



Effect of Different Distance Measures in Result of Cluster Analysis

Master's Thesis

Aalto University School of Engineering,
Department of Real Estate, Planning and
Geoinformatics,
Date: 04.08.2015

Sujan Dahal
Degree Program in Geomatics

Supervisor: Professor Kirsi Virrantaus
Instructor: D.Sc.(Tech) Jussi Nikander

*In Loving Memory Of My Grandmother
Dhana Laxmi Tripathhee
(1923-2014)*

Author Sujan Dahal

Title of thesis Effect of Different Distance Measures in Result of Cluster Analysis

Degree programme Geomatics

Major/minor Geoinformatics **Code** M3002

Thesis supervisor Prof. Kirsi-Kanerva Virrantaus, Department of Real Estate, Planning and Geoinformatics, Aalto University.

Thesis advisor(s) D.Sc. (Tech) Jussi Nikander, Principal Research Scientist, Natural Resources Institute Finland

Date 04.08.2015 **Number of pages** 77 **Language** English

Abstract

The objective of this master's thesis was to explore different distance measures that could be used in clustering and to evaluate how different distance measures in K-medoid clustering method would affect the clustering output. The different distance measures used in this research includes Euclidean, Squared Euclidean, Manhattan, Chebyshev and Mahalanobis distance. To achieve the research objective, K-medoid method with different distance measures was applied to a spatial dataset to explore relative information revealed by each distance measure. The effect of each distance measure on output is documented and the output was further compared with each other to reveal the differences between each distance measure.

The study starts with literature review of cluster analysis process where necessary steps for performing cluster analysis are explained. In literature section, different clustering methods with particular characteristics of each method are described that would serve as basis for choice of clustering method. Data description and data analysis is included thereafter which is followed by interpretation of clustering result and its use for Terrain analysis. Terrain analysis has its significance in forest industry, military as well as crisis management and is usually concerned with off-road mobility of a vehicle or a group of vehicles between given locations. In case of terrain analysis, clustering could be used to group the similar areas and determine the off-road mobility of a particular vehicle. This result could be further categorized according to suitability of the item in the cluster and interpreted using expert evaluation in order to reveal useful information about mobility in a terrain.

Cluster Validation measures were applied to output of clustering to determine the differences between different distance measures. The findings of this study indicate that in the study area, there exists some level of differences in the result of clustering when different distance measures are used. This difference is then interpreted with the help of input dataset and expert opinion to understand the effect of different distance measures in the dataset. Finally, the study provides basis for mobility analysis with help of clustering output.

Keywords Spatial Analysis, Clustering, Similarity Measures, Distance Measures, K-medoid Clustering, Cluster Validation, Mobility Analysis

Table of Contents

LIST OF ABBREVIATIONS	III
LIST OF FIGURES	IV
LIST OF TABLES.....	VI
1. INTRODUCTION	1
1.1 THESIS STRUCTURE.....	4
1.2 RELATED WORKS	4
1.3 OBJECTIVES OF RESEARCH AND RESEARCH QUESTIONS.....	7
1.4 MOBILITY ANALYSIS	8
1.5 METHODS AND MATERIALS	10
2 CLUSTER ANALYSIS	11
2.1 APPLICATION AND OBJECTIVE OF CLUSTER ANALYSIS	12
2.2 WORK FLOW IN CLUSTER ANALYSIS.....	13
2.3 ASSUMPTIONS IN CLUSTER ANALYSIS	15
2.4 CLUSTERS ANALYSIS AS MEASURE OF SIMILARITY.....	15
2.5 DISTANCE MEASURES.....	16
2.6 MINKOWSKI DISTANCE.....	17
2.6.1 <i>Euclidean Distance</i>	19
2.6.2 <i>Squared Euclidean Distance</i>	20
2.6.3 <i>Manhattan Distance</i>	21
2.6.4 <i>Chebyshev Distance</i>	22
2.7 MAHALANOBIS DISTANCE.....	24
2.8 SELECTING THE BEST DISTANCE MEASURE.....	25
2.9 CLUSTERING METHODS.....	26
2.9.1 <i>Partitioning Methods</i>	26
2.9.2 <i>Hierarchical Methods</i>	32
2.9.3 <i>Density Based Methods</i>	34
2.9.4 <i>Grid Based Methods</i>	36
2.9.5 <i>Fuzzy Clustering</i>	37
2.10 NUMBER OF CLUSTERS AND HETEROGENEITY MEASUREMENT.....	37
2.11 DATA STANDARDIZATION	38
2.12 CLUSTER VALIDATION	39
2.12.1 <i>External Criteria</i>	40

3 DATA ANALYSIS	43
3.1 ANALYSIS DESIGN.....	43
3.2 STUDY AREA AND DATA DESCRIPTION	45
3.3 DATA EXPLORATION	46
3.4 COMPUTATIONAL ANALYSIS	48
4 RESULT INTERPRETATION	50
4.1 CLUSTER MAP	50
4.2 CLUSTER VALIDATION	55
4.2.1 <i>Misclassification Matrix and Visual Analysis</i>	55
4.2.2 <i>Cluster Validation</i>	66
5 DISCUSSION.....	68
5.1 CHALLENGES	70
5.2 FUTURE RESEARCH	71
6 CONCLUSION	72
REFERENCES.....	74
APPENDIX 1	A
APPENDIX 2	B
APPENDIX 3	C

LIST OF ABBREVIATIONS

KDD	Knowledge Discovery in Database
CLARA	Clustering Large Application
CLARANS	Clustering Large Applications based on Randomized Search
AGNES	Agglomerative Nesting
DIANA	Divisive Analysis
DBSCAN	Density Based Spatial Clustering of Application with Noise
OPTICS	Ordering Points to Identify the Clustering Structure
DENCLUE	Density-based Clustering
STING	Statistical Information Grid-based method
DEM	Digital Elevation Model
PCP	Parallel Coordinate Plot

LIST OF FIGURES

Figure 1 Different steps in knowledge discovery process (University of Florida, 2015)	2
Figure 2 Different Steps in Cluster Analysis (Halkidi et al., 2001).....	13
Figure 3 Different forms of Minkowski distance.....	18
Figure 4 Unit circles with various values of 'p'	18
Figure 5 Euclidean distance between two points	19
Figure 6 The local pattern of a Manhattan network and real life examples of orthogonal streets of Manhattan and Barcelona (Dalfo et al., 2007)	21
Figure 7 Manhattan distance (red); equivalent Manhattan distance (yellow and blue) and Euclidean distance (green) between two points (Wiktionary, 2013)	21
Figure 8 Chebyshev distance between two points	22
Figure 9 Unit Circle representation of Chebyshev distance	23
Figure 10 Comparison between Euclidean distance and Mahalanobis distance (Maesschalck et al., 2000).....	24
Figure 11 K-Means clustering algorithm steps (Miller & Han, 2001).....	27
Figure 12 A dendrogram showing two distinct clusters (Manchester Metropoliton University,)	32
Figure 13 Agglomerative and Divisive Clustering on a set of data object (p,q,r,s,t) (Miller & Han, 2001)	34
Figure 14 DBSCAN method and cluster detection (Miller & Han, 2001)	35
Figure 15 Grid Based Clustering (Patentdocs, 2011)	36
Figure 16 Workflow of Analysis Process	43
Figure 17 Vegetation, Soil Type and Slope raster layers.....	46
Figure 18 Normalized Slope Layer.....	47
Figure 21 Normalized Soil Type Layer	48
Figure 22 Cluster map created using Euclidean distance	51
Figure 23 Cluster map created using squared Euclidean distance	51
Figure 24 Cluster map created using Manhattan distance	52
Figure 25 Cluster map created using Mahalanobis distance.....	52
Figure 26 Cluster map created using Chebyshev distance.....	53
Figure 27 Difference between clusters created using different distance measures. (a) Between Euclidean and squared Euclidean distance and (b) Between Euclidean and Manhattan distance.....	57

Figure 28 Difference between clusters created using different distance measures. (a) Between Euclidean and Mahalanobis distance and (b) Between Euclidean and Chebyshev distance.....	59
Figure 29 Difference between clusters created using squared Euclidean and Manhattan distance	60
Figure 30 Difference between clusters created using different distance measures. (a) Between squared Euclidean and Mahalanobis distance and (b) Between squared Euclidean and Chebyshev distance.....	62
Figure 31 Difference between clusters created using different distance measures. (a) Between Manhattan and Mahalanobis distance and (b) Between Manhattan and Chebyshev distance.....	64
Figure 32 Difference between clusters created by Mahalanobis and Chebyshev distance.....	65

LIST OF TABLES

Table 1 Misclassification Matrix of Euclidean and Squared Euclidean Distance	56
Table 2 Misclassification Matrix of Euclidean Distance and Manhattan Distance	57
Table 3 Misclassification Matrix of Euclidean Distance and Mahalanobis Distance..	58
Table 4 Misclassification Matrix of Euclidean Distance and Chebyshev Distance	58
Table 5 Misclassification Matrix of Squared Euclidean Distance and Manhattan Distance.....	60
Table 6 Misclassification Matrix of Squared Euclidean Distance and Mahalanobis Distance.....	61
Table 7 Misclassification Matrix of Squared Euclidean Distance and Chebyshev Distance.....	61
Table 8 Misclassification Matrix of Manhattan Distance and Mahalanobis Distance	63
Table 9 Misclassification Matrix of Manhattan Distance and Chebyshev Distance ...	63
Table 10 Misclassification Matrix of Mahalanobis Distance and Chebyshev Distance	64
Table 11 Comparison of external indices for different distance measures	66

1. Introduction

Recent advancements in data acquisition and storage technologies have resulted in growth of huge databases. This advancement ranges in different areas from credit card usage data, telephone call data, government statistics, astronomical data, molecular database as well as geographic databases (Hand et al., 2001). Research in the field of medicine, science and engineering are rapidly accumulating data that is key to important new discoveries. This progress has been induced by the fact that systems are often been used in different fields that we do not know in depth and need more information about them. This lack of knowledge should be compensated by the mass of the stored data that is available nowadays. The available data have induced the need to process and use it. The data reflects the behavior of the analyzed system; therefore there is a theoretical potential to obtain useful information and knowledge from the data (Abonyi & Feil, 2000). However, extracting useful information from available dataset is extremely challenging. Often, traditional data analysis methods, which are based on hypothesize-and-test paradigm, cannot be used because of size of data. Also, non-traditional nature of data means that traditional approaches cannot be applied even if the dataset is relatively small. Most of non-traditional methods are motivated by the desire to automate the process of hypothesis generation and its evaluation. Further, there can be situations where questions that need to be answered cannot be addressed using existing data analysis techniques thus, new methods are required to be used in order to extract useful information from huge datasets (Tan et al., 2006).

One of the areas of advancement in data acquisition is in the field of geography where advancement in data collection methods like photogrammetry and remote sensing has led to acquisition of huge amount of data. Thus, geography has moved towards data-rich and computation –rich environment. The scope, coverage, and volume of geographic datasets are rapidly growing. Geographical data are unique in nature due to special characteristics such as geographic measurement framework, spatial autocorrelation, heterogeneity, complexity of spatial objects and relationships

and diversity of data. So, it requires unique tools for analysis and provides unique research challenge. Formal and computational representation of the geographic information requires adoption of implied topological and geometric measurement framework, which affects measurement of geographic attributes and consequently the patterns that can be extracted. Thus, because of inductive nature and ability to handle heterogeneous datasets, data mining is appropriate tool for exploring geographical databases. (Miller & Han, 2001)

Data mining is the analysis of large observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. Data mining is often set in broader context of *Knowledge Discovery in Database* (KDD). The KDD process involves different stages: selecting the target data, preprocessing the data, transforming if necessary, data mining to extract patterns and relationships, and finally interpreting and assessing the discovered structures. (Hand et al., 2001)

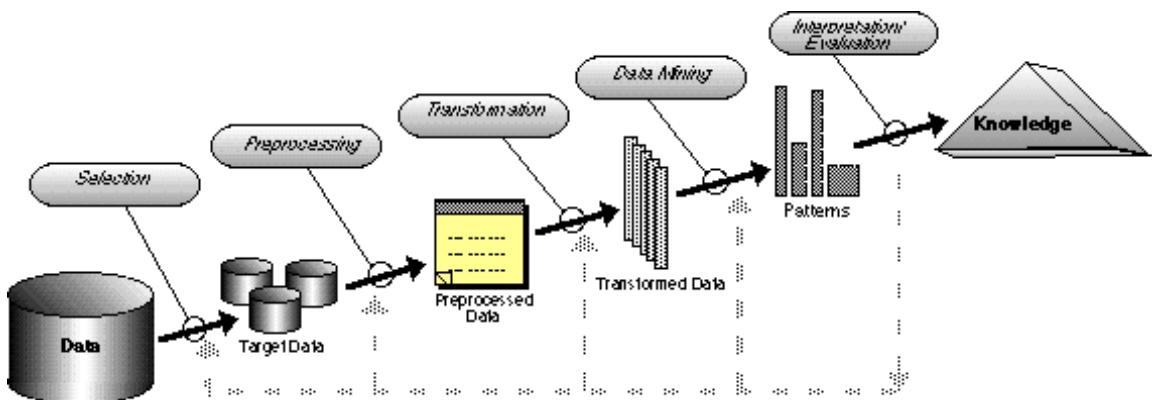


Figure 1 Different steps in knowledge discovery process (University of Florida, 2015)

The extraction of information is useful for understanding the overall knowledge discovery process. There are different possible ways to extract patterns and determine relationships in the dataset. Clustering is partition of a dataset into subsets so that the data in each subset shares some common trait. (Abonyi & Feil, 2000)

Clustering is one of the most primitive mental activities of humans, used to handle huge amount of information we receive everyday. Processing every piece of information as a single entity would be impossible. Thus, humans

tend to categorize entities into clusters where each cluster is characterized by common attributes of the entities it contains. (Theodoridis & Koutroumbas , 2003). Mostly cluster analysis reveals meaningful groups in data, revealing natural structure of the data. For understanding the dataset, cluster analysis has an important role in wide variety of fields like psychology, social sciences, biology, statistics, pattern recognition, information retrieval, geosciences and many more. (Tan et al., 2006)

Most of the geographical datasets are multivariate in nature and contains different attributes as well as geographic information. To reveal important patterns from such dataset is a challenging task as the patterns may have various forms (linear or nonlinear) and involve multiple spaces (e.g. multivariate space and geographic space). To be considered as a multivariate data, all the variables must be random and interrelated in such ways that their different effects cannot meaningfully be interpreted separately. Thus, multivariate analysis techniques are applied to such dataset to analyze multiple variables in a single relationship or a set of relationships. (Hair et al., 2006)

There are different forms of geographic data of which, point data is the simplest form. The location of a spatial object or an event is represented as a point. In the case of clustering, point data is generally used to calculate distance between data objects. Distance is one of the widely used measures in the process of spatial analysis. Especially, in clustering, to determine the clusters, distance is one of the important parameter. The goal of clustering is to find clusters from unlabeled data so that the data element that belongs to same clusters is as similar as possible whereas data belonging to different clusters are as dissimilar as possible. It is one of the main exploratory data analysis methods where similar items are grouped as close as possible. Distance measures can be then applied to compute the similarity between objects. (Miller & Han, 2001)

Similarity represents the degree of correspondence among objects across all of the characteristics used in the analysis. (Hair et al., 2006). Similarity enables each observation to be compared to each other and determine the objects that are similar to each other according to certain criteria. The term *proximity* is often used as a general term to denote either a measure of similarity or

dissimilarity (Hand et al., 2001). The term '*distance-similarity metaphor*' (Montello et al., 2003) is often used to relate distance measure with similarity and is reminiscent of Tobler's 'First Law of Geography' (Tobler, 1970). Tobler's first law of geography contends that one can predict the similarity of geographic features based on their distance to other features on the Earth's surface. Thus, distance is correlated with similarity, in most cases because distance determines similarity (Fabrikant & Montello, 2008).

The research will analyze different distance methods used for multivariate data specially focusing on using similarity measures in clustering and determine the corresponding effect on result.

1.1 Thesis Structure

The thesis consists of three distinct parts: introduction, theoretical background and data analysis.

Initially, background information about the cluster analysis and related works in field of cluster analysis is presented. Further, research question for analysis is purposed and methods and materials used in the research are defined.

In second part, theoretical background related to clustering methods along with its applications and validity measures are presented. This chapter incorporates all the concepts and methods used throughout the thesis. Furthermore, the application of clustering in this research is explained.

Further, third part, Data analysis, explores the characteristics of dataset and aims to understand the dataset. This chapter deals with formulation of analysis process, application of clustering method in the dataset along with result analysis and comparison between results.

Finally, conclusion and discussion along with future works in relation to this research is presented.

