancy in those regions. For clusters generated by mixed combination of weighted and unweighted features, clusters appear to be suggestive of real world scenario. In figure 4.3 (d), clusters E, F, G, H, I are suggestive of high occupancy buildings appearing on well-connected streets. Clusters C and D are suggestive of mid or small sized buildings appearing on preferred connection links that form urban core. Regions in clusters A and B appear to have suffered natural boundary effects.

As we move towards global radius, the clusters generated from mixed features capture global movement patterns and very cleanly distinguish the regions based on their infrastructural set up. This is clearly observable from figure 4.4 (c) and 4.5 (c).

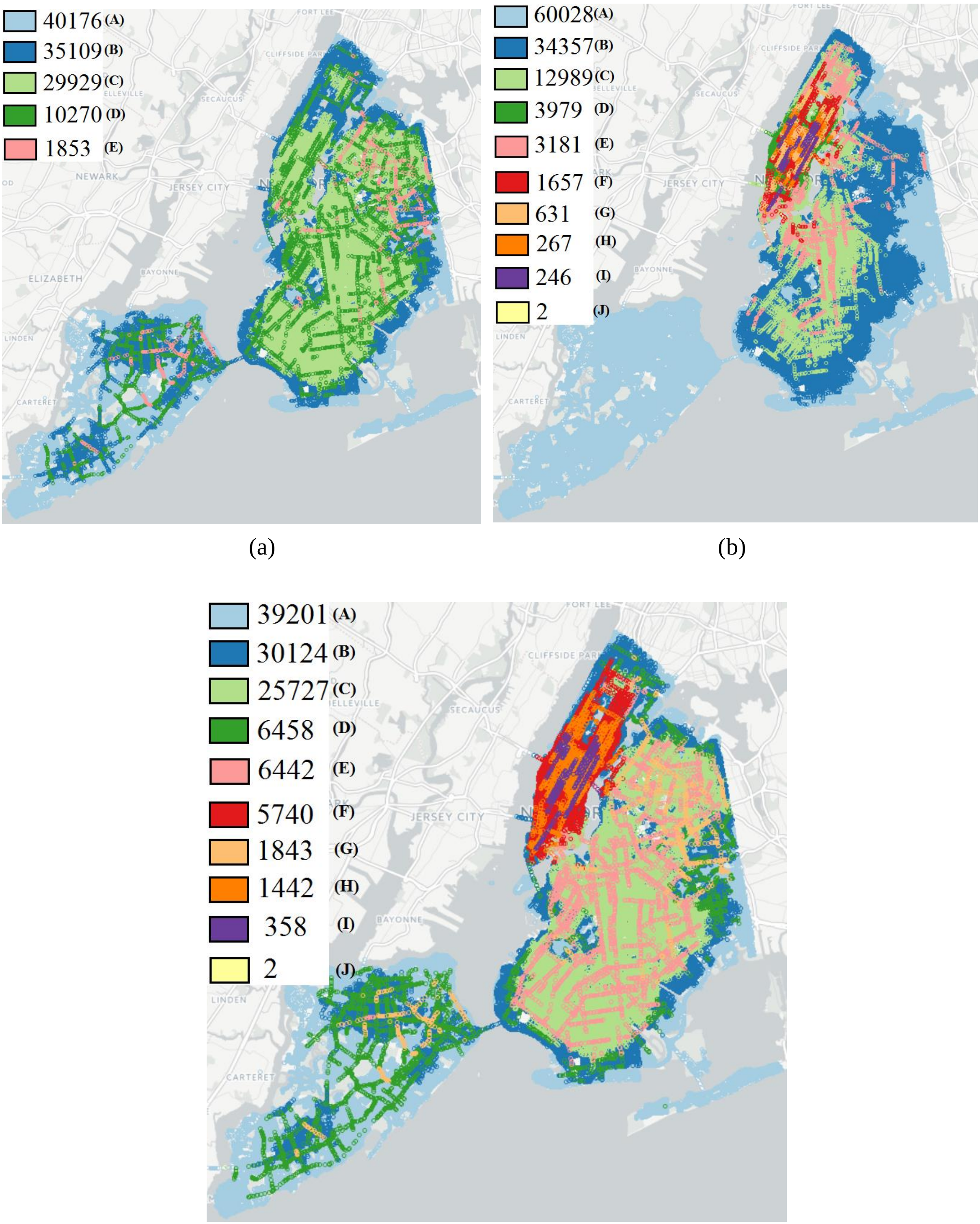
The core represented by clusters F, G, H and I in figure 4.5 (c) closely resemble the high traffic core represented for traffic in figure 4.5 (d) which shows the average number of vehicles passing through streets in a single day. This proves that the mixed cluster model generated from weighted and unweighted segment angular analysis attributes captures global and local movement patterns.

The infrastructural set up of regions represented by ‘3’ in figure 4.5 (c) is shown in figure 4.6 which is more of residential area with single storied building set up. The regions represented by ‘2’ in figure 4.5 (c) are mixed residential and economic areas with 2-3 storied buildings in compact set up as shows in figure 4.7, and regions marked by ‘1’ in figure 4.5 (c) are fully commercial zones with very high rise buildings as shown in figure 4.8. All these variabilities in movement and infrastructure is not captured in case of clusters generated from standard segment angular analysis.



(a) (b) (c) (d)

Figure 4.3 Clusters at radius of 2km for unweighted segment angular analysis configuration (a), Weighted configuration (b, c), mixed configuration (d)