1.2 Related works

The notion of proximity is fundamental component of any comprehensive ontology of space (Worboys, 2001). The proximity of objects with a number of attributes is typically defined by combining the proximities of individual attribute (Tan et al., 2006). The attributes could be used to assess similarity of geographic events and process based on their spatiotemporal characteristics,

(Mcintosh & Yuan, 2005), to obtain optimal values of performance parameters, or to calculate class centroids for interpretation of similarities and differences between classes (Gorsevski et al., 2005). There has been a lot of research in the field of clustering and different similarity measures (Worboys, 2001; Fabrikant & Montello, 2008; Morales-Esteban et al., 2014; Pollard, 1981). However, there have been very limited comparisons between different distance measures used for clustering and its impact on result is still a subject of research. Most of the previous research is particularly focused on individual similarity measures (Fabrikant & Montello, 2008) and its application in clustering.

K-means is a very well known and relatively simple clustering method. K-means divides a set of data items into k clusters, where the number of clusters must be provided beforehand (Miller & Han, 2001). Each item belongs to the cluster with nearest mean. K-means has its use in different field like medical (Wilmer et al., 2008), spatial (Morales-Esteban et al., 2014; Borruso, 2008) and many other fields. Typically, K-means clustering method uses Euclidean distance to determine k number of clusters (Wilmer et al., 2008). There has been research on use of Mahalanobis distance (Morales-Esteban et al., 2014), as well as research on comparison between distance measures (Dong et al., 2013) in relation to optimization of parameters in k-means.

K-Medoids is another clustering method where, the dataset of n object is clustered with K number of clusters provided by the user. K-medoids method is the modified form of K-means method. Unlike K-means method, in K-medoids method, instead of calculating the mean values of the objects in a cluster as reference point, actual object from the data also called as medoid is selected to represent the cluster (Miller & Han, 2001). K-medoids has its application in computer science (Alarcon-Aquino et al., 2014), (Park & Jun, 2009), geo science (Ding et al., 2009), medicine (Zadegan et al., 2013) and other different fields. Kaufman and Rousseeuw, (1990), developed a new partitioning algorithm PAM (Partitioning Around Medoids) which used K-medoid method, to overcome the drawbacks of K-means method and create cluster with the help of medoid. Adnan et al., (2010) has presented comparison between efficiency of K-medoid method over K-means method in large multidimensional spatial data. Similar to K-means method, different

distance measures could be incorporated with K-medoid method. Jung et al., (2013) has used modified Hausdorff distance, a pattern based distance, with K-medoid clustering method for image-based scenario modeling of fractured reservoirs for flow uncertainty quantification. Also, Alarcon-Aquino et al., (2014) analyzed Minkowski distance with K-medoid method. However, comparison between different distance measures with K-medoid method is a subject of research.

One of the related applications of similarity used in this research is mobility analysis. Mobility analysis is the process of analyzing the off-road mobility of vehicle with the goal to create a map representing how difficult it is for a specified vehicle to advance over terrain. Mobility analysis has application in crisis management, military movement as well as other various areas. The goal of mobility analysis is to create a mobility map, which is a type of cost surface, where the value of each pixel represents the amount of resources required for specific activity at the location that depicts the maneuverability of the terrain in the operational area. To analyze the mobility, the cluster produced could be categorized into good or bad mobility by assigning a mobility value to each cluster which corresponds to similarity in location and hence, similarity measures can be used for solving the problem (Nikander et al., 2012; Nikander, 2012).

There are also many clustering methods used in field of Geoinformatics (Miller & Han, 2001; Theodoridis & Koutroumbas , 2003; Park & Jun, 2009; Pollard, 1981; Zhai et al., 2014; Kaufman & Rousseeuw, 1990). However, most of them are focused on using standard distance measure to create clusters.

In this research, the concept of similarity is determined with the help of attribute values of the object and is used to analyze and compare different similarity measures. This research focuses on clustering with use of different distance measures and evaluates different distances for determining similarity between data objects. Further, application of different distance measures on K-Medoids clustering and affect of similarity measures on a dataset is analyzed, the result of which can be used to analyze mobility on a given terrain.

1.3 Objectives of Research and Research Questions

The thesis is based on idea that knowing different distance measures and its use on clustering would provide good insight to result interpretation of the clustering and subsequently reveal interesting knowledge about the dataset.

The main research question answered by this research is “*What are different distance measures used in clustering?*” The research analyzes the different distance measures and provides answer to “*How the use of different distance measures affects the result of spatial analysis in clustering?*“ Further, the research will also provide insight into “*Which distance measures would provide the best result in case of clustering?*” Also, the research will provide an answer to “*Can a distance measure provide proper insight about similarity of the cluster and reveal useful information?*” The thesis aim to evaluate different distance measures by applying the K-medoid method to a dataset and explore the relative information revealed through each distance measures.

The objective of the research is to perform literature review about clustering methods that uses distance as important input parameter. This is followed by explanation and use of different distance measure in clustering. Further, different distance measures are then applied to the available dataset for evaluating difference/similarity and corresponding difference/similarity is documented and compared. Different distances to be used in research include *Mahalanobis distance* and *Minkowski distance*, whose extension includes *Euclidean distance*, *Squared Euclidean distance*, *Chebyshev distance* and *Manhattan Distance*.

This research is an extension into similar research on mobility analysis by Nikander et.al. (2012), performed in Aalto University and provides insight on use of different distance measures for clustering which is further used for mobility analysis of given terrain.

In addition, the research has similar limitations as research performed by Nikander et.al, (2012). Here, the research is limited to spatial problems where input data can be transformed into a format where there are no explicit spatial dependencies between locations. Thus, the knowledge and information about spatial correlation between layers, as well as spatial autocorrelation between locations is not explicitly inserted into the process. If such knowledge is

required, other computational methods or user knowledge is used to analyze these phenomena (Nikander et al., 2012). Also, there are different clustering methods (Hair et al., 2006; Miller & Han, 2001) and algorithms (Theodoridis & Koutroumbas , 2003) that could be used for the given dataset. This research is limited to use of K-medoid method to determine the clusters.

1.4 Mobility Analysis

Travelling is intrinsic part of human society. Advancement in technology, has allowed us to plan and to analyze how to travel more efficiently. With increase in digital spatial data, their use varies from consumer applications like online maps to find best routes to complex analysis like in the military or forest industry to plan how to move outside road. So, analysis of problems related to vehicle mobility has different application areas. Vehicle mobility is the capability of a vehicle to move between locations and is dependent on both vehicle and environment it is moving through. Thus, mobility analysis is a spatial analysis problem concerned with the movement of vehicles between the locations. Mobility problems include computing the best route between locations by calculating a measure of how easy it would be for a vehicle to move through a location in the target area.

Vehicle mobility can be divided into two categories: on-road and off-road mobility. On-road mobility of vehicle is limited by road type, traffic, and maximum speed of vehicle whereas off-road mobility is limited by the ability of a vehicle to travel in rough terrain, soil type, slope, amount and type of vegetation in the given terrain.

Off road mobility is important in fields such as military (Nikander et al., 2012) and forestry. Off-road mobility has its typical application in crisis management, which is linked to military movement. In military application, the problem area can often be large and large part of the route may be traversed outside the existing road network. Thus, the route selected requires a terrain that is trafficable even after passage of several vehicles and the route must be such that all relevant vehicles are able to traverse over it. Also, routes wide enough for several vehicles to move in a row may be of interest. Damage to the terrain caused by the passage of vehicles may be an issue, depending on the situation. (Davis et al., 1991).

Vehicle mobility is modeled using *mobility map*, which is a type of cost surface, where the value of each pixel represents the amount of resources required for specific activity at that location that depicts the maneuverability of the terrain in the operational area. (Nikander et al., 2012) The specific activity could be movement of troops between locations, rescue in case of emergency situations or a training scenario for military.

The application of clustering in this study is to analyze the terrain based on mobility of specific vehicles and representing how difficult it is for a specified vehicle to advance over terrain. To analyze the mobility, the cluster produced could be categorized into good or bad mobility by assigning a mobility value to each cluster. By assigning each cluster a mobility value, it could be used as mobility map. This division into mobility categories is an example of *suitability problem* where the goal is to find a location best suited for a given activity, or to categorize locations according to their suitability. Here, assigning different categories to different location depends on the input, type of vehicle and different other consideration however, all locations belonging to a category have similar overall suitability scores. This corresponds to similarity in location and hence, similarity measures can be used for solving the problem. For multivariate data, similarity is calculated as a distance in multi-dimensional space. So the suitability problem can be solved by combining similar locations into classes and giving each class a suitability value.

The goal of cluster analysis is to see whether the data can be divided into natural subsets, which are clearly distinct from each other. In case of mobility analysis, the goals could be to visualize whether there is a subclass of good mobility that is clearly distinct of classes of bad mobility, to see whether there is a subclass of fair mobility, and what are the differences between these; to see whether there are clearly distinct subclass of bad mobility, and what prevents mobility in these classes. Further, clustering does not directly solve the suitability problem. The result of clustering is a class or a cluster representing a set of similar data items. These clusters need to be categorized according to the suitability of the items in the cluster. Thus, the clustering result needs to be interpreted in order to reveal useful information from it. (Nikander, 2012; Nikander et al., 2012)

1.5 Methods and materials

The thesis purposes different distance measures used to determine similarity in case of cluster analysis and analyze different distance measures in relation to K-medoid clustering method. The clustering method is applied to analyze the similar areas in given terrain for the purpose of mobility analysis.

The dataset used in the research includes slope information, total cross-sectional areas of trees from 1 to 3 meters in height and soil type of the given terrain. Using the above datasets and K-medoid clustering method, cluster map is created which, is used to compare between different distance measures.

The following software have been used in this thesis:

- **Matlab R2013b** for clustering and computational tasks,
- **ArcGIS 10.1** for visualization and computational analysis,
- **R** for cluster validation.

2 Cluster Analysis

In this chapter different related concepts and methods used throughout the thesis are presented. It covers the theories related to cluster analysis, similarity measures, clustering methods, cluster validation and its application on mobility analysis on a terrain is explained. The chapter provides the insight to the methods that are used in the analysis process and facilitates the interpretation of process and results, contributing to better understanding of result.

Cluster analysis is a multivariate data analysis technique whose primary purpose is to group objects based on characteristics they possess. It classifies objects so that each object is similar to other in the cluster based on a set of selected characteristics (Everitt, 2011). The resulting clusters should exhibit high internal homogeneity within a cluster and high external heterogeneity between clusters (Hair et al., 2006). Cluster analysis is a tool of discovery, which can be used to reveal association, and structures in data, which, though not previously conceived, are nevertheless sensible and useful when found. The result can contribute to the development of a classification method; they may suggest general models to describe other samples and ultimately the parent population; or they may simply provide definitions of size and measures of change in what previously were notional categories (Backer, 1995). Cluster analysis is thus concerned with exploring the dataset and generalizing it meaningfully with small number of clusters of individual observation that represent general characteristics of the group of data.

Cluster analysis is also referred as data segmentation in some applications as it partitions large dataset into groups according to similarity. Cluster analysis can be used to gain insight to data distribution, observe characteristics of each cluster and focus on particular set of clusters for further analysis. Further, it may also be used as a preprocessing step for other algorithms such as characterization, attribute subset selection and classification, which would then operate on the detected clusters and selected attributes or features (Miller & Han, 2001).

Cluster analysis could be regarded as a form of a classification as it creates a labeling of objects with class (cluster) labels (Tan et al., 2006). Thus, if

classification is successful, the objects within clusters will be close together when plotted geometrically and different clusters will be far apart. In cluster analysis, concept of variate is central issue. The *cluster variate* is a set of variables representing the characteristics used to compare the objects in cluster analysis. As the cluster variate includes only the variables to compare object, it determines the character of the objects (Hair et al., 2006).

2.1 Application and Objective of Cluster Analysis

Cluster analysis has wide applications including market research, pattern recognition, data analysis and image processing (Theodoridis & Koutroumbas , 2003). Apart from this, there are different application areas where cluster analysis method could be applied. Here, few basic directions are determined where clustering could be used in general.

- Data reduction

In most of the cases, the data available is very large hence the simplification of data is very demanding and time consuming. Cluster analysis could be used to group the data into number of “sensible” clusters and each cluster could be processed as a single entity.

- Hypothesis generation

Here, the cluster analysis is applied to a data set to infer some hypotheses concerning the nature of the data. Cluster analysis could be used to suggest certain hypotheses, which are further verified using other datasets.

- Hypothesis testing

In this context, cluster analysis is applied for the verification of a specific hypothesis. For a given set of problem defined by different variables, clustering method could be used to group similar set of variables that could be used to verify the given hypothesis (Theodoridis & Koutroumbas , 2003). With a cluster, it is possible to reveal the relationship among the observations, which is typically not possible to obtain with individual observations. The simplified structure from cluster portrays relationships not revealed otherwise. (Hair et al., 2006)

- Prediction based on groups

Cluster analysis is applied to the available dataset and the resulting clusters are characterized based on the characteristics of the patterns by which they

are formed. For an unknown pattern, the corresponding cluster could be determined and could be characterized based on characterization of respective cluster (Theodoridis & Koutroumbas , 2003).

Selection of clustering variable is one important objective of analysis. Whether the objective is exploratory or confirmatory, the possible results are effectively constrained by selection of variables. The derived clusters reflect the inherent structure of the data and are defined only by the variables. Thus, selection of variable is done with regard to theoretical and conceptual as well as practical considerations (Hair et al., 2006)

2.2 Work Flow in cluster Analysis

Cluster analysis is the supervised learning process where all patterns are represented in terms of features (Theodoridis & Koutroumbas , 2003)

Figure 2 provides different steps of clustering and how clustering can be used for knowledge discovery from a data.

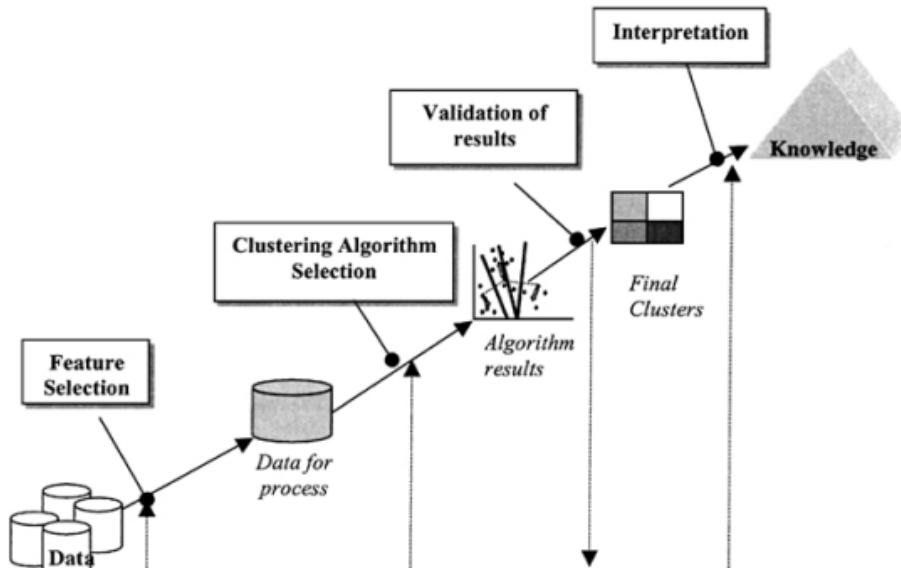


Figure 2 Different Steps in Cluster Analysis (Halkidi et al., 2001)

The basic steps that must be followed to perform cluster analysis are as follows:

1. Feature Selection

Features must be selected properly in order to encode as much information as possible concerning the task of interest. The major goal is minimum information redundancy among the features. As in supervised classification,

preprocessing of features may be necessary prior to their utilization in subsequent stages.

2. Proximity Measure

Proximity measure quantifies how “similar” or “dissimilar” two features are. It ensures that all selected features contribute equally to the computation of the proximity measure and there are no features dominating other features. Thus, selection of proximity measure is one important step in cluster analysis. This research focuses on evaluation of different proximity measures.

3. Clustering Criterion

Clustering criterion depends on the interpretation of the expert as “sensible” based on the type of clusters that are expected to underlie the dataset. The clustering criterion may be expressed as cost function or some type of rule depending on the dataset.

4. Clustering Algorithms

Here, a specific algorithm is selected that reveals the clustering structure of data set based on previously selected proximity measure and clustering criterion.

5. Validation of Results

The clustering algorithm provides the result of dataset as clusters, which needs to be verified using appropriate tests.