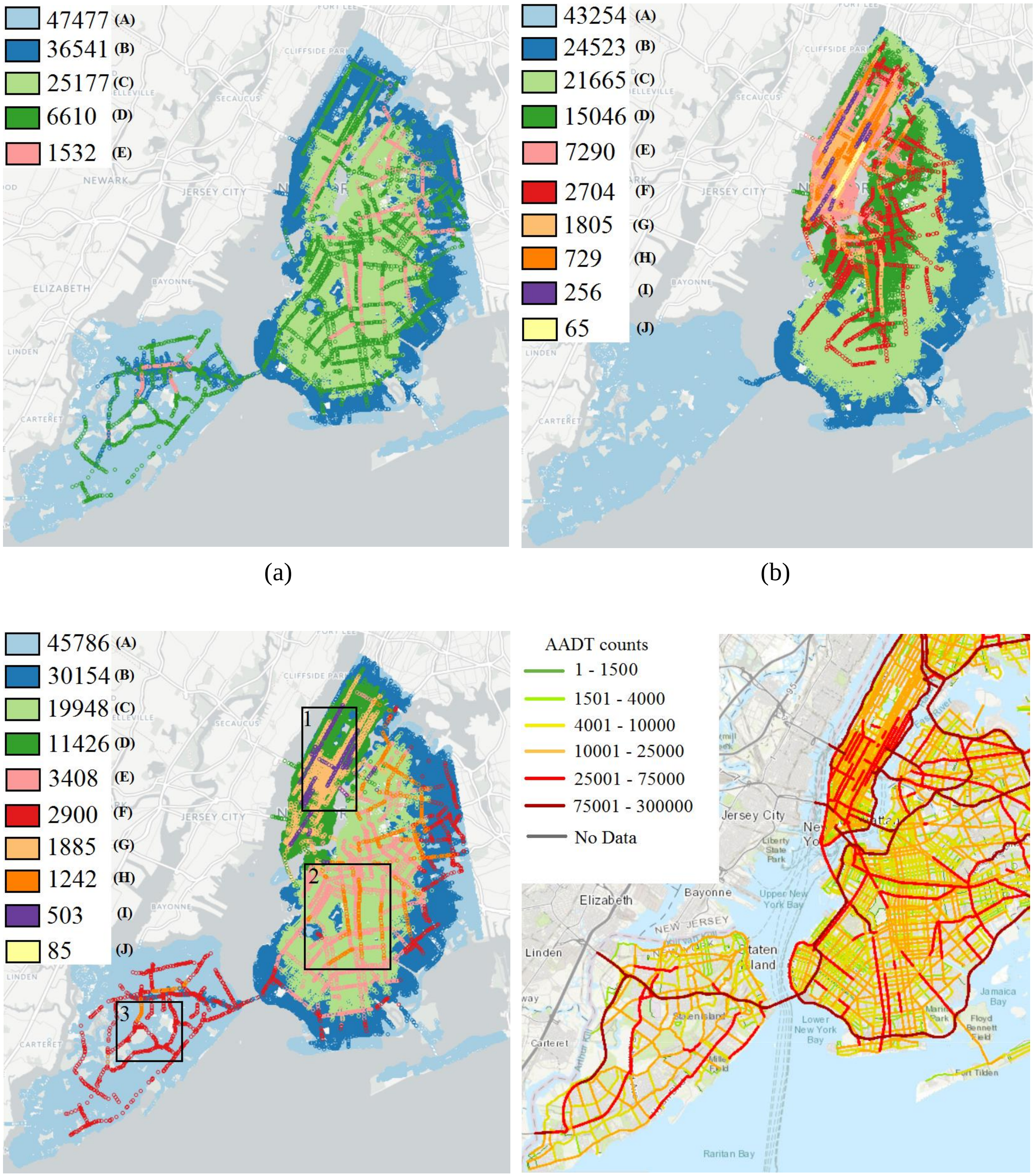


(a)

(b)

(c)

Figure 4.4 Clusters at radius of 5km for unweighted segment angular analysis configuration (a), Weighted configuration (b), mixed configuration (c)



(a) (b)

(c) (d)

Figure 4.5 Clusters at radius of 10km for unweighted segment angular analysis configuration (a), Weighted configuration (b), mixed configuration (c), Average Annual Daily Count of vehicular traffic representations for analysis area (d)