6. Interpretation of the Results

The resulting clusters must integrate the results of clustering with other experimental evidence and analysis in order to draw right conclusions. In most of the case, expert in application field is required to integrate the result. Further, in some cases, *Clustering Tendency* should be involved which includes various test that indicate whether or not the available data possess a clustering structure. This step is particularly important in case of completely random data where trying to find a cluster would be meaningless. The choice of features, proximity measures, clustering criteria and clustering algorithm is important as they may lead to totally different clustering results (Theodoridis & Koutroumbas , 2003).

2.3 Assumptions in Cluster Analysis

Cluster Analysis is a method for quantifying the structural characteristics of a set of observations and has strong mathematical properties but does not have strong statistical foundation. Thus, there are certain assumptions to be made with respect to variables in the cluster variate.

One of the assumptions is representativeness of the sample. Usually, a sample case is obtained for clustering rather than the whole census data. Thus, it is assumed that the given sample of observation is the true representation of the population and the results are general to the population of interest.

Another assumption in cluster analysis is the impact of multicollinearity. Multicollinearity is the statistical phenomena where two or more variables are strongly correlated. In case of cluster analysis, the effect of multicollinearity is the form of implicit weighing and acts as a weighting process, which is not apparent to the observer but affecting the analysis. For example, when there are many variables in a dataset, and multicollinearity is examined with two sets of variables where one dataset has more variables than other. The effect on similarity measure would be large with dataset containing more variables. This is due to fact that each variable is weighted equally in cluster analysis. Thus, it is suggested to examine the variables used in cluster analysis for substantial multicollinearity and if present, either the number of variables is reduced to equal numbers in each set or one of distance measures such as Mahalanobis distance that compensates for the correlation is used (Hair et al., 2006).

2.4 Clusters analysis as measure of similarity

Similarity represents the degree of correspondence among objects across all of the characteristics used in the analysis (Hair et al., 2006). The concept of similarity is fundamental to cluster analysis and inter-object similarity is an empirical measure of correspondence, or resemblance between objects to be clustered (Hair et al., 2006).

Similarity provides the concept of proximity and presents how ‘close’ the observations are to each other or how ‘far’ are the observations. Similarity measures are most commonly used for the dataset with categorical variables.

The measures are generally scaled to be in the interval [0, 1], although occasionally they are expressed as percentages in the range 0–100%. Two individuals i and j have a similarity coefficient s_{ij} of 1 if both have identical values for all variables. A similarity value of 0 indicates that the two individuals differ maximally for all variables. (Everitt, 2011)

The input for a clustering method is a set of data vectors, where each data vector is a multidimensional set of data elements. The data vectors are compared for similarity and similar vectors are combined into clusters. Similarity in clustering is measured using a function, which takes two data vectors as input and returns a similarity value for them (Nikander, 2012). Similarity measure could be applied to identify the outliers in dataset, which could be observations with large distance from all other observations, or appear in cluster as single member or a small cluster (Hair et al., 2006).

There are two ways of obtaining measures of similarity. First method is to directly obtain the similarity about the objects like by market survey or food testing experiment. Alternatively, similarity can be obtained indirectly from vectors of measurements or characteristics describing each object. However, in second case, it is necessary to define the idea of ‘similar’ in order to calculate formal similarity measure. (Hand et al., 2001)

For measurement of similarity, three measures could be used namely, correlational measure, distance measure and association measure. Correlation measure uses correlation coefficient between the variables where higher correlation indicates similarity and lower correlation indicates lack of similarity. Association measure is used to compare objects whose characteristics are measured only in nonmetric term. (Hair et al., 2006)

In this research, distance measure used to measure the similarity and is described in section 2.5.

2.5 Distance measures

Distance is the most common similarity measure used in case of cluster analysis. It is not entirely clear how a ‘cluster’ is recognized when displayed in the plane, but one feature of the recognition process would appear to involve the assessment of relative distances between points. The distance measures represent similarity as the proximity of observations to one another across the

variables in cluster variate. Distance measures are actually a measure of dissimilarity for continuous variables (Everitt, 2011), where a larger value denotes less similarity and is converted into a similarity measure by using an inverse relationship. The distance measure best represents the concept of proximity as it focuses on the magnitude of the values and portrays similar cases of the objects that are close together. As the characteristics measured by metric variables are used, distance measure is the best method to assess similarity in clusters. (Hair et al., 2006)

The different distance measures used in this research are explained in section 2.6 and 2.7.

2.6 Minkowski distance

Minkowski distance is the generalized form of different distance measures like Euclidean distance, Manhattan distance, Chebyshev distance and Hamming distance. Minkowski distance of order P is defined as:

$$d(g_1, g_2) = \left(\int_Y |g_1(x) - g_2(x)|^p dx \right)^{\frac{1}{p}}, \quad p \geq 1$$

Where, $g_1(x)$ and $g_2(x)$ are functions of x and Y is the range of the integration. If Y is an index set, $Y=\{1,2,\dots,n\}$ and $g_1(x)$ and $g_2(x)$ are real then, above equation can be written as:

$$d(g_1, g_2) = \left(\sum_{i=1}^n |g_1^i - g_2^i|^p \right)^{\frac{1}{p}}$$

Which is the distance between two points $(g_1^1, g_1^2, \dots, g_1^n)$ and $(g_2^1, g_2^2, \dots, g_2^n)$ in R^n . Alternatively, if $X=[a, b]$ and $g_1(x)$ and $g_2(x)$ are L_P integrable which corresponds to: $\int_a^b |g_1(x)|^p |dx| < \infty$ and, $\int_a^b |g_2(x)|^p |dx| < \infty$, then

$$d(g_1, g_2) = \left(\int_a^b |g_1(x) - g_2(x)|^p dx \right)^{\frac{1}{p}}$$

is the L_P distance between functions $g_1(x)$ and $g_2(x)$.

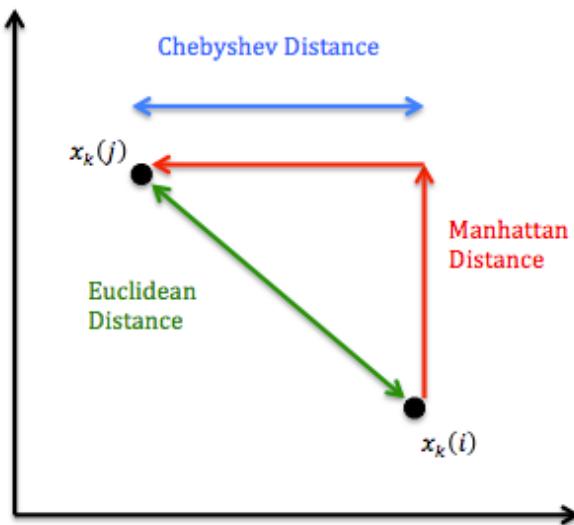


Figure 3 Different forms of Minkowski distance

Minkowski distance is often used when variables are measured in ratio scales with absolute zero value. The Minkowski distance reduces to the rectilinear distance, Euclidean distance and Chebyshev distance when the order p equals to 1, 2 and ∞ , respectively (Figure 3). The Minkowski distance is a general formula where $p=1$ results in Manhattan distance, $p=2$ results in Euclidean distance and $p = \infty$ results in Chebyshev distance (Zhai et al., 2014). The different forms of Minkowski distance between two points is represented in figure 3 and explained in this section.

The value of ‘ p ’ in Minkowski distance has its effect on type of clusters a distance measure produces. Figure 4 shows unit circles with various values of ‘ p ’ and corresponding shape of clusters produced as a result of ‘ p ’.

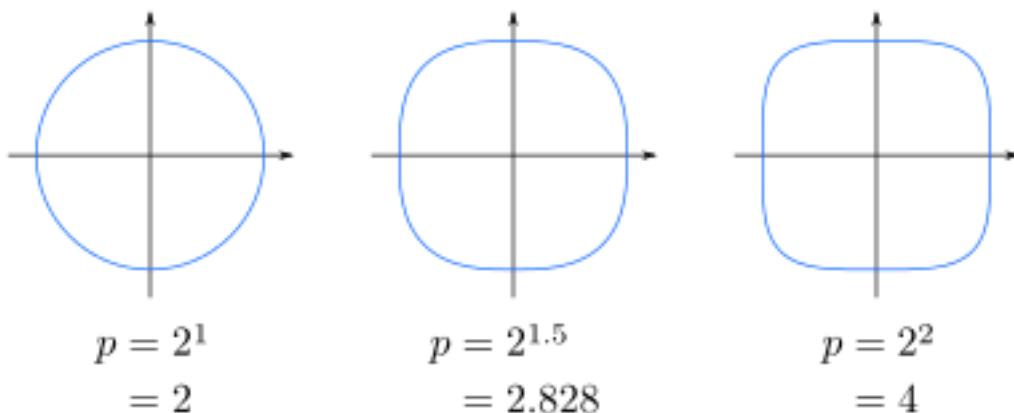


Figure 4 Unit circles with various values of 'p'

For $p = 2$, assumes circular cluster shapes, $p = 1$ assumes cluster in shape of square in two dimensions or diamond like in three or more dimensions and $p = \infty$ assumes clusters in the form of a box with sides parallel to the axes. Also, $p = \infty$ and $p = 1$ could be particularly useful in cases where the data structures have shapes with sharp edges. (Groenen & Jajuga, 2001). For practical application, it is intuitive to use $p = 1$, $p = 2$, or $p = \infty$ for Minkowski distance (Zhai et al., 2014). Thus, these three different variations of Minkowski distance are used as three different distance measures in this research.

For a dataset with n data objects with p real valued measurements on each object, the vector of observations for the i^{th} object is denoted by $x(i) = (x_1(i), x_2(i), \dots, x_p(i))$, $1 \leq i \leq n$, where the value of the k^{th} variable for the i^{th} object is $x_k(i)$. The different distance measures are defined below.

2.6.1 Euclidean Distance

This is the most commonly used distance between two points. It is simply the geometric distance in the multidimensional space and is extension to Pythagoras theorem.

Then the Euclidean distance between i^{th} and j^{th} object is defined as:

$$d(i, j) = \left(\sum_{k=1}^n (x_k(i) - x_k(j))^2 \right)^{\frac{1}{2}}$$

In simpler terms, Euclidean distance is the shortest distance between two given points as represented in figure 5.

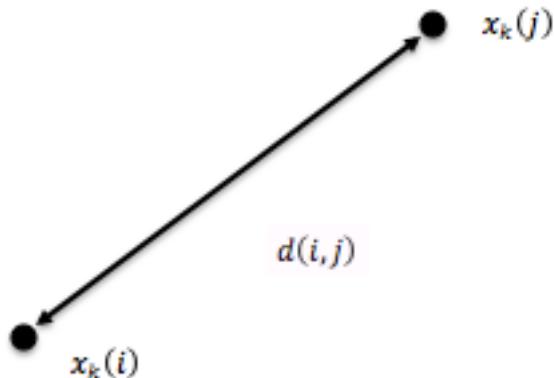


Figure 5 Euclidean distance between two points

Euclidean distance can be interpreted as a physical distance between two p -dimensional points in Euclidean space. The Euclidean distance between two vectors takes its minimum value $d_0 = 0$ when the vectors coincide. (Theodoridis & Koutroumbas , 2003)

This measure assumes some degree of commensurability between different variables. Thus, it would be effective if each variable were measured using same unit. Since the dataset often has non-commensurate variables, the arbitrariness of choice of unit must be overcome. A common strategy is to standardize the data by dividing each of the variables by its sample standard deviation so they are all regarded as equally important. Alternatively, data could be standardized by taking into account the covariance between the variables. The Euclidean distance is additive in the sense that the variables contribute independently to the measure of distance. (Hand et al., 2001)

2.6.2 Squared Euclidean Distance

Squared Euclidean distance is the sum of the squared differences without taking the square root. The squared Euclidean distance has the advantage of not having to take the square root, which speeds the computations markedly (Hair et al., 2006). The squared Euclidean distance is used more often than Euclidean distance to place progressively greater weight on objects that are further apart (Sage Publications, 2008). The values are calculated for each object pair by summing the squared difference between the observations.

The Squared Euclidean distance between i^{th} and j^{th} object is defined as:

$$d(i,j) = \sum_{k=1}^n (x_k(i) - x_k(j))^2$$

Square Euclidean distance is not a metric, as it does not satisfy the triangle inequality, however it is used in problems in which only distance have to be compared. Generally, Squared Euclidean distance and Euclidean distance is computed from raw data and not from standardized data (Sage Publications, 2008). It has wide spread use among researchers in the social and behavioral sciences. (Gore Jr., 2000)

2.6.3 Manhattan Distance

Manhattan distance is based on Manhattan network which is a unidirectional regular mesh structure resembling locally the topology of the avenues and streets of Manhattan (Dalfo et al., 2007). Manhattan distance can be defined as distance between two points in Euclidean space with fixed Cartesian coordinate system. It is the sum of the lengths of the projections of the segment between the points into the coordinate axes (Wikia, 2013).

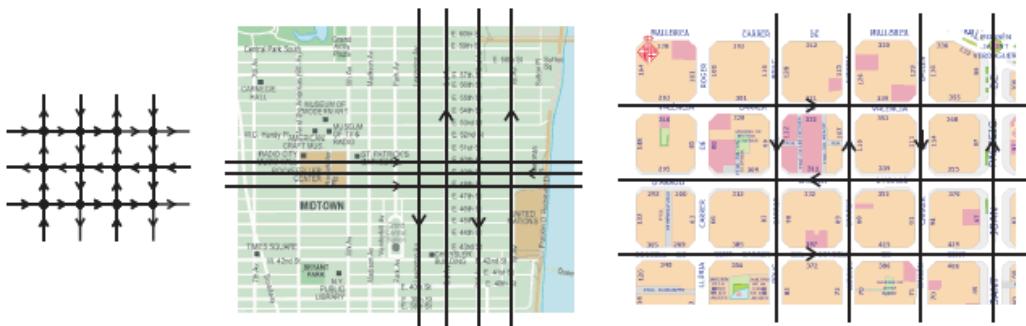


Figure 6 The local pattern of a Manhattan network and real life examples of orthogonal streets of Manhattan and Barcelona (Dalfo et al., 2007)

Manhattan distance is the distance between two points measured along axes at right angles and is often referred as city block distance as it measures distances travelled in street configuration. The Manhattan distance between i^{th} and j^{th} object is defined as:

$$d(i,j) = \sum_{k=1}^n |x_k(i) - x_k(j)|$$

In simpler terms, Manhattan distance is the sum of horizontal and vertical components between the given points in a plane.

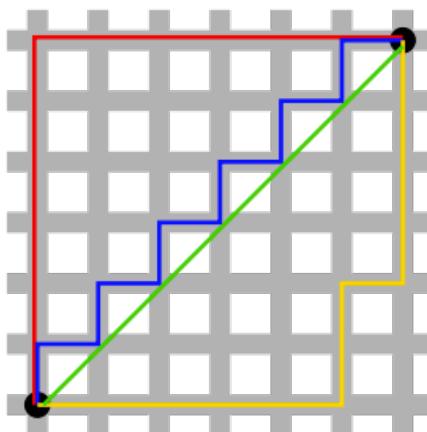


Figure 7 Manhattan distance (red); equivalent Manhattan distance (yellow and blue) and Euclidean distance (green) between two points (Wiktionary, 2013)

In figure above, the path represented by the red, blue or yellow lines, which is to be followed to reach from the point of origin to destination point in a Manhattan network is the Manhattan distance. It is also known as rectilinear distance, L_1 distance or l_1 norm, city block distance or Manhattan length.

Manhattan distance depends on the choice of the rotation of the coordinate system, but does not depend on the translation of the coordinate system or its reflection with respect to a coordinate axis (Wikia, 2013). It uses the sum of absolute differences of the variables and is simple to calculate but may lead to invalid clusters if the clustering variables are highly correlated (Hair et al., 2006).

2.6.4 Chebyshev Distance

The Chebyshev distance is one extreme case of Minkowski distance where $p = \infty$ is used that makes the distance equal to the single largest attribute value difference (Cichosz, 2015). The Chebyshev distance is also known as maximum value distance and calculates absolute magnitude between values of two objects. It is appropriate in cases when two objects are to be defined as “different” if they are different in any one dimension (University of Texas, 2000).

Chebyshev distance is a metric defined on a vector space where distance between two vectors is the greatest of their difference along any coordinate dimension.

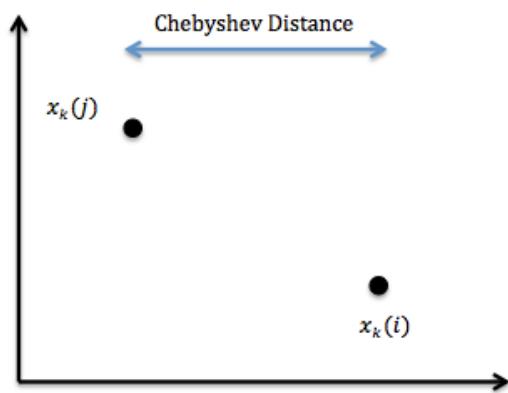


Figure 8 Chebyshev distance between two points

The Chebyshev distance (Figure 8) represents the distance along the largest dimension between two points. The Chebyshev distance are piecewise linear and in process of clustering, it ensures that the next considered points are

potentially located at the border of neighborhood of point in one dimension, and these point usually discover an unexplored area of the search space. (Dillmann et al., 2010)

Then the Chebyshev distance between i^{th} and j^{th} object is defined as:

$$d(i, j) = \lim_{p \rightarrow \infty} \left(\sum_{k=1}^n (x_k(i) - x_k(j))^p \right)^{\frac{1}{p}} = \max |x_k(i) - x_k(j)|$$

The Chebyshev distance is also known as Chess distance and is the distance between squares, in terms of move necessary for a King to go from one square to another. The circle of radius r , in Chebyshev metric is a square with side of length $2r$ parallel Euclidean distance (Agarawal & Sahoo, 2008).

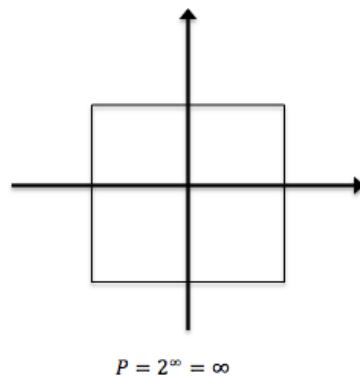


Figure 9 Unit Circle representation of Chebyshev distance

Chebyshev distance is often used in cases where the execution speed is so critical that the time involved in calculating the Euclidean distance is unacceptable. The contour lines of equal Chebyshev distance from a point are squares in two dimensions (Webb, 1999). It reduces the unit circle to a square with sharp edges (figure 9) hence, is useful to determine clusters with sharp edges. The major advantage of the Chebyshev distance is that it requires less time to decide the distances between the datasets. However, with the Chebyshev distance, one single feature is allowed to represent a dataset. This one single largest feature might not offer enough description of the dataset to lead to accurate neighborhood selection and final predictions (Filipe & Cordeiro, 2011). Thus, there might be case where Chebyshev distance could favor one dataset over another when different dataset are combined for clustering purpose.

2.7 Mahalanobis Distance

Mahalanobis distance is a generalized distance that accounts for the correlation among variables in a way that weights each variable equally. It also relies on standardized variables (Hair et al., 2006). A dataset could be standardized using covariance between the variables. The covariance between variable X and Y is calculated as:

$$Cov(X, Y) = \frac{1}{n} \sum_{i=1}^n (x(i) - \bar{x})(y(i) - \bar{y})$$

Where, \bar{x} is mean of X values and \bar{y} is the mean of Y values.

When the covariance matrix is incorporated in definition of distance, Mahalanobis distance between two p -dimensional measurements $x(i)$ and $x(j)$ is obtained which is defined as:

$$d_{MH}(i, j) = \left((x(i) - x(j))^T \sum^{-1} (x(i) - x(j)) \right)^{\frac{1}{2}}$$

Where, T represents the transpose, Σ is the $p \times p$ sample covariance matrix, and Σ^{-1} standardizes the data relative to Σ . (Hand et al., 2001)

In simpler terms, Mahalanobis distance is distance between a point and a distribution of data (Mahalanobis, 1936). It measures how many standard deviations away a point is from the mean of distribution. As represented in Figure 10 (plot b) below, when the point is in the mean of distribution of data, Mahalanobis distance is zero and as the point moves away from the mean of distribution of data, Mahalanobis distance increases.

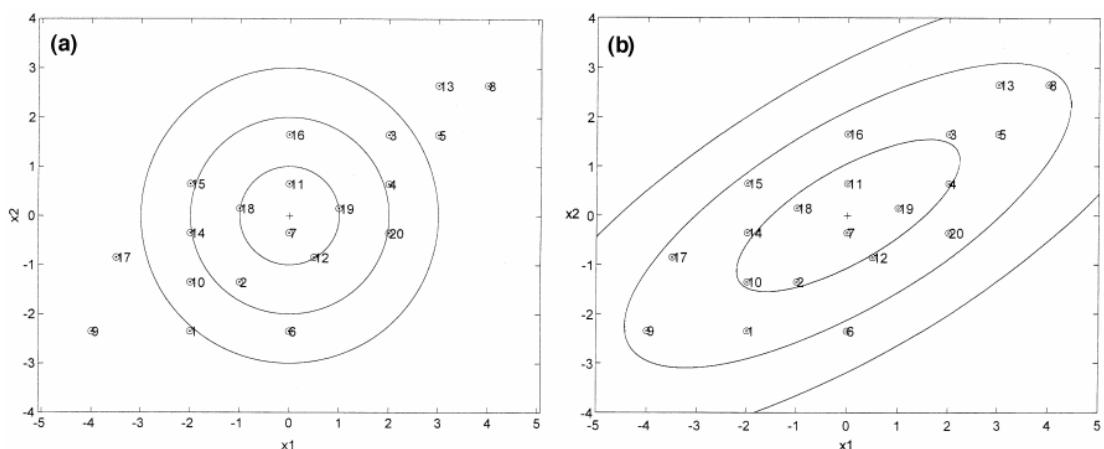


Figure 10 Comparison between Euclidean distance and Mahalanobis distance (Maesschalck et al., 2000)

Since, it is intuitive to understand Euclidean distance, figure 10 provides the comparison between Euclidean distance and Mahalanobis distance. In figure, part (a) is the plot of the simulated data for two variables x_1 and x_2 together with the circles representing equal Euclidean distance towards the center point. Part (b) is the plot of the simulated data for two variables x_1 and x_2 together with the ellipses representing equal Mahalanobis distance towards the center point. When Euclidean distance is used, the set of point equidistant from a given location is a sphere whereas; the Mahalanobis distance stretches the sphere to correct for the respective scales of the different variables and to account for the correctional among the variables providing ellipsoidal shape (Manly, 1986). As Mahalanobis distance takes covariance as well as direction of covariance of data into account, distance varies according to spread of the data. (Maesschalck et al., 2000).

The Mahalanobis distance increases with increasing distances between the two group centers and with decreasing within-group variation. Mahalanobis distance takes account of shape of the clusters by employing within-group correlation. (Everitt, 2011)

2.8 Selecting the best distance measure

As there are different distance measures for analyzing similarity, the ideal question rises about the selection of the best measures. However, selection of best distance measure is not straightforward and rather depends on different factors of the observed dataset. Everitt (2011) mentioned the influence of nature of data on choice of proximity measure. Also, the choice of measure should depend on scale of data as they provide different cluster solutions. For continuous data, distance or correlation-type dissimilarity measures should be used. Further, the clustering method used might have certain implications for the choice of parameters.

Hair (2006) has purposed few issues to be taken into consideration while selecting the best distance measure. As mentioned earlier by Everitt (2011) change in scale of variables may lead to different cluster solutions and comparing the result with theoretical or known pattern provides better solution to given problem. When the variables are correlated, Mahalanobis distance

measure is likely to be most appropriate as it adjusts for correlation and weights all variable equally (Hair et al., 2006).

2.9 Clustering Methods

There are different methods in determining and describing a cluster. However, different methods or even a same method with different parameter configurations can produce different clustering result. Clustering methods differ in many ways including definition of distance between data items, definition of ‘cluster’, the strategy to group or divide data items into clusters and, the data type that can be analyzed (numerical, categorical) and application-specific context and constraints (Miller & Han, 2001).

In general, the clustering method can be categorized into following five categories

1. Partitioning Methods
2. Hierarchical Methods
3. Density Based Methods
4. Grid Based Methods
5. Fuzzy Clustering

2.9.1 Partitioning Methods

For the dataset of n objects, partitioning methods creates the K number of partition for the given dataset where each partition corresponds to a cluster. In this method, each partition should contain at least one object and each object should only belong to one partition. The clusters are formed to optimize an objective-partitioning criterion, such as a dissimilarity function based on distance. It creates an initial partitioning and uses iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. (Miller & Han, 2001)

There are different partitioning methods, two of which are described below.

K-means

K-means is a typical partitioning method where the given dataset is partitioned into K number of clusters. In k-means method, cluster similarity is measured with respect to the mean value of the objects in a cluster, which can be viewed as the cluster’s centroid or center of gravity.

K-means algorithm is used to determine the clusters. In the algorithm, it randomly selects K of the objects, each of which initially represents a cluster center. For each remaining objects, an object is assigned to the cluster to which it is the most similar based on the distance between the object and the cluster mean. It then computes the new mean for each cluster. The algorithm iterates until the center of the clusters does not change, which corresponds to convergence of criterion function. The main aim of K-means clustering is the optimization of the objective function based on input parameters. The algorithm attempts to determine k partitions that minimize the objective function used, which is defined by square-error criterion.

The square-error criterion is used as stoppage criteria of the algorithm, which is defined as,

$$E = \sum_{i=1}^k \sum_{p \in C_i} |p - m_i|^2$$

Where, E is the sum of square-error for all the objects in the dataset; P is the point representing a given object and m_i is the mean of cluster C_i .

Here, for each object in each cluster, the distance from the object to its cluster center is calculated, and the distances are summed up. This criterion tries to make the resulting K clusters as compact and as separate as possible. The criterion function attempts to minimize the distance of each point from the center of the cluster to which the point belongs.

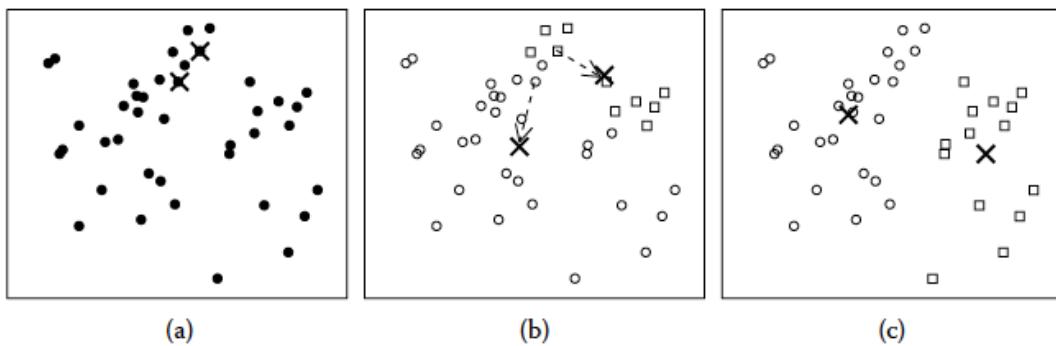


Figure 11 K-Means clustering algorithm steps (Miller & Han, 2001)

Figure 11 represents clustering process for the k-means algorithm with two clusters. In figure 11 (part a), for the dataset, two random cluster centers are selected (marked with 'x'). Further, when the points are assigned to a cluster based on its distance to cluster centers, the cluster centers starts to move

from its initial location as represented in part b. Finally, when there are no more points remaining to assign to the cluster, the center of cluster does not change and final solution, containing two distinct clusters and two cluster centers are obtained as represented in part c. The pseudo code for K-means algorithm is presented in Appendix 1.

In K-means, random initialization of centroids is used, thus, different runs of K-means can produce different clusters. Selecting the proper initial centroids is the key step of the basic K-means method. Since, the initial centroids are selected randomly, it might not be possible to replicate exact cluster in different runs of algorithm. Thus, to solve the problem, one effective approach is to take a sample of points and cluster them using a hierarchical clustering technique. From hierarchical clustering, K clusters are extracted and centroids of those clusters are used as initial centroids for K-means clustering. Another approach is, by selecting first point at random or by taking the centroid of all points. Then for each successive initial centroid, the point that is farthest from any of the initial centroid is selected. By this approach, the initial centroids are guaranteed not only to be randomly selected but also well separated. But this approach can select outliers, rather than points in cluster and also it is expensive to compute the farthest point from the current set of initial centroids. Thus to overcome these problems, this approach is applied to sample of the points as outliers are rare, they tend not to show up in a random sample. (Tan et al., 2006)

The algorithm works well when the clusters are compact clouds that are well separated from one another. The method is relatively scalable and efficient in processing large dataset because the computational complexity of the algorithm is $O(nkt)$, where n is the total number of objects, k is number of clusters, and t is the number of iterations. The method terminates at a local optimum. The output of algorithm produces clusters and cluster center is represented by the mean value of the objects in the cluster.

Although the algorithm partitions the given dataset into desired number of clusters, specifying the number of clusters in advance can be a disadvantage. Further, the method is not suitable for discovering clusters with non-convex shapes of clusters of different size and is sensitive to noise and outlier data

points because a small number of such data can substantially influence mean value. (Miller & Han, 2001)

The output of K-Means clustering is not affected if Euclidean distance is replaced with Euclidean squared. However, the output of hierarchical clustering is likely to change.

K-Medoids

K-Medoids is another clustering method where, the dataset of n object is clustered with K number of clusters provided by the user. The K-medoids method works on principle of minimizing the dissimilarities between each and every object on the dataset. (Sood & Bansal, 2013)

K-means is sensitive to outliers as an object with extremely large value may substantially distort the distribution of data, which is exacerbated due to the use of square-error criteria. K-medoids method is the modified form of K-means method to diminish the sensitivity to outlier. Unlike K-means method, in k-medoids method, instead of calculating the mean values of the objects in a cluster as reference point, actual object from the data is selected to represent the cluster center. This point is called *medoid* or a *representative object*. The medoid is the most centrally located object within the cluster. In K-medoid method, not every selection of K representative object creates “good” clustering. The clue for obtaining good cluster is to select the representative objects that are centrally located in the cluster they define. Here, each remaining objects are clustered with the medoid to which it is the most similar. Partitioning method is then performed based on the principle of minimizing the sum of the dissimilarities between each object and its corresponding representative object called *absolute error criterion*. In simpler terms, with absolute error criterion, the average distance of the representative object to all other objects of the same cluster is minimized. The absolute error criterion used is defined as,

$$E = \sum_{j=1}^k \sum_{p \in C_j} |p - o_i|$$

Where, E is the sum of absolute-error for all the objects in the dataset; P is the point representing a given object and o_i is the representative object of cluster C_i . (Miller & Han, 2001)

K-medoid method selects initial representative objects arbitrarily. The iterative process of replacing representative objects (medoids) by non-representative (non-medoids) ones continues as long as the quality of resulting cluster is improved. The quality of a clustering is measured by the average dissimilarity between an object and the representative object of its cluster. The method computes the difference in the absolute-error value if a current representative object is swapped with a non-representative object. The total cost of swapping is the sum of differences incurred by all non-representative objects. If the total cost is negative, then the representative object is replaced with the non-representative since the actual absolute error would be reduced. Else, current representative object is considered acceptable and nothing is changed in iteration. (Miller & Han, 2001)

K-medoid method is more robust clustering method as it minimizes the sum of dissimilarities. It allows good characterization of all the clusters that are not too elongated and makes it possible to isolate outliers in most situations. (Kaufman & Rousseeuw, 1990)

There are different algorithms to perform K-medoids clustering. Most of the algorithm using K-medoids method is based on Partition Around Medoid (PAM) algorithm that operates on dissimilarity matrix of given dataset. There are two ways of entering the data in PAM. The most common way is by means of a matrix of measurement values. The rows of this matrix represent the objects and the columns correspond to the variables, which must be on an interval scale. Alternatively the program can be used by entering a matrix of dissimilarities between objects, which can be obtained in several ways with variables that are not necessarily on an interval scale but on binary, ordinal or nominal scale. PAM is useful to isolate the representative objects, which may be useful for data reduction or characterization purpose. (Kaufman & Rousseeuw, 1990)

Typically, PAM has three phases, the *build phase*, where an initial set of K representative objects are selected. The first selected object is the one for which the sum of dissimilarities to all other objects is as small as possible which is the dataset medoid. Other medoids are selected subsequently, one at a time, considering the object that most decrease the objective function. The second phase is called the *swap phase*, which computes the total cost

for all pairs of objects. The final phase is the *selection phase*, where a pair minimizing total cost is selected. If the minimum total cost is negative, swap is carried out and the algorithm re-iterates else, for each non-selected object, the most similar medoid is found and the algorithm stops. (Camila et al., 2008) For $n \times d$ data set, the PAM algorithm first computes the dissimilarity matrix, and then searches the optimal set of K data points as cluster prototypes to minimize the objective function by swapping all non-medoid data points and medoids. (Wire, 2012)

Initially, let us consider two representative objects O_i and O_j . If O_i is replaced with a non-representative object O_h , for all objects I , that are originally in the cluster represented by O_i , the most similar representative object is to be calculated. The PAM algorithm creates K clusters for the object and computes the total cost TC_{ih} of swapping for every pair of objects O_i and O_h . It then selects the pair of O_i and O_h that achieves the minimum of TC_{ih} . If the minimum is negative, O_i is swapped with O_h and the process is repeated until no swapping occurs. The final sets of representative objects are in the respective medoids of the clusters. The pseudo code for PAM algorithm is presented in Appendix 2.