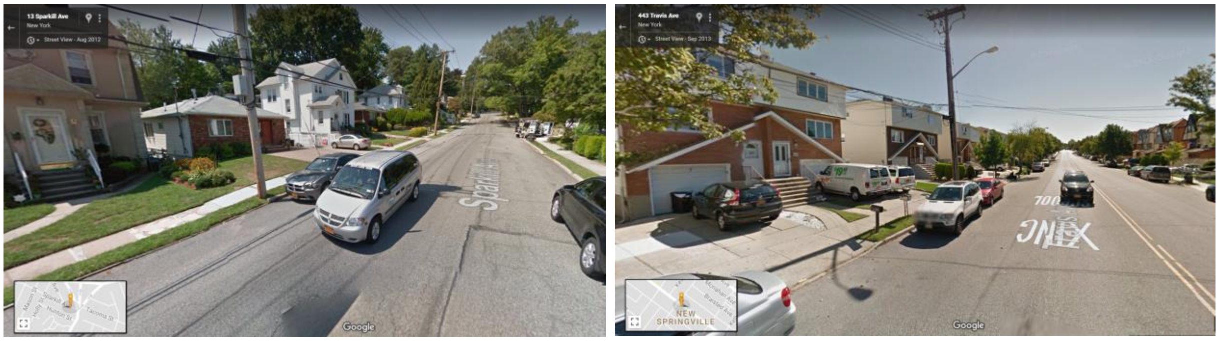


Figure 4.6 Western New York regions

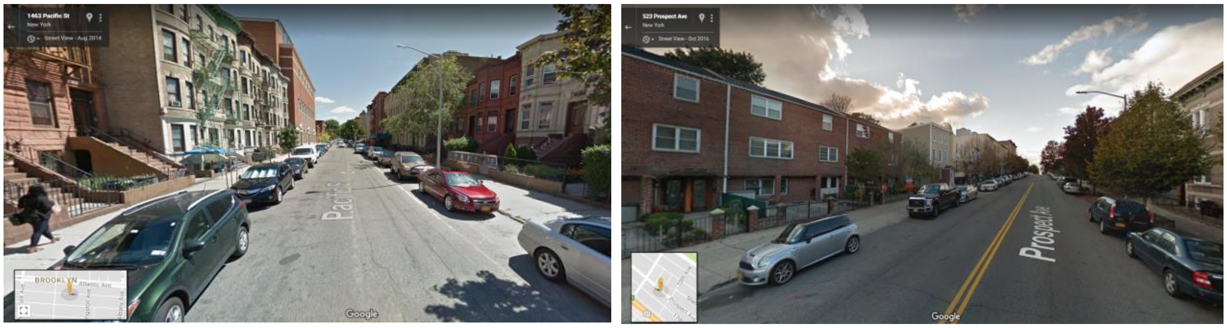


Figure 4.7 Brooklyn and nearby regions

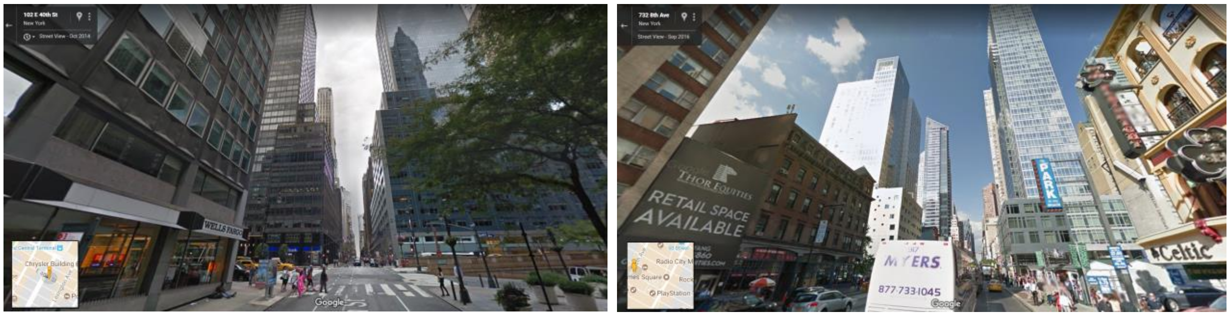


Figure 4.8 Manhattan (Near Empire State Building and Times Square)
(The images are courtesy of Google Maps – Street View)

4.6 Classification for predicting movement affinities

Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, based on a training set of data containing observations (or instances) whose category membership is known. In the terminology of machine learning, classification is considered an instance of supervised learning, i.e. learning where a training set of correctly identified observations is available. In context of this study the road segments are given labels of ‘high’, ‘medium’ and ‘low’ according to Annual Average Daily Traffic counts obtained from New York traffic data repository [40]. These labels are then predicted using classification techniques using configurational features from segment angular analysis, both weighted and unweighted. Further sections discuss the advantages of weighted metric proposed.

4.6.1 Average Annual Daily Traffic (AADT) Count Data

New York city has placed vehicle counters on several primary and secondary roads of the city. The data exposed for public use is a single GIS layer of road elements that contains the daily vehicle counts averaged over entire year days captured by these counters. The reference map for AADT data layer is represented by figure 4.5 (d).

4.6.1.1 Labelling AADT data set

The peak times for vehicular traffic are from 8:00 AM to 1:00 PM and from 4:00 PM to 7:00 PM. The vehicular traffic during the 10 hours of constant peak hours traffic is almost equal to the total traffic during the day (Appendix A). The mean and maximum decelerations for vehicles are discussed in [44], we consider the maximum deceleration values of cars for deciding threshold limits for low, medium and high amount of traffic. However, the deceleration rates also tend to vary by the nature of streets but this study uses 1.625m/s^2 as suggested for cars as the common deceleration rate. All the roads are considered bi-lane streets. Going by this logic, a vehicle going at nearly 16m/s would get to a Zero speed in 80m. This is assumed to be safe distance for vehicles moving at 16m/s. With this speed, a vehicle can cover nearly a Km in a minute. And a separation of 80m between 2 vehicles would cause around 25 vehicles to pass through a point (the vehicle counter in this case) in a minute considering both the lanes. As discussed above, this traffic is nearly equal to 10 hours of peak time traffic (Appendix A). This come to 15000 vehicles. Any count of AADT greater than this treated as High Traffic. The street segments that carry lesser than 6000 vehicles in a day are labelled as low and all other street segments are labelled with medium traffic carriage.

We use classification to predict these traffic labels using the configurational attributes as training features. Following are the three feature sets, we will be using throughout the study:

Feature set 1: {Unweighted_Choice_2km, Unweighted_Choice_5km, Unweighted_Choice_10km, Unweighted_Integration_2km, Unweighted_ Integration _5km, Unweighted_ Integration _10km}

Feature set 2: {Weighted_Choice_2km, Weighted _Choice_5km, Weighted _Choice_10km, Weighted _Integration_2km, Weighted _ Integration _5km, Weighted _ Integration _10km}

Feature set 3: {Unweighted_Choice_2km, Unweighted_Choice_5km, Unweighted_Choice_10km, Unweighted_Integration_2km, Unweighted_ Integration _5km, Unweighted_ Integration _10km, Weighted_Choice_2km,Weighted_Choice_5km,Weighted_Choice_10km,Weighted_Integration_2km, Weighted _ Integration _5km, Weighted _ Integration _10km}

Following sections discuss the use of different techniques used to classify road segments to get the model with best accuracy.

4.7 Logistic Regression

Logistic regression is a model where the categorical dependent variable is predicted using set of independent variables [45]. It is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables (which may be real-valued, binary-valued, categorical-valued, etc.).

The logistic classifier uses a discrete target variable y. For each observation, the probability that y=1 is modeled as the logistic function of a linear combination of the feature values. Given a set of features xi, and a label yi ∈ {0,1}, logistic regression interprets the probability that the label is in one class as a logistic function of a linear combination of the features.

$$f_i(\theta)=p(y_i=1|x)=\frac{1}{1+\exp(-\theta^T x)} \tag{4.11}$$

An intercept term is added by appending a column of 1’s to the features. Regularization is often required to prevent over fitting by penalizing models with extreme parameter values. The logistic regression module supports l_1 and l_2 regularization, which are added to the loss function.

The composite objective being optimized for maximizing the accuracy of the model in terms of predicting true labels for training sample is the following.

$$\min_{\theta} \sum_{i=1} f_i(\theta) + \lambda \parallel \theta \parallel_1 + \ \lambda_2 \parallel \theta \ \parallel_2^2 \tag{4.12}$$

Where, λ_1 is l_1_penalty the and λ_2 is the l_2_penalty. Discussion of penalties and overfitting are not under the scope of this study. The class for which the probability is maximum is assigned to the observation point while testing. With each iteration, the training accuracy increases as error keeps on reducing. The stopping criterion for this study is number of iterations, when testing error either becomes constant or starts to increase.

4.7.1 Results and discussions for classification using Logistic Regression

Feature Set	Accuracy
Feature Set 1	Training Accuracy = 43%, Testing Accuracy = 43%
Feature Set 2	Training Accuracy = 41%, Testing Accuracy = 40%
Feature Set 3	Training Accuracy = 47%, Testing Accuracy = 46%

Table 4.1 Classification accuracy using logistic regression

The best trained model using logistic regression does not give high accuracy as seen in table 4.1. The testing accuracy on model is 43% using training set 1, whereas accuracy is 40% using feature set 2. The accuracy increases to 46% when feature set 3 is used. This gives clear indication that, like in the case of clustering, using the configurational features from both weighted and unweighted segment angular analysis is a better syntactic representation of urban environment.

4.8 Decision Trees

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility [46]. A decision tree is a flowchart-like structure (as shown in figure 4.9) in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The decision is usually taken such that the resultant split maximizes the accuracy of training accuracy of classification of all intermediate states of the final tree. Thus, in the process, next best feature to split on is considered at all the states of tree formation. The state when no more splits are made further is called the stopping criterion which is out the scope of this study. The max depth of the tree we have considered is 6 as we have very few features to split on. The paths from root to leaf represent classification rules.

In decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.

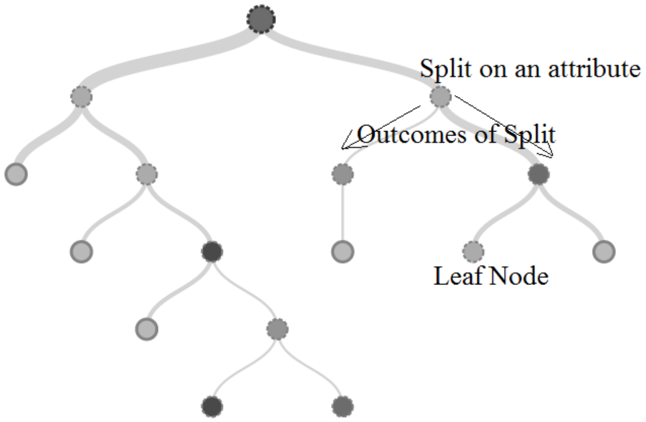


Figure 4.9 Decision tree representation

Decision trees can be described also as the combination of mathematical and computational techniques to aid the description, categorization and generalization of a given set of data.

Data comes in records of the form: $(x, Y) = (x_1, x_2, x_3 \ldots\ldots, x_k, Y)$

The dependent variable, Y, is the target variable that we are trying to understand, classify or generalize. The vector x is composed of the input variables, x_1, x_2, x_3 etc., that are used for that task.

Thus, Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). The stopping criterion for training in our case is depth of the tree/ or when the number of features used in training the model have been exhausted which usually is the case as we have limited number of features in feature set.

4.8.1 Results and discussions for classification using Decision Tree classification

Feature Set	Accuracy
Feature Set 1	Training Accuracy = 47%, Testing Accuracy = 47%
Feature Set 2	Training Accuracy = 49%, Testing Accuracy = 48%
Feature Set 3	Training Accuracy = 50%, Testing Accuracy = 49%

Table 4.2 Classification testing accuracy using decision trees

The classification accuracy using decision trees is more than that using logistic regression as seen in table 4.2. The best trained model using decision trees is 47% using training set 1, whereas accuracy is 48% using feature set 2. A similar trend of increase in classification accuracy is observed when feature set 3 is used for training, the testing accuracy is 49% in this case.

4.9 Ensemble based gradient boosted trees

Gradient boosting is a technique which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. gradient boosting combines weak "learners" into a single strong learner in an iterative fashion. It is easiest to explain in the least-squares regression setting, where the goal is to "teach" a model ‘F’ to predict values in the form, $\hat{y} = \mathrm{F}\ (\mathrm{x})$ by minimizing the mean squared error $(\hat{\mathrm{y}} - \mathrm{y})^2$, averaged over some training set of actual values of the output variable y [47].

At each stage m, 1 <= m <= M, of gradient boosting, it may be assumed that there is some imperfect model F_m (at the outset, a very weak model that just predicts the mean y in the training set could be used). The gradient boosting algorithm improves on F_m by constructing a new model that adds an estimator h to provide a better model: $F_{m+1}(x) = F_m(x) + h(x)$. To find h, the gradient boosting solution starts with the observation that a perfect h would imply $F_m+1(x)=F_m(x)+h(x) = y$ or, equivalently, $h(x) = y - F_m(x)$. Therefore, gradient boosting will fit h to the residual $y - F_m(x)$. The number of iterations correspond to number of trees in the model. With each iteration, a weak learner is added to the model that increases the resultant training accuracy. The number iterations is the stopping criterion for training. We experiment with 5, 10, 50, 100, 200 and 500 iterations to get the threshold where testing accuracy does not increase significantly or begins to decrease causing overfitting.

4.9.1 Results and discussions for classification using gradient boosting Classification

The classification models trained using gradient boosting gives even better solutions compared to decision trees. The convergence for all the three training sets occurs at close to 100 iterations of training. The best trained models have training accuracies of 54%, 58% and 63% respectively for feature sets 1,2 and 3 as shown in table 4.3.

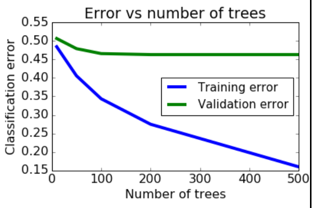
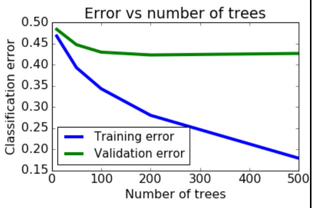
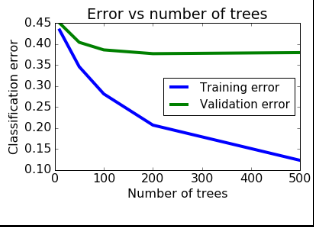
Feature Set	Iterations	Accuracy	Error Graph
Feature Set 1	5	Training Accuracy = 49%, Testing Accuracy = 48%	
	10	Training Accuracy = 51%, Testing Accuracy = 49%	
	50	Training Accuracy = 52%, Testing Accuracy = 52%	
	100	Training Accuracy = 66%, Testing Accuracy = 53%	
	200	Training Accuracy = 73%, Testing Accuracy = 54%	
	500	Training Accuracy = 84%, Testing Accuracy = 54%	
Feature Set 2	5	Training Accuracy = 49%, Testing Accuracy = 50%	
	10	Training Accuracy = 52%, Testing Accuracy = 52%	
	50	Training Accuracy = 54%, Testing Accuracy = 55%	
	100	Training Accuracy = 67%, Testing Accuracy = 57%	
	200	Training Accuracy = 75%, Testing Accuracy = 58%	
	500	Training Accuracy = 84%, Testing Accuracy = 58%	
Feature Set 3	5	Training Accuracy = 52%, Testing Accuracy = 53%	
	10	Training Accuracy = 57%, Testing Accuracy = 55%	
	50	Training Accuracy = 54%, Testing Accuracy = 59%	
	100	Training Accuracy = 67%, Testing Accuracy = 61%	
	200	Training Accuracy = 75%, Testing Accuracy = 63%	
	500	Training Accuracy = 84%, Testing Accuracy = 62%	