The complexity of each iteration of PAM is $O(k(n - k)^2)$ and for large values of n and k , the computation is very costly. Thus, for the larger datasets, a sampling based method called CLARA (Clustering LARge Application) (Kaufman & Rousseeuw, 1990) can be used. In CLARA, instead of finding representative object for entire dataset, a sample of dataset is drawn and PAM is used on sample to find medoid of the sample. If the sample is drawn sufficiently random way, the medoid of the sample would approximate the medoid of the entire dataset. Thus, for better approximation, CLARA draws multiple samples and gives the best clustering as the output. Here, for accuracy, the quality of a clustering is measured based on the average dissimilarity of all the objects in the entire dataset. However, CLARA cannot find the best clustering if any of the best K-medoids are not selected during sampling. To overcome quality and scalability issue, another algorithm called CLARANS (Clustering Large Applications based on RANDomized Search) could be used. (Miller & Han, 2001)

2.9.2 Hierarchical Methods

In Hierarchical clustering method, a hierarchical classification of data is produced where the data items are not partitioned into a particular number of classes or group in single step. Instead, the classification consists of a series of partitions that may run from a single cluster containing all individuals, to n clusters, each containing a single individual (Everitt & Hothorn, 2011). The basic idea in hierarchical clustering is to link points that are close together into the same 'branch' in a tree representation of the distance. An important characteristic of hierarchical procedure is that the results at an earlier stage are always nested within the result at a later stage (Hair et al., 2006). Hierarchical methods suffer from fact that once a step (merge or split) is performed, it cannot be undone. Hierarchical process has drawbacks on selecting merge or split points in creating a cluster. The split point is critical, as the process at next step will only operate on the newly generated cluster. It is not possible to undo previously created cluster or swap objects between clusters in hierarchical method, thus, if merge or split decision is not well chosen, it may lead to low quality clusters. (Miller & Han, 2001)

Hierarchical method produces the tree like diagram to illustrate the arrangement of clusters called *dendrogram*.

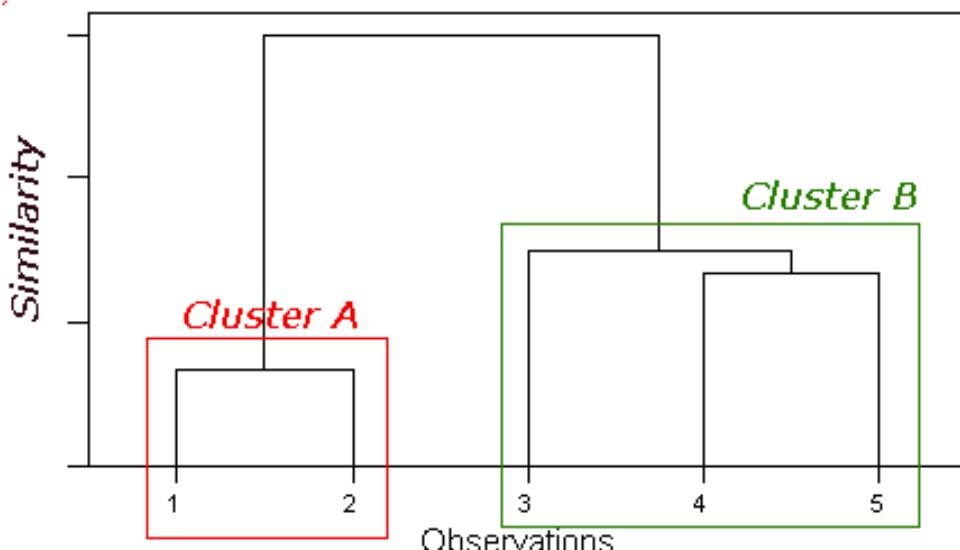


Figure 12 A dendrogram showing two distinct clusters (Manchester Metropoliton University,)

The tree structure in dendrogram is not a single set of cluster but is a multilevel hierarchy, where clusters at one level are joined as clusters at the

next level. In a dendrogram, the height of the lines indicates the distance between the objects that are connected.

Unlike partitioning method, hierarchical method gradually merges objects or divides a cluster. On the basis of merging or dividing of an object, hierarchical method can be classified as agglomerative or divisive method.

At each stage, the algorithm joins the two clusters that are closest together and uses the distance between clusters. The common distances used in hierarchical method are

$$\text{Minimum Distance: } d_{min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} |p - p'|$$

$$\text{Maximum Distance: } d_{max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} |p - p'|$$

$$\text{Mean Distance: } d_{mean}(C_i, C_j) = |m_i - m_j|$$

$$\text{Average Distance: } d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} |p - p'|$$

Where, $|p - p'|$ is distance between two objects or points p and p' ; m_i is the mean for cluster C_i and n_i is the number of objects in C_i .

When the algorithm uses the minimum distance to measure distance between clusters, it is called *single-linkage algorithm* or nearest neighbor clustering algorithm. Further, when the algorithm uses maximum distance, it is called *complete-linkage algorithm* or farthest neighbor clustering algorithm. Here, the maximum and minimum distance used for clustering tends to be overly sensitive to outliers or noise. Thus, mean or average distance is used to overcome outlier sensitivity problem. (Miller & Han, 2001)

Agglomerative Method

Agglomerative method of hierarchical clustering is also called bottom-up method and the process starts by placing each object in its own cluster. Here, individual clusters are then merged into larger cluster based on similarity until all of the objects are in one cluster or until certain criteria are fulfilled. (Miller & Han, 2001)

Agglomerative methods are based on measures of distance between clusters, where nearby clusters are merged to form reduced number of clusters. This is repeated each time merging the two closest clusters until just one big cluster of the entire data object is created (Hand et al., 2001)

Divisive Method

Divisive method of hierarchical clustering is also called top-down method and the process starts by assigning all objects to a single cluster. Here, single cluster is then subdivided into smaller clusters until each object forms a cluster of its own or until certain criteria is fulfilled. (Miller & Han, 2001) AGNES (Agglomerative Nesting) and DIANA (Divisive Analysis) are two hierarchical algorithms to produce the clusters.

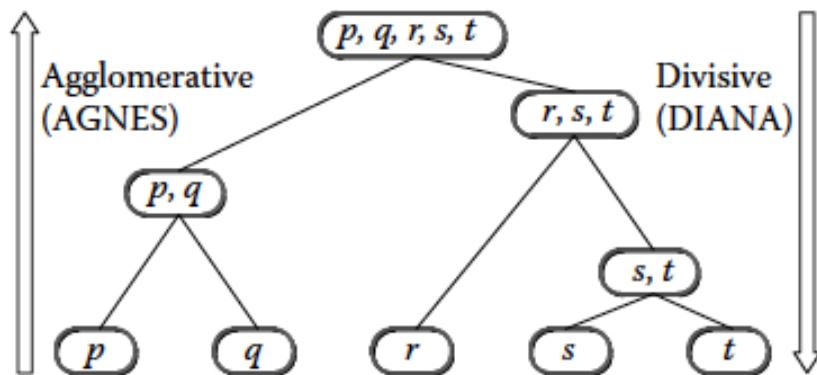


Figure 13 Agglomerative and Divisive Clustering on a set of data object (p,q,r,s,t) (Miller & Han, 2001)

Figure 13 shows different steps of AGNES and DIANA algorithm. For a dataset containing elements (p,q,r,s,t), DIANA algorithm starts by placing all data elements in single cluster which is subdivided into small cluster until each elements forms its own cluster. Alternatively, AGNES algorithm starts by placing each object in its own cluster, which is combined to bigger cluster until single cluster containing all element is formed.

2.9.3 Density Based Methods

In density based clustering method, clusters are the dense region of objects in the data space that are separated by regions of low density. It was developed to discover clusters with arbitrary shape. The idea is to continue growing a given cluster as long as density in the ‘neighborhood’ exceeds a threshold. Density based method is able to filter out noise and discover clusters of arbitrary shape.

DBSCAN (Density Based Spatial Clustering of Application with Noise) is a density based clustering method based on connected regions with sufficiently

high density. The method grows region with sufficiently high density into clusters and discovers clusters of arbitrary shape in spatial databases with noise. It defines clusters as a maximal set of density connected points. Here, the neighborhood area of radius ϵ is defined around each object called *epsilon-neighborhood*. If ϵ of an object contains at least a minimum number of objects (*Minpts*) then the object is a *core object*. Also, for given set of object, an object p is *directly density reachable* from another object q if p is within the ϵ of q and q is a core object. Further, an object p is *density-reachable* from another object q with respect to ϵ and *MinPts* in set of object D if there is a chain of objects p_1, \dots, p_n , $p_1 = q$, and $p_n = p$ such that p_{i+1} is directly density-reachable from p_i with respect to ϵ and *MinPts*, for $1 \leq i \leq n$, $p_i \in D$. Finally, an object p is *density-connected* to object q with respect to ϵ and *MinPts* in a set of objects, D , if there is an object $o \in D$ such that both p and q are density-reachable from o with respect to ϵ and *MinPts*. Thus, a density-based cluster is a set of density connected objects that are maximal with respect to density reachability and every object not contained in any cluster is considered to be noise.

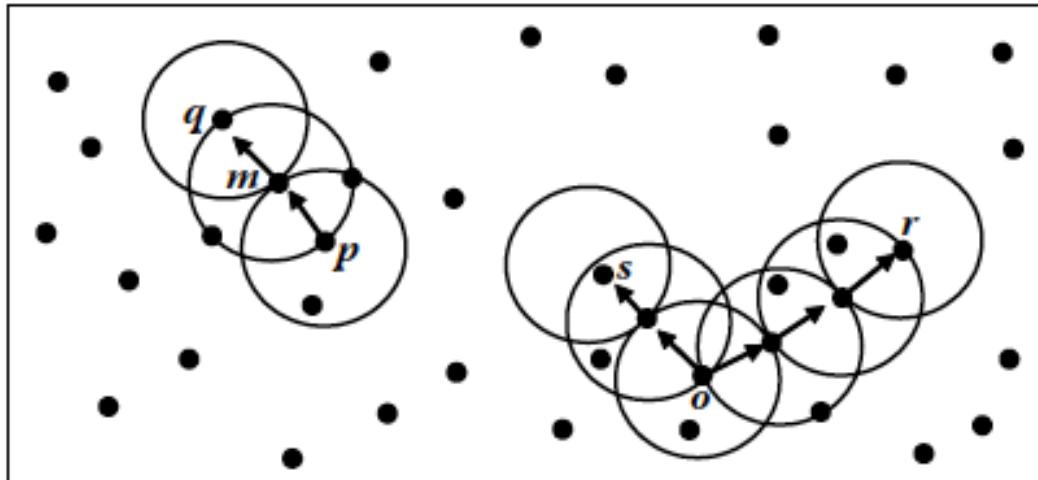


Figure 14 DBSCAN method and cluster detection (Miller & Han, 2001)

Figure 14 represents the basic idea behind the DBSCAN. In figure, epsilon-neighborhood is represented by circle around each point, which is used to determine density reachability and density connectivity and finally to obtain density based clusters.

OPTICS (Ordering Points to Identify the Clustering Structure) is a density-based method, which is used to reveal clusters with different local densities in

different regions of the data space. The OPTICS computes an augmented cluster ordering for automatic and interactive cluster analysis. The cluster ordering can be used to extract basic clustering information as well as provide intrinsic clustering structure.

DENCLUE (DENsity-based CLUstEring) is a clustering method based on set of density distribution function. The method is built on idea that the influence of each data point can be formally modeled using a mathematical function called influence function, which describes the impact of the data point within its neighborhood. Further, the overall density of the data space can be modeled analytically as the sum of the influence function applied to all the data points and clusters can be determined mathematically by identifying density attractors where density attractors are local maxima of the overall density function. (Miller & Han, 2001)

2.9.4 Grid Based Methods

In grid-based method, the object space is quantized into finite number of cells that form a grid structure. The clustering operations are then performed on the grid structure. The main advantage of grid based method is its fast processing time which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in quantized space (Miller & Han, 2001)

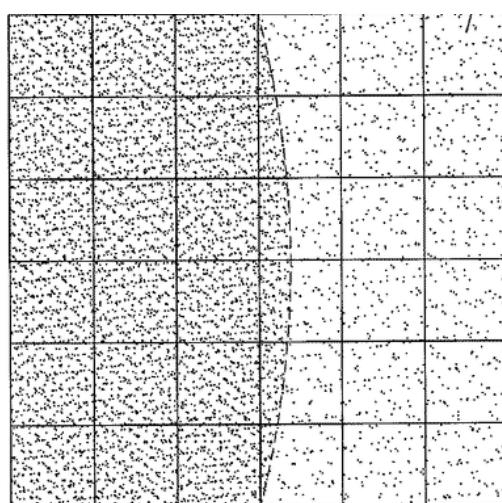


Figure 15 Grid Based Clustering (Patentdocs, 2011)

In Figure 15, geographical area is divided into number of rectangular grids. The clusters are then determined based on concentration of data points on a individual cell in the grid.

STING (Statistical Information Grid-based method) is a grid-based method where the spatial area is divided into rectangular cells using hierarchical structure. The algorithm computes statistical parameters (mean, variance) of each numerical feature of the object within cells and generates a hierarchical structure of the grid cells so as to represent the clustering information at different levels. Based on the structure, STING enables the usage of clustering information to search for queries or the efficient assignment of a new object to the clusters. (Halkidi et al., 2001)

2.9.5 Fuzzy Clustering

In fuzzy clustering method, objects are not assigned to a particular cluster rather they have a membership function indicating the strength of membership in all or some of the clusters. In all other methods, the strength of membership of an object to be in a particular is either one or zero which corresponds to whether the given object belongs to a certain clusters or not. In fuzzy cluster analysis, the numbers of subsets are assumed to be known, and the membership function of each object in each cluster is estimated using an iterative method, which is usually, a standard optimization technique based on a heuristic objective function. In general, membership functions do not obey the rules of probability theory, although, once found, memberships can be scaled to lie between zero and one, and can then be interpreted as probabilities (Everitt, 2011). The concept of membership function is derived from fuzzy logic, which is an extension of Boolean logic, where the concept of true and false is replaced by concept of partial truth. The connection between fuzzy cluster analysis and fuzzy logic is usually only through the application of membership function.

2.10 Number of clusters and heterogeneity measurement

For selection of optimum value of K , there is no standard objective selection procedure that exists. For determining the optimum number of clusters, one

of the measures to be considered is to check large increase in average within-cluster distance. When there is large increase, prior cluster solution with smaller within-cluster distance is selected as it's combination caused large increase in heterogeneity. This method has been shown to provide fairly accurate decisions in empirical studies but it is not uncommon for a number of cluster solution to be identified by these large increases in heterogeneity. Thus, the final cluster solution depends upon the dataset and could be a subject of expert evaluation.

Each cluster solution must be viewed for its description of structure balanced against the heterogeneity across the cluster. Heterogeneity measure should represent the overall diversity among observations in all of the clusters. As the observation are combined to form clusters, heterogeneity increases thus the measure of heterogeneity should start with value of zero and increase to show the level of heterogeneity as clusters are combined. A large increase in heterogeneity indicates that two dissimilar clusters were joined. (Hair et al., 2006)

2.11 Data Standardization

In many clustering applications, variables describing the clusters will not be measured in same unit. There may be variables of different type, which creates problem in interpretation of data. Hence, to deal with the problem of different units of measurement, each variable is standardized to unit variance prior to analysis (Everitt, 2011). Data standardization provides solution to complication in comparison between variables. However, in some cases, the standardization process could remove some natural relationship reflected in the scaling of the variables. The decision to standardize should be based on research objectives and the empirical qualities of the data. It must be considered that most cluster analysis using different distance measures are sensitive to different scales or magnitude among the variables. Generally, the variables with large standard deviation have more impact on the final similarity value. Thus, with the use of similarity measure, the prospect of data standardization must be known (Hair et al., 2006)

Standardization of data can be performed by three different methods. First method is standardizing the variables by conversion of each variable to

standard scores by subtracting the mean and dividing the standard deviation for each variable. It is the general form of normalized distance function and converts each raw data score into a standardized value with a mean of 0 and a standard deviation of 1, which eliminates the bias introduced by the differences in the scales of several attributes or variables. The second method is by using a standardized Mahalanobis distance measure. It not only standardizes the data by scaling in terms of the standard deviation but also sums the pooled within-group variance-covariance, which adjusts for correlations among the variables. The third method is standardizing by observation. It helps to identify the groups according to their response style (question and its response) (Hair et al., 2006).

2.12 Cluster Validation

Cluster validation is the method for quantitative evaluation of the result of clustering algorithm (Theodoridis & Koutroumbas , 2003). The question concerning the evaluation of goodness of the resulting cluster is important in order to avoid finding patterns in noise, to compare clustering algorithms, to compare two sets of clusters or to compare two clusters. Validity of cluster is important as it evaluates how well the result of analysis fits the data with or without reference to external information, compares the result of two different sets of cluster to determine the better one, to determine the correct number of clusters and to find the partitioning that best fits the underlying data.

To investigate the cluster validity, two different approaches are used. The first one is based on the *external criteria*, which implies the evaluation of the result of clustering algorithm based on the pre-specified structure, which is imposed on a dataset and reflects the intuition about the clustering structure of the dataset. The second approach is based on the *internal criteria*, where the results of clustering algorithm in terms of quantities that involve the vectors of the data set themselves are evaluated.

For selection of an optimal clustering scheme, two criteria are purposed namely compactness and separation. Compactness of cluster measures how close the members of each cluster are and is measured using variance. Separation on the other hand measures how distinct is a cluster from other cluster.

However, validation method only provides an indication of the quality of the resulting partitioning and thus, can only be considered as a tool for experts to evaluate the clustering results. (Halkidi et al., 2001)

There are different internal criteria available for cluster validation like, Ball-Hall Index, Banfeld-Raftery Index, Calinski-Harabasz Index (Desgraupes, 2013), Silhouette Index (Rousseeuw, 1987). However, all of the internal indices used to validate the clusters use the Euclidean distance between clusters or from a point to a cluster (Desgraupes, 2013; Rousseeuw, 1987; Halkidi et al., 2001) to calculate the value of given index. Thus, according to expert opinion, there might be cases where internal criteria could favor one distance measure over another. Since, the research compares different distance measures, use of indices that uses distance, as input was deemed unsuitable. Thus, in this research, only the external criteria were used to validate the clusters.

2.12.1 External Criteria

External criteria of cluster validation are indices designed to measure the similarity between two partitions. They take into account only the distribution of the points in the different clusters and do not allow measuring the quality of this distribution. (Desgraupes, 2013)

In this approach, the basic idea is to test whether the points of the dataset are randomly structured or not. The analysis is based on the *Null Hypothesis* H_0 , expressed as a statement of random structure of a data. The hypothesis is tested in two fold. First, a reference data population under random hypothesis is generated which is a data population that models a random structure. Second, appropriate statistic, whose values are indicative of the structure of a dataset and compare the value that results from the dataset against the value obtained from the reference population. (Theodoridis & Koutroumbas , 2003)

There are different external indices that could be used to validate given cluster. Different indices for external validation method are dependent on Misclassification matrix representing the count of pairs of points depending on whether they are considered as belonging to the same cluster or not.

For the clustering structure $C=[C_1\dots C_m]$ of a dataset X and $P= (P_1\dots P_s)$ is defined partition of the data. The pair of points (X_v, X_u) from the dataset, can be referred using following terms:

- **SS**: if both points belong to the same cluster of the clustering structure **C** and to the same group of partition **P**
- **SD**: if points belong to the same cluster of **C** and to different groups of **P**
- **DS**: if points belong to different clusters of **C** and to the same group of **P**
- **DD**: if both points belong to different clusters of **C** and to different groups of **P**

Now, let's assume a , b , c and d is the number of SS, SD, DS and DD pair respectively. Then, maximum number of all pairs in the dataset is given by,

$$M = a + b + c + d = \frac{N(N - 1)}{2}$$

Where, N is the total number of points in the dataset.

Using the above values, different external indices could be defined to measure the degree of similarity between cluster **C** and partition **P**.

Rand Statistics

Using above notations, Rand Statistics is defined as,

$$R = \frac{(a + d)}{M}$$

Rand statistics has value between 0 and 1. The higher value indicates greater similarity between **C** and **P**.

Jaccard Coefficient

Using above notations, Jaccard Coefficient is defined as,

$$J = \frac{a}{(a + b + c)}$$

Jaccard Coefficient has value between 0 and 1. The higher value indicates greater similarity between **C** and **P**.

Folkes and Mallows Index

Using above notations, Folkes and Mallows index is defined as,

$$FM = \frac{a}{\sqrt{m_1 m_2}} = \sqrt{\frac{a}{a + b} \times \frac{a}{a + c}}$$

Where, $m_1 = a + b$ and $m_2 = a + c$

Folkes and Mallows Index also has value between 0 and 1 and as Jaccard Coefficient, higher values of the indices indicate greater similarity between **C** and **P**. (Halkidi et al., 2001)

3 Data Analysis

Data analysis process starts with design of analysis process followed by description of study area and related dataset used for cluster analysis. Further, use of K-medoid with different distance measure in dataset is presented. Finally, comparison between different distance measures, validation of clusters, and interpretation of cluster results are presented.

3.1 Analysis Design

In this section, a simple workflow of the analysis process is documented. The input for the analysis process would be different data layers where k-medoid clustering method would be applied and the final result will be a classification of area into different categories.

For solving the suitability problem, two types of knowledge is required. First is domain knowledge, which is the knowledge about the problem, the factors that affect it and how these factors affect one another. Second is GIS knowledge, which is the knowledge about how to use and analyze spatial data and how to use spatial data in problem solving. (Nikander et al., 2012)

The overview of different analysis steps used in this research is presented in figure 16.

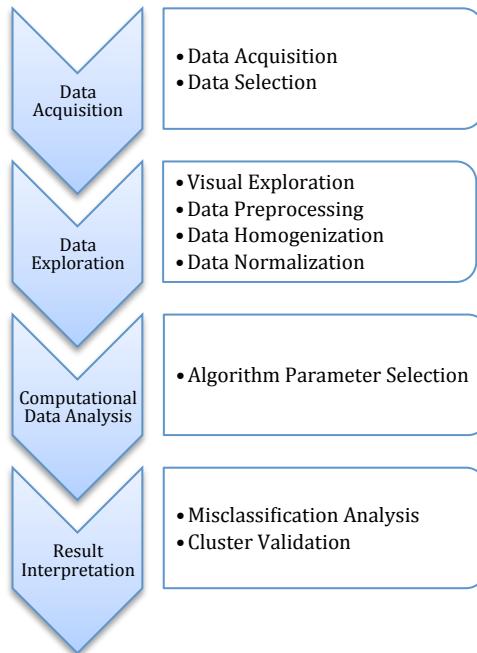


Figure 16 Workflow of Analysis Process

The analysis process starts with *Data Acquisition*, where different geographic datasets required to solve the given problem are acquired. In this phase, domain knowledge is required for analyzing the problem and deciding which data and information is required to solve the problem. GIS knowledge is required in analyzing the possible input data, and finding which of these are available as spatial datasets and where such spatial data can be found.

After Data acquisition, *Data Selection* is performed, where the acquired data sets are categorized according to the use in analysis process. Here, domain knowledge is required to know how a given input data layer affects the problem and GIS knowledge is required to know how the input data can be used in analysis.

After selecting the data for input, different details about the data is visualized and modified if required. This process of familiarizing with details of input data and its modification is called *Data Exploration*. Through *Visual Exploration*, different visualizations of the input data are produced to explore and familiarize with the details of input data. In order to transform the input data layers to be used in further steps of analysis, *Data preprocessing* is performed. Further, *Data Homogenization* is done to transform the input layers so that all the layers use same coordinate projection and have same resolution, and thus can be used in analysis process. Finally, to prepare the data for further processing, *Data Normalization* is performed so that the data layers are in normalized format and the data values of various layers can be compared. In data exploration process, GIS knowledge is used to understand the contents of various visualizations that are used for exploring the data, and drawing inferences from it, to select appropriate preprocessing for input layers and to know how to preprocess the layers. Domain knowledge is used to understand how different input layers affect the problem, to know what the data layers need to be present after preprocessing and how each data layer independently affects the problem and thus how the particular layer should be normalized.

In *Computational Data analysis*, the input data is used for computation of clusters by selecting appropriate algorithm and parameters.

The output from the computational data analysis is reviewed in *Result Interpretation* phase. In this step, *Validation methods* are used to compare

between different clusters and to provide the answer to the research questions presented in section 1.3. In this phase, domain knowledge is required mainly for interpreting how well each part of the output suits the activity that is being analyzed and what sort of suitability value should be given. Further, GIS knowledge is required to explore the algorithm output. (Nikander et al., 2012)

In this research, data exploration is performed with the help of ArcMap and Computational data analysis is performed with Matlab. In data analysis, K-medoid method with different distance measures will be used to obtain clusters, which would be further analyzed to obtain difference between in cluster output between different distance measures.

3.2 Study Area and Data Description

The study area in the research is of 16 X 20 Square Kilometer from Lahti, Finland. The area used in the experiment was a part of central Finland that had both wilderness and urban areas.

For the purpose of mobility analysis, the factors affecting the mobility are determined which are *soil type, amount and type of vegetation* (represented by total cross-sectional areas of trees from 1 to 3 meters in height), *degree of slopes, roads and buildings*.

The data layers require pre-processing and are used in different ways during the analysis process. The effect of road and buildings on mobility is not influenced by other input layers thus are not included in analysis process for this study. The three dataset are combined to know the overall effect that they have on mobility and are used throughout the analysis process. (Nikander et al., 2012)

The slope of terrain varies from 0-41 degrees. The vegetation layer, which is the diameter of vegetation from 1-3 m height, has values of 0-39 m^2/ha . Higher value for vegetation layer corresponds to dense vegetation areas. Soil type consists of information about different type of soils present in the study area and varies from bedrock, clay, sand, water and many more. The data used in this study is assumed to contain no spatial dependencies, and thus can be analyzed without taking spatial autocorrelation into account.

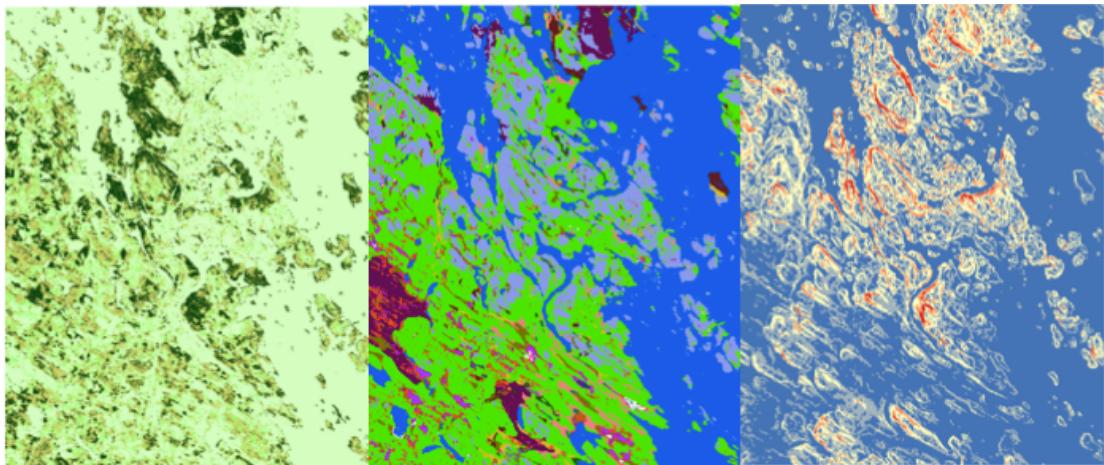


Figure 17 Vegetation, Soil Type and Slope raster layers

These dataset were imported to ArcMap for visualization. In Figure 17, for vegetation layer, darker the green color, more vegetation is present. For the soil type, different colors represent different soil type. The blue color in soil type represents water. In Slope layer, deeper the blue color, less steeper the slope. Extremely steep slopes are represented in red color.

3.3 Data Exploration

After data acquisition and classification, the data is explored and modified to be comparable and usable as input for computational method. Here, the data is transformed into usable form for rest of the process. Initially, Slope data is obtained from the DEM (Digital Elevation Model) through preprocessing.

All the dataset was converted into appropriate coordinate system (ETRS-TM35) and resampled to have same resolution of 20 X 20 m. Further; the dataset is normalized between 0 and 10 with one-digit precision to make the dataset comparable. Here, 10 represents perfect mobility and 0 represents no mobility. Also, expert knowledge is used to assign different values to different classes of individual layer of the dataset. (Nikander et al., 2012)

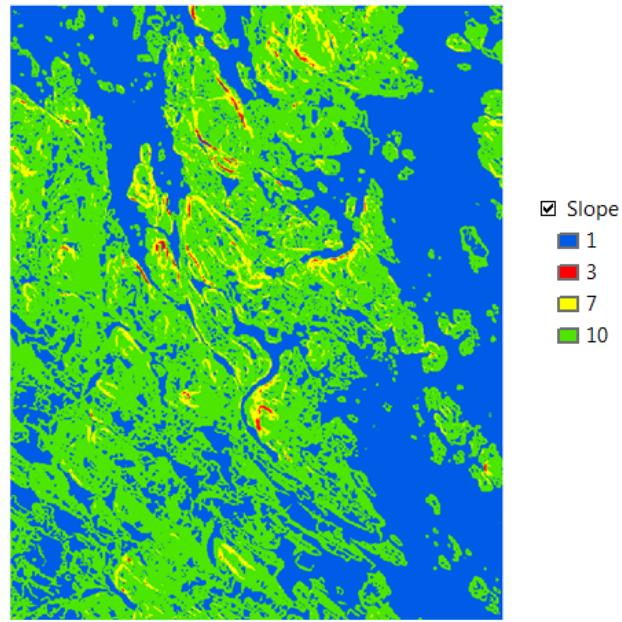


Figure 18 Normalized Slope Layer

During normalization, for slope layer (Figure 18), areas with smaller slopes were assigned higher mobility values, thus have better mobility, which decreases as the slope increases. In figure 18, blue color corresponds to area with very steep slope or water area whose combination has limited mobility, green color corresponds to flat land or areas with very gentle slope, and red color with areas of highest slope, which is less suitable for mobility. Since, areas with water and steep slope layers are combined together, blue color was used for their visualization, as area with steep slope was fairly less than water areas.

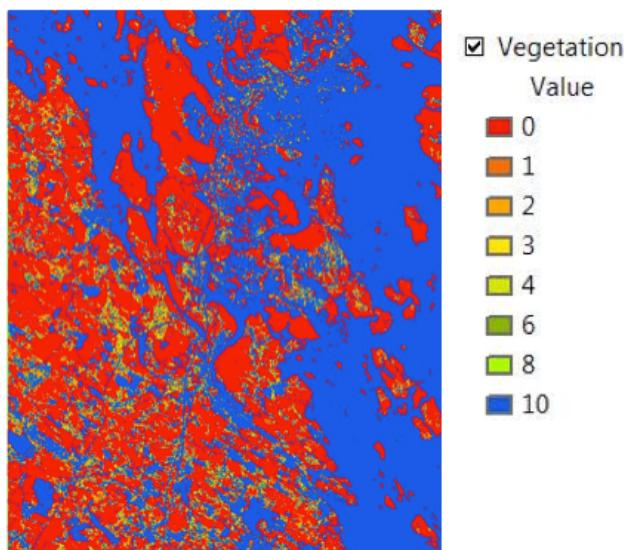


Figure 19 Normalized Vegetation Layer

Further, for vegetation layers (Figure 19), areas with dense vegetation (red areas) were assigned lower values, thus has lower mobility. The mobility value increase as the amount of vegetation decreases. In figure 19, red area has dense vegetation and hence has lower value for mobility; the blue area corresponds to water which lacks vegetation thus has higher mobility value. Further, areas with other colors have different mobility depending upon the amount of vegetation in particular area.