Table 4.3 Classification accuracy using ensemble gradient boosting

4.10 Towards increasing Classification Accuracy

The decision tree models rely on the number of observation points with similar features to put them into same classes or deciding the split in that manner. A split on certain feature may sometime result in points from same class being put into leaves representing separate classes or vice versa. Not having enough data-points representing a class may lead to low accuracy decision tree model. In this case study, the network layer is used, though the weighted measures have considered building capacities for all calculations but it is still prone to similar configurational attributes for different symmetries of arrangement. Considering a case, where large buildings are far apart from each other opposed to where relatively smaller buildings are very closely spaced. These two areas are significantly different from structural perspective but configuration suggests that areas are similar. The may actually observe different patterns of vehicular movements but may end being classified in same class by the model. Moving further, we superimpose the configurational features on buildings using spatial join as discussed in section 3.7. Each building has all configurational features of road segment nearest to it computed under weighted and unweighted scenarios along with AADT labels of the roads. We train the gradient boosted tree model using the feature set of all configurational attributed with AADT labels as target classes and analyze the accuracies of models under varying iterations.

4.10.1 Results and discussions for gradient boosting classification used on buildings

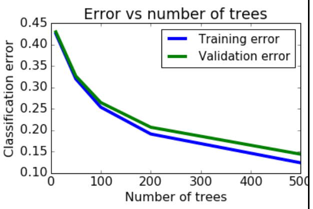
Feature Set	Iterations	Accuracy	Error Graph
Feature Set 3	5	Training Accuracy = 54%, Testing Accuracy = 54%	
	10	Training Accuracy = 57%, Testing Accuracy = 55%	
	50	Training Accuracy = 68%, Testing Accuracy = 67%	
	100	Training Accuracy = 75%, Testing Accuracy = 74%	
	200	Training Accuracy = 80%, Testing Accuracy = 79%	
	500	Training Accuracy = 88%, Testing Accuracy = 86%	

Table 4.4 Classification accuracy using ensemble gradient boosting on buildings

Target label	Predicted label	Count
High	High	**22416**
High	Medium	1552
High	Low	1339
Medium	High	1340
Medium	Medium	**21995**
Medium	Low	2103
Low	High	1161
Low	Medium	1596
Low	Low	**22552**

Table 4.5 Confusion Matrix for classification on testing data using gradient boosting on buildings

As expected, the classification performance increases tremendously by using buildings as the training layer. The gradient boosted model built using feature set 3 using buildings layer gives over 85% (table 4.4) of testing accuracy even when it has not converged fully after 500 iterations. The confusion matrix (table 4.5) confirms that only a small fraction of points with ‘high’ labels are misclassified. This model is highly efficient and can be used for prediction and planning.

4.11 Summary

This chapter proposed a building weighted metric on road network for urban analysis. This approach helped in capturing buildings information on road. The metric was used for weighted calculation of segment angular analysis. This approach is more efficient in a way as it does not require to calculate large number of building to building shortest distances as that in the hybrid approach discussed in previous chapter. Further clustering and classification approaches were used to denominate attractor values of spaces with high accuracy.

Chapter 5

Discussions and Conclusions

The thesis discusses major challenges and limitations of approaches that are used till date to understand urban configuration. The configurational attributes discussed in past have often been discussed pertaining to individual spatial units which themselves are characterized on the basis of individual configurational attributes. This study proposed a framework that utilizes the configuration in a hybrid form that captures the intrinsic interaction of road network and building. The work also focusses on discussing behavioral characteristics of areas as whole instead of individual spatial units of either road or building, by pipelining autocorrelation in the framework using all relevant configurational attributes at once. The approach helped us in capturing homogeneity in urban environment at local scales, and at the same time it helped in capturing the high amount of variability that may exist in any urban setting. All this is achieved by understanding the semantics of the configurational attributes and designing semantic rules for capturing movement attractor values of spaces. The framework is devised such that it captures local as well as global scale of movement affinities for spaces based on various trip distances.

The framework that we proposed is built on top of buildings layer which is effective in capturing semantic properties at scale, but this hybrid model too has limitations are majorly the computation time to generate configurational attributes from urban network analysis as it considers building to building trip distances and buildings are too many in number as compared to number of network nodes. Secondly, it misses out on regions that have network but no buildings. To tackle these challenges, the study proposed a building weight metric on top of road network and recomputed configurational parameters under weighted setting. This model when clubbed with unweighted model for urban configuration is better representation that is more robust and has computation budget less than that of the proposed hybrid approach.

Using the outcomes from the autocorrelation analysis approach that there exists continuity of similarities in urban regions, the study further aimed that generating clustering and classification models that could capture movement affinities and could easily dissociate dissimilar areas into separate categories. The similar models can be utilized for urban plans such as road widening, emergency evacuation, cycling routes and building an entirely new urban space.

The case studies hinted at heterogeneity in global urban environment but homogeneity at local scales. The cities were earlier developed in an opportunist fashion where needs drove the development. But the cities these days are more planned and designs are more symmetric. A future work could be understanding the viability and sustainability of the two design approaches for city development. As mentioned above, the application utilities of methodologies proposed is this study is too high as they can used at any stage of site development starting from planning to making improvements in the existing setting.

The approaches discussed in the study are very important from the perspective of city management as they consider a thorough accessibility analysis of spaces with exact load estimations using both layers of roads and buildings. Also, with the GIS data being easily available these days, from the planning stages of the cities, the approaches suggested by us can definitely come in handy to pre-evaluate the plan with valid justifications.

The study majorly discussed attractor values in terms of movement patterns, approaches to capture other spatial attractor values using syntactic configuration can be devised in future.