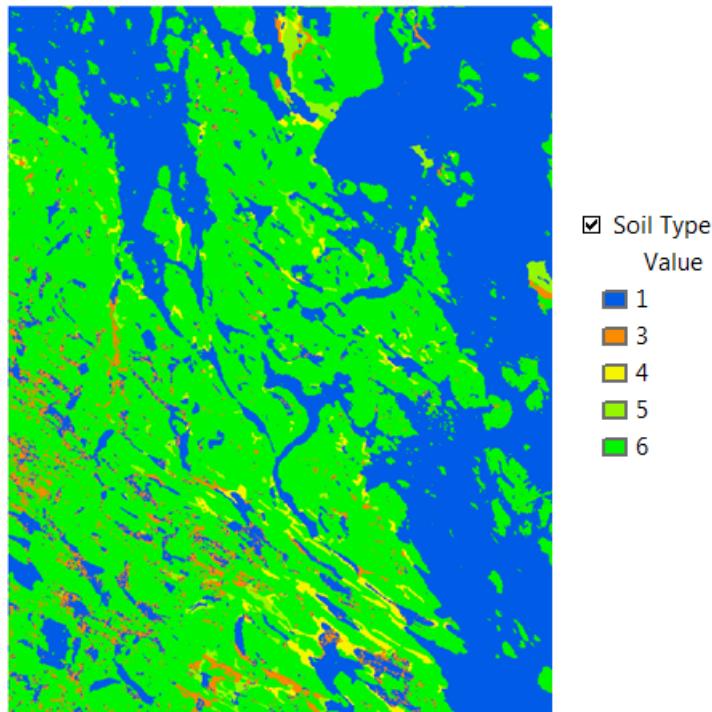


Figure 20 Normalized Soil Type Layer

Finally, for soil type (Figure 20), different soil types were assigned different values based on expert knowledge. In figure 20, areas with water (blue areas) have zero as mobility value whereas other soil types were either combined together for a mobility value or were given single mobility value based upon type of soil. Generally, soil type with good mobility, like different types of moraine soils is assigned higher mobility values (green areas). The detail about different classes used for different layers is provided in Appendix 3.

3.4 Computational analysis

For clustering, the reclassified datasets were converted into vector format. Conversion of raster data into vector format is required as for K-medoid clustering; raster data could not be converted into clusters. A new dataset

was created by combining vegetation, soil type, and slope data in vector and is used for clustering.

After normalization of data and prepared for clustering, computational analysis method is applied in order to obtain different clusters. Since K-medoid method uses random selection of initial centroids, for proper comparison between the different distance measures, initial centroid along with number of clusters needs to be same for different distance measures. The initial centroids were selected by performing preliminary clustering on random subsample of dataset. Since, the clue for obtaining good cluster is to select the representative objects that are centrally located in the cluster they define, the initial centroid was further analyzed so that it represents the centrally located objects in the given cluster.

The number of clusters was selected iteratively by executing the algorithm with different number of clusters and selecting the best number of clusters from the output. The number of clusters is to be determined experimentally, k=8 was selected based on experimental evaluation and expert opinion of the cluster result. With k=8 and same initial centroid, different distance measures were applied in K-medoid method to obtain the cluster output of different distance measures. The output k-medoids provides 8 clusters including cluster centers and members. The output of cluster analysis is further processed to obtain a map representing the geographic distribution of clusters.

4 Result Interpretation

The interpretation of clustering result is performed in two steps. First cluster map is created from the cluster output. Then, validation method is used to determine differences between clusters produced from different distance measures.

In this section, distance measure corresponds to the clusters produced as a result of using particular distance measure. For example, a cluster produced using Euclidean distance is represented just by 'Euclidean distance'.

4.1 Cluster Map

The output of computational method requires user/expert interpretation to obtain knowledge about cluster and finally about mobility in given terrain. Initially, the output must be visualized for showing clustering result. The output of cluster analysis is further processed to obtain a map representing the geographic distribution of clusters called the *Cluster Map*. The Cluster Map serves as the visual representation of clusters in the given geographical area. It is also the starting point to determine the mobility in given terrain. Each cluster is represented by unique color that aggregates the similar area.

In this study number of clusters, $k=8$, is used thus creating eight different clusters from the dataset. The exact information obtained from the cluster map is subject to user/expert interpretation as well as the expected outcome of the analysis process. In cluster map, blue color represents clusters of water area whereas other colors simply represent different clusters in given geographic area. In this study, the aim is to determine mobility in given terrain, hence, cluster map containing different clusters could be converted into mobility map by providing a mobility value for each cluster.

The cluster map created from different distance measure is illustrated in figure 21-25.

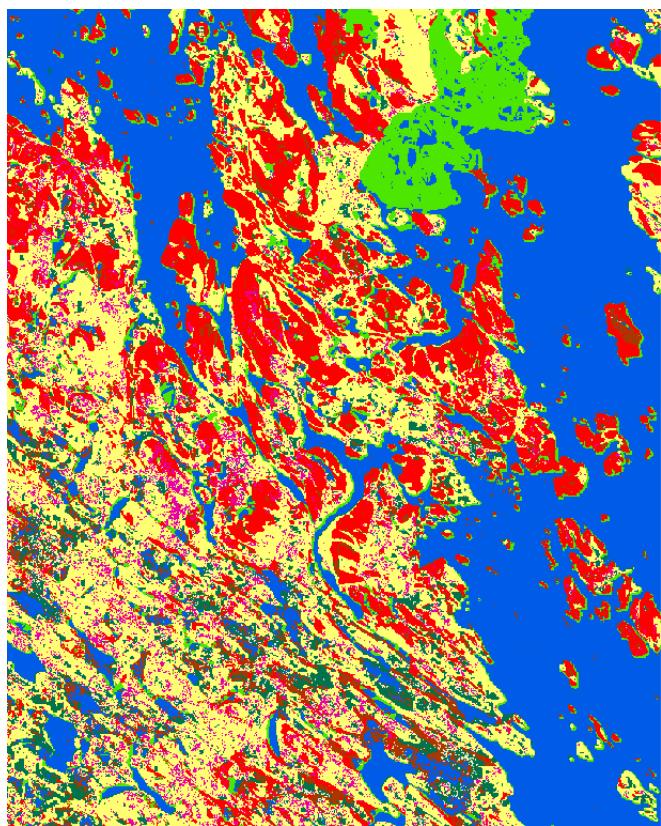


Figure 21 Cluster map created using Euclidean distance

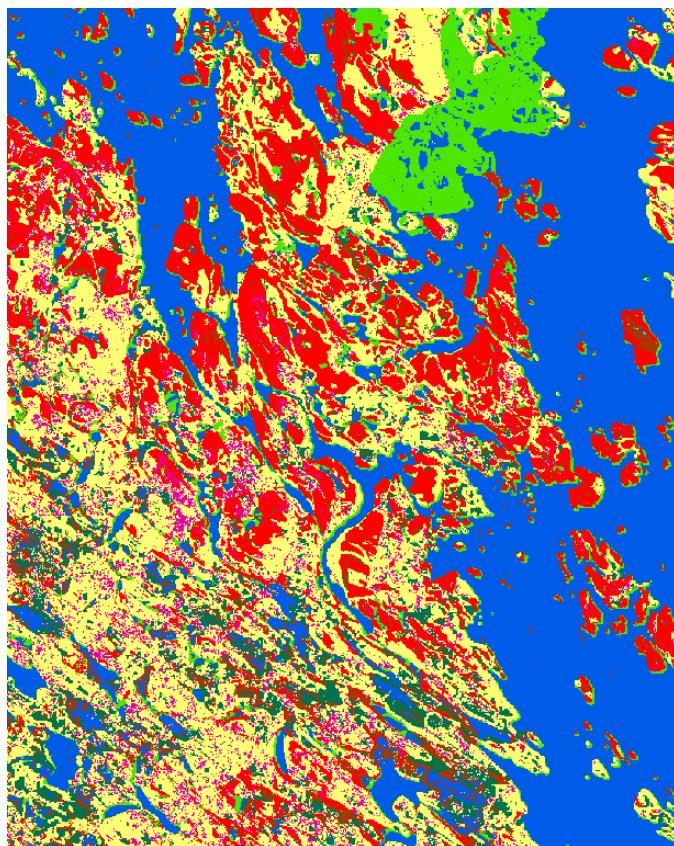


Figure 22 Cluster map created using squared Euclidean distance

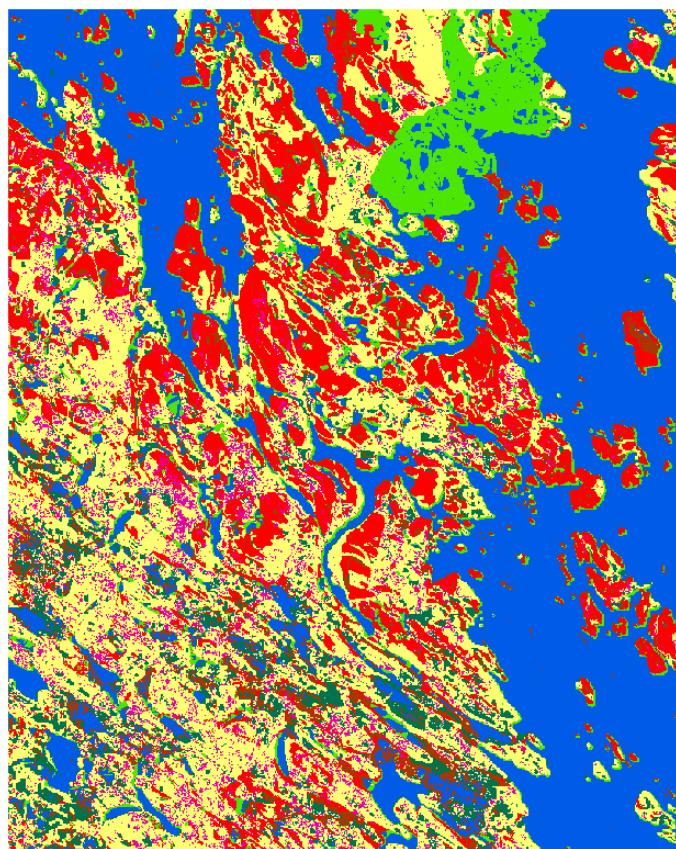


Figure 23 Cluster map created using Manhattan distance

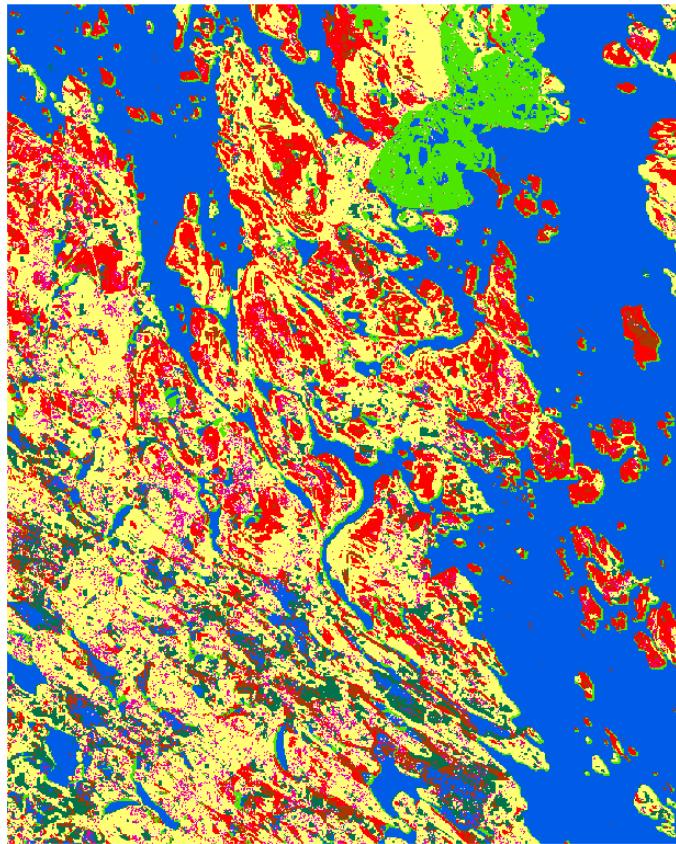


Figure 24 Cluster map created using Mahalanobis distance

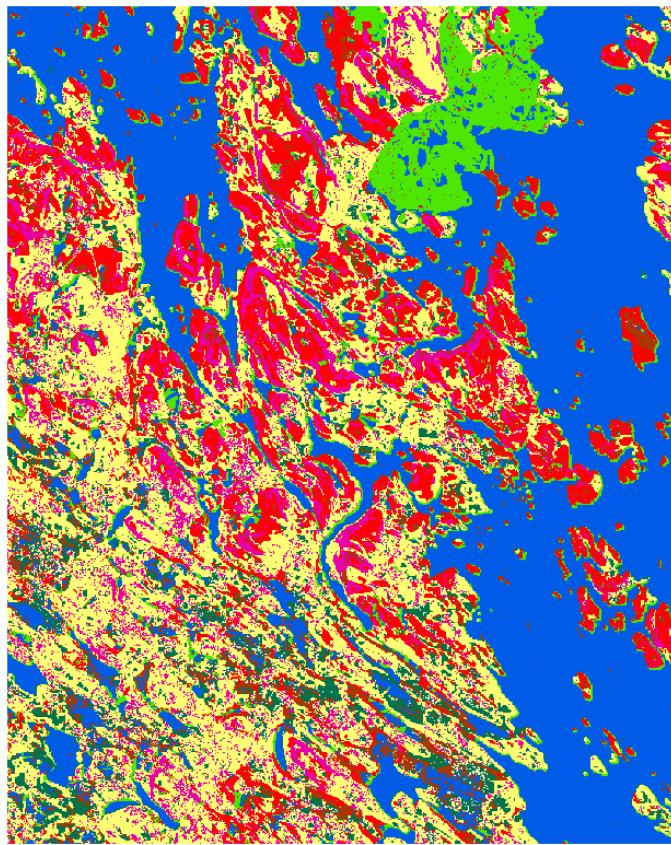


Figure 25 Cluster map created using Chebyshev distance

As the dataset was normalized based on mobility values, the final result of clustering depends on normalization of data as well as selection of ‘k’, optimum number of clusters. With different values of normalization and different values of ‘k’, the obtained cluster map will be different.

For providing mobility value to a cluster, visualization of clustering result is required. Cluster result could be visualized using *Parallel Coordinates Plot* (PCP), which is used to show n-dimensional data points with polylines that have vertices on the parallel axis. Here, the dataset is 4-dimensional points where each point contains a value from each three input layer and a cluster number. The cluster result is then linked with topographic map view so that when an area from PCP is highlighted, corresponding points from the topographic maps are also highlighted. The topographic map provides additional information about the topography and the PCP is used to provide easy visualization of the cluster’s data content.

Since mobility analysis is suitability problem, the suitability can be evaluated by combining similar locations obtained from clustering into classes, and giving each class a suitability value. Thus, to solve the suitability problem, the

cluster needs to be categorized according to the suitability of the items in the clusters and the clustering result needs to be interpreted. (Nikander et al., 2012). For interpretation of result for mobility, each cluster is visualized using different colors and individual cluster is assigned a mobility value based on its location in topographic map and expert knowledge. For assigning mobility value, it is required to view both attribute values of cluster and geographic distribution of cluster. After each cluster has been given a mobility value, the cluster map can be turned into a mobility map. The mobility value provided for the clusters could be divided into three categories: NO GO, for the areas that cannot be crossed, GO SLOW for areas, where maximum practical speed is slow and GO for areas where it is possible to drive fast or alternatively based on the requirement of end users. Practically, the clusters produced as a result of good mobility values (8-10) in input layers correspond to GO areas, values of 4-7 correspond to GO SLOW areas and values of 0-3 correspond to NO GO areas. However, these values could be changed based on expert evaluation of the cluster map as well as normalization of input layers.

The conversion of cluster map into mobility map is not straightforward. There has to be link between input layers, different clusters and its corresponding location in the map in order to assign mobility value for given cluster. Further, assigning mobility value to a set of cluster is ambiguous and requires specialized tools that could connect the cluster to corresponding location in the topographic map, as well as expert knowledge about the dataset. In other words, the cluster should be linked to the topographic map area by the help of PCP. With lack of available tools for such view, it is not possible to assign particular cluster a mobility value and hence not possible to create the mobility map.

One such toolkit providing linking cluster with topography is *Infovis 2005* (Jean-Daniel, 2004) which is an interactive Graphics Toolkit written in Java to aid information visualization. The toolkit provides different visualization method including scatter plot, time series, PCP, node-link diagram, tree maps and adjacency matrices. However, the toolkit is a decade old and is not regularly updated and hence cannot support the computation problems faced today. Another such toolkit is *Riskigis* developed by Jussi Nikander, as a research prototype (Nikander, 2012). It could incorporate various source

datasets in the form of gridded map layers as input to create the mobility map using clustering based on similarity. However, in the software prototype, only K-means and DBSCAN method could be used to determine clusters, hence, it was not a feasible solution to research problem in this case. Thus, due to lack of toolkit, conversion of cluster map into mobility map was not possible. Thus, alternative method is used that reveals the difference between the clusters produced.

4.2 Cluster Validation

Result of cluster analysis is validated by two ways. First, by using misclassification matrix with visual analysis method and second, by using cluster validation indices.

4.2.1 Misclassification Matrix and Visual Analysis

For investigating difference in result between clusters created using different distance measures in K-medoid clustering, *misclassification matrix* was created by comparing the cluster map created by one distance method with the cluster map created by another distance measure. The misclassification matrix represents misclassification between each class of clusters between different distance measures. From the misclassification matrix, *Kappa Index* is calculated which has values between zero and one and shows how much better the classification is compared to a totally random distribution of data values. Zero corresponds to totally random distribution and one to a perfect match between classifications. As $k=8$ is used for calculating clusters, each different distance measures produces eight different clusters. Since K-medoids clustering method assigns all data elements into clusters, there are no unclassified elements in the result of K-medoids clustering.