Related Publications

1. Exploring the impact of road traffic impedance and built environment for vulnerability mapping of evacuation areas – Case study of Hyderabad city. Rajesh Chaturvedi, K S Rajan. In Proceedings of 10th Space Syntax Symposium, 2015. London, UK

2. Traffic flow estimates based on structure of built environment, a case study of New York City. Rajesh Chaturvedi, K S Rajan.
   Accepted at - 15th International Conference on Computers in Urban Planning and Urban Management, 2017, Adelaide, Australia

Bibliography

[1] Sayed K. Al, Turner A., and Hanna S. Cities as Emergent Models - The Morphological Logic of Manhattan and Barcelona. 2009. 7th International Space Syntax Symposium. Edited by Koch D., Marcus L., and Steen J., Stockholm: KTH, 2009. Ref 001:1- 001:12.

[2] Major M. D., Penn A., Spiliopoulou G., Spende N., Doxa M., and Fong P. In with the right crowd - crowd movement and space use in Trafalgar Square during the New Year's Eve celebrations. 1999. 2nd International Space Syntax Symposium, Brasília, March-April 08.1-08.17.

[3] Hillier B. Space is the machine - Space Syntax. 1996. Cambridge University 1-368.

[4] Figueiredo L., Amorim L. 2007. Decoding the urban grid: or why cities are neither trees nor perfect grids. Proceedings, 6th International Space Syntax Symposium, İstanbul, 006:1-16.

[5] Hyeyoung K., and Chul M. J. Spatial Analysis of the relationship between space syntax and land use density. 2013. Proceeding of Ninth International space syntax Symposium, Seoul 125:1-125:9.

[6] Jingnan H. Study on Spatial Structure of Large Scale Retails Store Based on Space Syntax Case study in Wuhan. 2009. The International Institute for Geo-Information Science and Earth Observation Enschede, Netherlands 1-79 March.

[7] Penn A. The complexity of the elementary interface: shopping space University College London, UK. 2007. Proceedings, 6th International Space Syntax Symposium, İstanbul, 103:1-103:12.

[8] Hillier B., and Sahbaz O. An evidence based approach to crime and urban design. Designing Sustainable Cities: Decision-making Tools and Resources for Design, Wiley Blackwell. 2009. pp. 163-186.

[9] Nubani L., and Wineman. J. The role of space syntax in identifying the relationship between space and crime. 2005. Proceedings of the Fifth International Space Syntax Symposium.

[10] Baran P. K., Smith W. R., and Toker U. The Space Syntax and Crime: Evidence from a Suburban Community. 2007. Short paper presented at the 6th International Space Syntax Symposium, Istanbul.

[11] Jiang B. Ranking spaces for predicting human movement in an urban environment. July 2009. International Journal of Geographical Information Science, vol. 23, issue 7, pp. 823-837.

[12] Hillier B. The Social Logic of Space. 1989. Bell & Bain, Cambridge University Press, pp. 1-281.

[13] Sevtsuk A. Analysis and Planning of Urban Networks. 2014. In Alhajj R. & Rokne J. (Eds.), Encyclopedia of Social Network Analysis and Mining, pp. 2437.

[14] Turner A. Getting Serious with Depthmap Segment Analysis and Scripting. January 2008, http://archtech.gr/varoudis/depthmapX/LearningMaterial/advanceddepthmap.pdf [Accessed on January 1, 2017]

[15] Isovist - https://en.wikipedia.org/wiki/Isovist#/media/File:Isovist.svg. [Accessed on January 1, 2017]

[16] Representation AK3 - http://representationak3.blogspot.in/2011/10/isovist-using-grasshopper-in-rhino.html. [Accessed on January 1, 2017]

[17] Batty M. The new science of cities article. December 2014. International Journal of Geographical Information Science, pp. 345-347. DOI: 10.1080/13658816.2014.937717.

[18] Jiang B. A space syntax approach to spatial cognition in urban environments, Cognitive Models of Dynamic Phenomena and Their Representations, October 29 - 31, 1998. University of Pittsburgh, Pittsburgh, PA.

[19] Ratti C. Space syntax: some inconsistencies. 2004 Environment and Planning B: Planning and Design, vol. 31, issue 4, pages 487-499.

[20] Ratti C. Suggestions for developments in space syntax. 2004 Senseable City Laboratory, MIT, USA.

[21] Ratti, C. Rejoinder to Hillier and Penn. 2004. Environment and Planning B - Planning and Design, vol. 31, issue 4, pp. 63-74. Pion Ltd. DOI:10.1068/b3019b.

[22] Hillier B., and Penn A. Rejoinder to Carlo Ratti. 2004 Environment and Planning B - Planning and Design, vol. 31, issue 4, pp. 501-511. Pion Ltd, DOI:10.1068/b3019a.

[23] Hillier B, Penn A, Hanson J., Grajewski T., and Xu J. Natural movement: or, configuration and attraction in urban pedestrian movement. 1993. Environment and Planning B: Planning and Design, vol. 20, pp. 29-66.

[24] Turner A. From Axial to Road-Centre Lines: A New Representation for Space Syntax and a New Model of Route Choice for Transport Network Analysis. 2007. Environment and Planning B, vol. 34, issue 3, pp. 539-555.

[25] Turner A. Could A Road-centre Line Be an Axial Line In Disguise? 2003. 4th International Symposium on Space Syntax, in van Nes (ed). University College London, UK.

[26] Sevtsuk A, and Mekonnen M. Urban network analysis. A new toolbox for arcgis. 2012. Revue Internationale de Géomatique, vol. 22, issue 2, pp. 287-305.

[27] Freeman C.L. A Set of Measures of Centrality Based on Betweenness. March 1977. Sociometry, vol. 40, issue 1, pp. 35-41.

[28] Sayed Al. K., Turner A., Hillier B., Iida S., and Penn A. 2014. Space Syntax methodology. http://discovery.ucl.ac.uk/id/eprint/1415080 [Accessed on January 1, 2017]

[29] Henley S. Nonparametric Geostatistics. 1981. Applied Science Publishers: London, UK.

[30] Haining R. Spatial Data Analysis in Social and Environmental Sciences. 1990. Cambridge University Press: New York, NY.

[31] Upton J. G. G., and Fingleton B. Spatial Data Analysis by Example. 1985. Point Pattern and Quantitative Data, vol. 1: Wiley: Chichester, UK.

[32] Varoudis, T., depthmapx Multi-Platform Spatial Network Analysis Software, Version 0.30 opensource. 2012. http://varoudis.github.io/depthmapx/ [Accessed on June 1, 2017]

[33] Anselin L. Local Indicators of Spatial Association-LISA. 1995. Geographical Analysis vol. 27, pp. 93-115.

[34] Anselin L. Under the hood issues in the specification and interpretation of spatial regression models. 2002. Agricultural economics, vol. 17, pp. 247–267.

[35] Jarrell M. G. A comparison of two procedures, the Mahalanobis Distance and the Andrews-Pregibon Statistic, for identifying multivariate outliers. 1994. Research in the schools, vol. 1, pp. 49-58.

[36] Hawkins D.M. Identification of outliers. 1980. London: Chapman and Hall.

[37] Evans V. P. Strategies for detecting outliers in regression analysis: an introductory primer. 1999. Advances in Social Science Methodology, ed. Thompson B., editor. (Stamford, CT: JAI Press), 271–286.

[38] Turner A., and Dalton N. A simplified route choice model using the shortest angular path assumption. 2005. Proceedings of the 8th International Conference on geo computation, Michigan USA.

[39] http://wiki.openstreetmap.org/wiki/New_York_City [Accessed on July 1, 2016].

[40] https://data.ny.gov/Transportation/Annual-Average-Daily-Traffic-AADT-Beginning-1977/6amx-2pbv/data [Accessed on July 1, 2016].

[41] http://wiki.openstreetmap.org/wiki/New_Jersey [Accessed on July 1, 2016].

[42] Lloyd, Stuart P. Least squares quantization in PCM. Information Theory. 1982. IEEE Transactions on 28.2, pp. 129-137.

[43] Arthur D., Vassilvitskii S. k-means++: the advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. 2007. Society for Industrial and Applied Mathematics Philadelphia, PA, USA, pp. 1027–1035.

[44] Maurya A. K., and Bokare P. S. Study of deceleration behavior of different vehicle types. 2012. International journal for traffic & transport engineering vol.2, issue 3.

[45] Freedman D. A. Statistical Models: Theory and Practice. 2009. Cambridge University Press. pp. 128.

[46] Rokach L, and Maimon O. 2008. Data mining with decision trees: theory and applications. World Scientific Pub Co Inc. ISBN 978-9812771711.

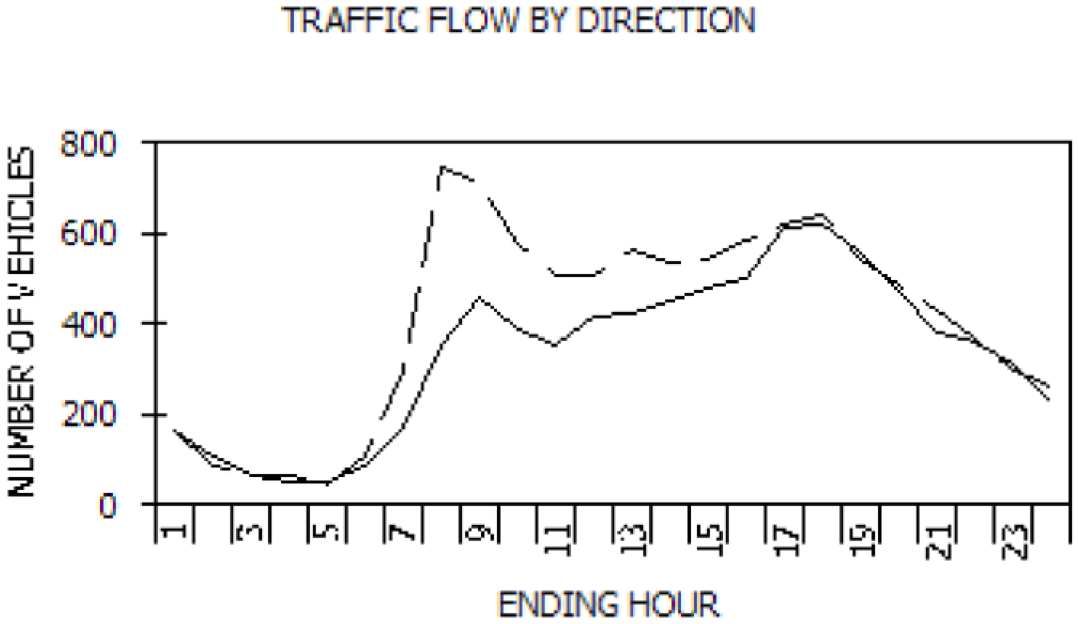
[47] Cheng Li. A Gentle Introduction to Gradient Boosting - http://www.ccs.neu.edu/home/vip/teach/mlcourse/4_boosting/slides/gradient_boosting.pdf [Accessed on June 26, 2017].

[48] Pearson K. Notes on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London. June 1895. pp. 240–242.

Appendix A

Following graphs show variations of traffic captured by traffic count stations at various time of the day in New York. It is observed that if the peak time traffic is assumed constant, then the total traffic counts of the day comes nearly equal to 10 hours of peak traffic. We have used this detail to calculate the minimum separation between the two vehicles to allow then to move at a constant speed of 16m/s.

STATION: 010007

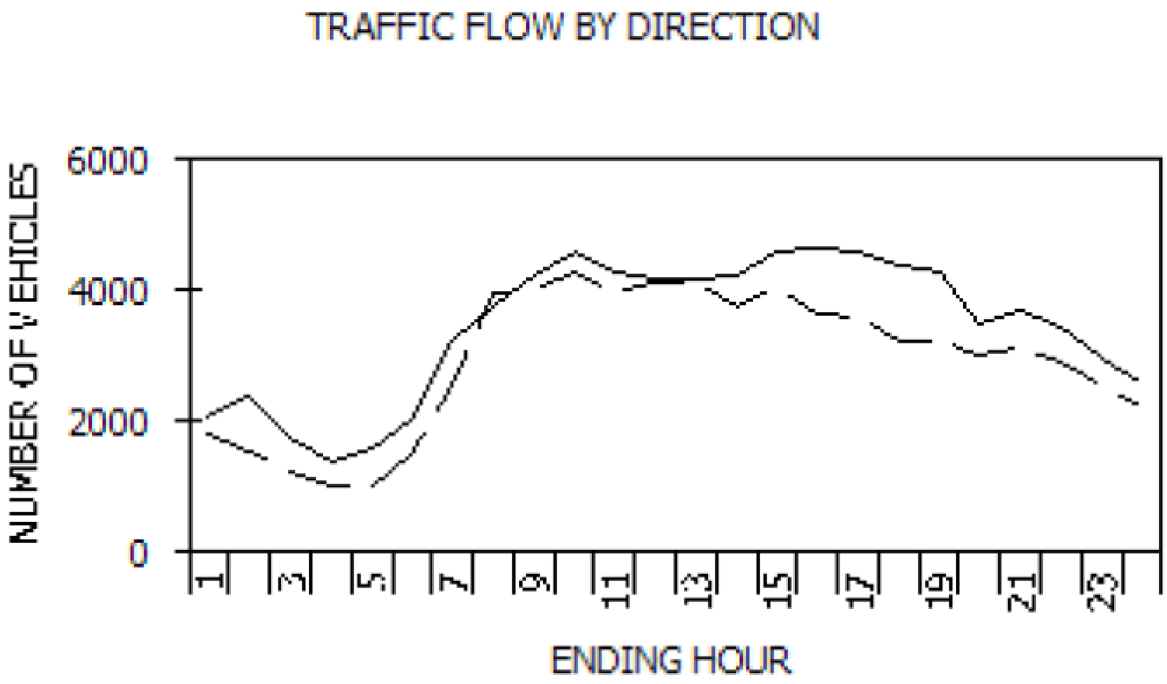


--- North - -South

PEAK HOUR DATA

DIRECTION	HOUR	COUNT	2-WAY	HOUR	COUNT
North	**18**	**618**	**A.M.**	**9**	**1172**
South	**8**	**746**	**P.M.**	**18**	**1257**

STATION: 010033



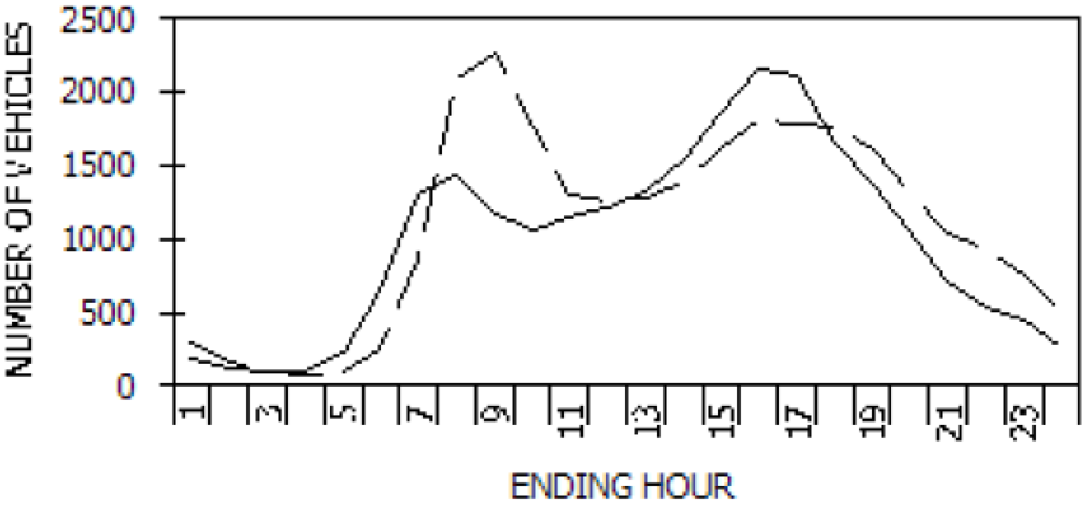
--- East - -West

PEAK HOUR DATA

DIRECTION	HOUR	COUNT	2-WAY	HOUR	COUNT
East	**16**	**4641**	**A.M.**	**10**	**8907**
West	**10**	**4297**	**P.M.**	**15**	**8695**

STATION: 010906

TRAFFIC FLOW BY DIRECTION



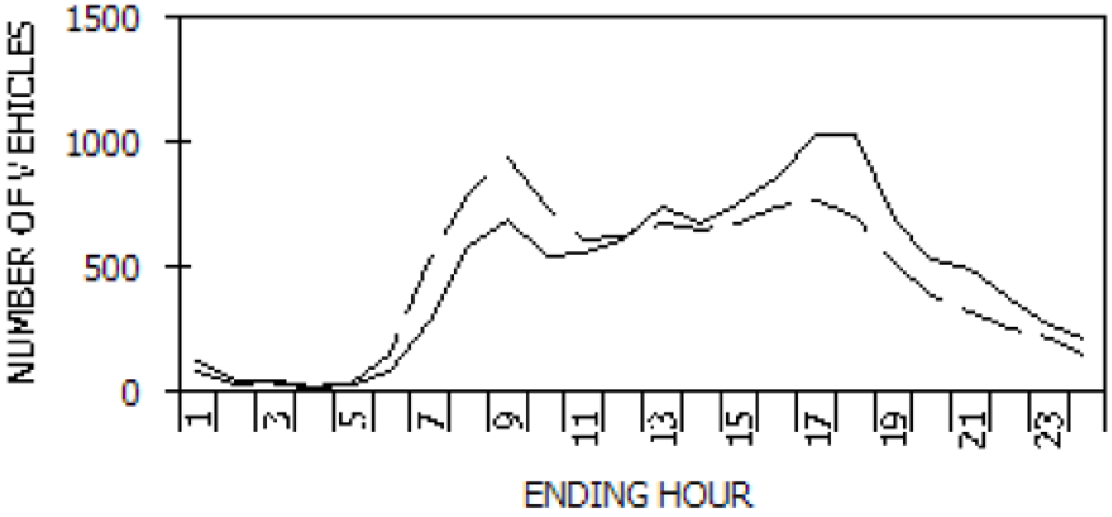
--- East - -West

PEAK HOUR DATA

DIRECTION	HOUR	COUNT	2-WAY	HOUR	COUNT
East	**16**	**2160**	**A.M.**	**8**	**3525**
West	**9**	**2272**	**P.M.**	**16**	**3966**

STATION: 434054

TRAFFIC FLOW BY DIRECTION



--- East - -West

PEAK HOUR DATA

DIRECTION	HOUR	COUNT	2-WAY	HOUR	COUNT
East	**17**	**1034**	**A.M.**	**9**	**1631**
West	**9**	**937**	**P.M.**	**17**	**1800**