Further, the result of cluster analysis is evaluated by applying visual analytics approach. The concept of evaluation is based on analysis of result with help of interactive visual interfaces using visual, mathematical or computational analysis method. The evaluation of cluster analysis result is based on expert reasoning in connection with use of interactive visual method in spatial context. The quality of result depends heavily on the knowledge and reasoning skill of analyst in visual analysis process (Hall et al., 2014). The visual analysis process could be simple visual comparison between two

datasets, mathematical analysis of the dataset for comparison or alternatively different form of visualization for the dataset in order to establish certain association. In this paper, only visual comparison between different cluster maps is considered.

To compare between different distance measures, initially a misclassification matrix is computed by comparing result of each distance measure with another. The misclassification matrix reveals how big the differences are between clusters of different distance measures. Each row in the table represents how one data value in clustering result is divided between the data values in different distance measures. The Kappa index is then calculated from the table. To visualize the difference between clustering results of different distance measures, the cluster map of two distance measures are combined. After combination, a map is created (Figure 26), where, those elements present in diagonal of the misclassification matrix, which are the data values that correlates in both result are assigned a single color (green) and all off-diagonal elements, which are the data values that does not correlate in both result, represents difference in clustering are assigned another color (red). The clusters representing water area in the map is assigned blue color. This map provides the information about the areas where the result varies between different distance measures. With further, interpretation of the map, the reason for difference could be identified.

In the research, five different distance measures are used for K-medoids clustering thus; ten different misclassification analyses are performed which are presented in table 1-10. Also, cluster created from each distance measure is compared with another visually and result of which is presented in figures 24-29.

		Squared Euclidean Distance								
		clusters	1	2	3	4	5	6	7	8
Euclidean Distance	1	276161	0	0	0	0	0	218	0	
	2	0	39038	0	0	0	0	0	0	
	3	0	0	204077	0	0	0	0	0	
	4	0	309	0	134515	0	0	0	0	
	5	0	0	0	0	31872	0	418	0	
	6	0	22	0	0	0	58868	0	0	
	7	0	0	0	0	349	0	4449	3788	
	8	0	0	0	0	0	0	0	45916	
	Kappa								0.9918	

Table 1 Misclassification Matrix of Euclidean and Squared Euclidean Distance

		Manhattan Distance								
		Clusters	1	2	3	4	5	6	7	8
Euclidean Distance	1	276379	0	0	0	0	0	0	0	0
	2	0	39038	0	0	0	0	0	0	0
	3	0	0	204077	0	0	0	0	0	0
	4	0	725	0	134099	0	0	0	0	0
	5	0	0	0	0	32290	0	0	0	0
	6	0	0	0	0	0	58890	0	0	0
	7	0	0	0	0	549	0	5408	2629	
	8	0	0	0	0	0	0	0	45916	
Kappa										0.9937

Table 2 Misclassification Matrix of Euclidean Distance and Manhattan Distance

From table 1 and 2, it can be observed that there are no significant differences between clusters created by Euclidean distance, squared Euclidean distance and Manhattan distance. The Kappa coefficient of >0.99 in both table 1 and 2 represents very small differences between the cluster solutions.

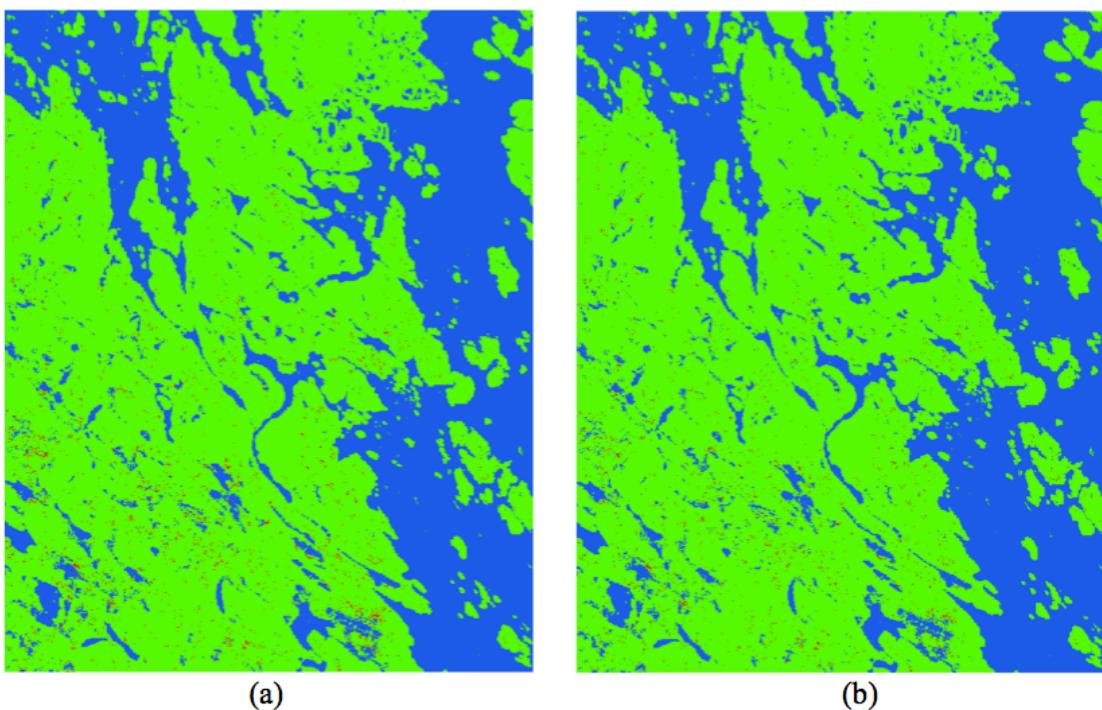


Figure 26 Difference between clusters created using different distance measures. (a) Between Euclidean and squared Euclidean distance and (b) Between Euclidean and Manhattan distance.

Now, when the map is created to visualize the difference (figure 26), the areas where difference exists (red areas), are randomly distributed throughout the map. The difference was found only in 0.638% of the study area between Euclidean and squared Euclidean distance, and in 0.487% of study area between Euclidean and Manhattan distance. Since, the difference in result between the clusters created from Euclidean distance, squared Euclidean distance and Manhattan distance is small and the difference is

distributed throughout the study area, it can be associated with difference in methods while using three different distance measures so no further evaluation of result is conducted.

		Mahalanobis Distance								
		Clusters	1	2	3	4	5	6	7	8
Euclidean Distance	1	274740	1014	0	0	0	0	625	0	
	2	0	27854	0	0	0	0	11184	0	
	3	0	0	204077	0	0	0	0	0	
	4	0	5787	19498	108631	0	908	0	0	
	5	4	795	0	0	31491	0	0	0	
	6	0	797	0	0	0	57158	935	0	
	7	0	2960	0	0	253	0	2744	2629	
	8	0	0	0	0	0	0	0	45916	
Kappa									0.923	

Table 3 Misclassification Matrix of Euclidean Distance and Mahalanobis Distance

		Chebyshev Distance								
		Clusters	1	2	3	4	5	6	7	8
Euclidean Distance	1	276368	0	0	0	0	0	11	0	
	2	0	39038	0	0	0	0	0	0	
	3	0	9801	194276	0	0	0	0	0	
	4	0	23440	0	111384	0	0	0	0	
	5	650	0	0	0	31640	0	0	0	
	6	6	259	0	670	0	57955	0	0	
	7	0	0	0	0	353	0	8233	0	
	8	0	0	0	0	0	239	45677		
Kappa									0.9432	

Table 4 Misclassification Matrix of Euclidean Distance and Chebyshev Distance

When Euclidean distance is compared with Mahalanobis and Chebyshev distance (Table 3 and 4), certain difference could be observed. The Kappa coefficient of 0.923 for Mahalanobis distance and 0.9432 for Chebyshev distance represents some differences between the clustering outputs.

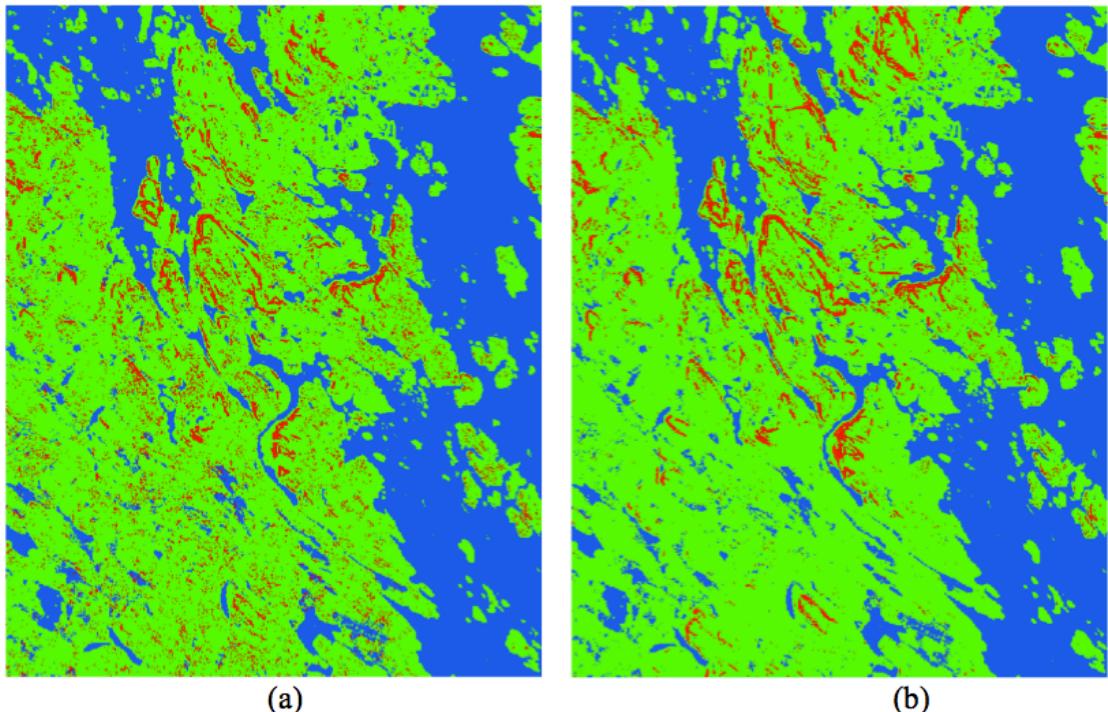


Figure 27 Difference between clusters created using different distance measures. (a) Between Euclidean and Mahalanobis distance and (b) Between Euclidean and Chebyshev distance.

When the difference is visualized in map (figure 27), the areas where difference exists (red areas), are randomly distributed throughout the map area with red clusters appearing in some particular area. The difference was found in 5.92 % of study area between Euclidean and Mahalanobis distance, and 4.427% of study area between Euclidean and Chebyshev distance. However, due to aggregation of those differences in particular areas, the difference is subject of further evaluation.

For evaluating the difference, result of Euclidean and Mahalanobis distance was compared with the input layers (Figure 17). The areas with difference have particularly steep slope (more than 30 degrees), dense vegetation and rocky soil. As steep slope, dense vegetation and rocky soil would correspond to bad mobility; Mahalanobis distance was particularly useful in determining correlation between these three layers in some parts of study area. Again, for Chebyshev distance the data objects which are different in any one of the dimension is dominant over other and tends to represent the values that is highest in certain dimension. When the normalized soil type, vegetation and slope layer is compared with the result of difference between Euclidean and Chebyshev distance, normalized slope layer (Figure 18) had higher value of mobility for the red areas in figure 27. Here, Chebyshev distance favors slope

layer over other layer in areas where slope layer have higher mobility value than other. This could be due to normalization process of slope layer as it has fewer classes than of vegetation and soil type layers after normalization. Hence, due to this tendency of Chebyshev distance, difference in result was obtained when compared with Euclidean distance.

		Manhattan Distance								
		Clusters	1	2	3	4	5	6	7	8
Squared Euclidean Distance	1	276161	0	0	0	0	0	0	0	0
	2	0	39347	0	0	0	0	22	0	0
	3	0	0	204077	0	0	0	0	0	0
	4	0	416	0	134099	0	0	0	0	0
	5	0	0	0	0	32221	0	0	0	0
	6	0	0	0	0	0	58868	0	0	0
	7	218	0	0	0	618	0	4249	0	0
	8	0	0	0	0	0	0	1159	48545	0
	Kappa									0.9961

Table 5 Misclassification Matrix of Squared Euclidean Distance and Manhattan Distance

When comparing squared Euclidean distance and the Manhattan distance, there were no significant differences found between the clusters created by squared Euclidean distance and Manhattan distance. The Kappa coefficient of 0.9961 represents very small differences between the cluster solutions.

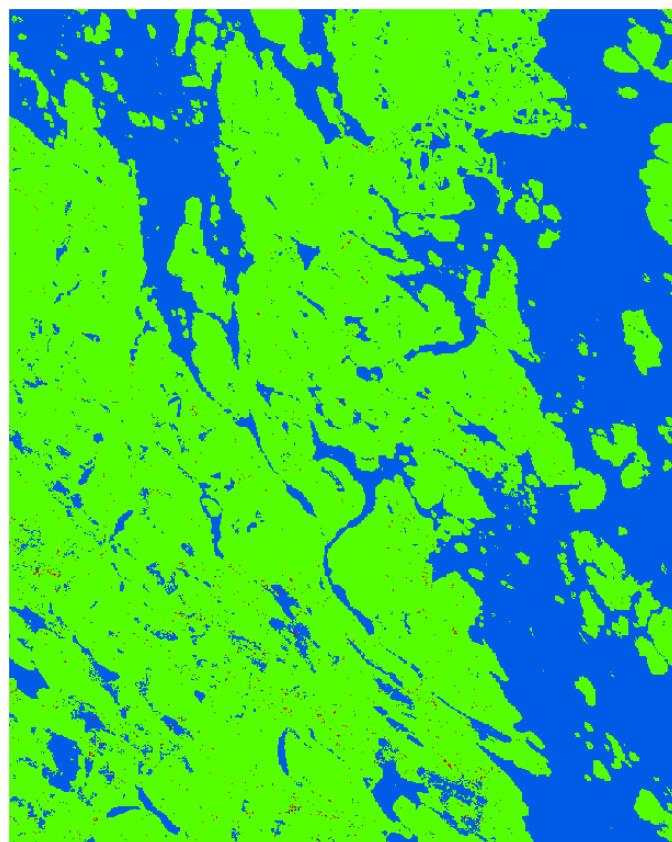


Figure 28 Difference between clusters created using squared Euclidean and Manhattan distance

To visualize the difference, a map is created (figure 28), which shows that the areas where difference exists are randomly distributed throughout the map. These differences were visible in 0.304% of the study area between squared Euclidean and Manhattan distance. Since, the difference in result between the clusters created from squared Euclidean distance and Manhattan distance is small and is the difference is distributed throughout the study area, it can only be associated with difference in methods while using three different distance measures.

		Mahalanobis Distance								
		Clusters	1	2	3	4	5	6	7	8
Squared Euclidean Distance	1	274740	1014	0	0	0	0	407	0	
	2	0	28163	0	0	0	0	11206	0	
	3	0	0	204077	0	0	0	0	0	
	4	0	5478	19498	108631	0	908	0	0	
	5	4	473	0	0	31744	0	0	0	
	6	0	797	0	0	0	57158	913	0	
	7	0	2123	0	0	0	0	2962	0	
	8	0	1159	0	0	0	0	0	48545	
Kappa									0.929	

Table 6 Misclassification Matrix of Squared Euclidean Distance and Mahalanobis Distance

		Chebyshev Distance								
		Clusters	1	2	3	4	5	6	7	8
Squared Euclidean Distance	1	276161	0	0	0	0	0	0	0	
	2	0	39060	0	309	0	0	0	0	
	3	0	9801	194276	0	0	0	0	0	
	4	0	23440	0	111075	0	0	0	0	
	5	232	0	0	0	31989	0	0	0	
	6	6	237	0	670	0	57955	0	0	
	7	625	0	0	0	4	0	4456	0	
	8	0	0	0	0	0	0	4027	45677	
Kappa									0.9369	

Table 7 Misclassification Matrix of Squared Euclidean Distance and Chebyshev Distance

When squared Euclidean distance is compared with Mahalanobis and Chebyshev distance (Table 6 and 7), some difference in some cluster classification was observed. The Kappa coefficient of 0.929 for Mahalanobis distance and 0.9369 for Chebyshev distance represents rather big differences than previous comparison.

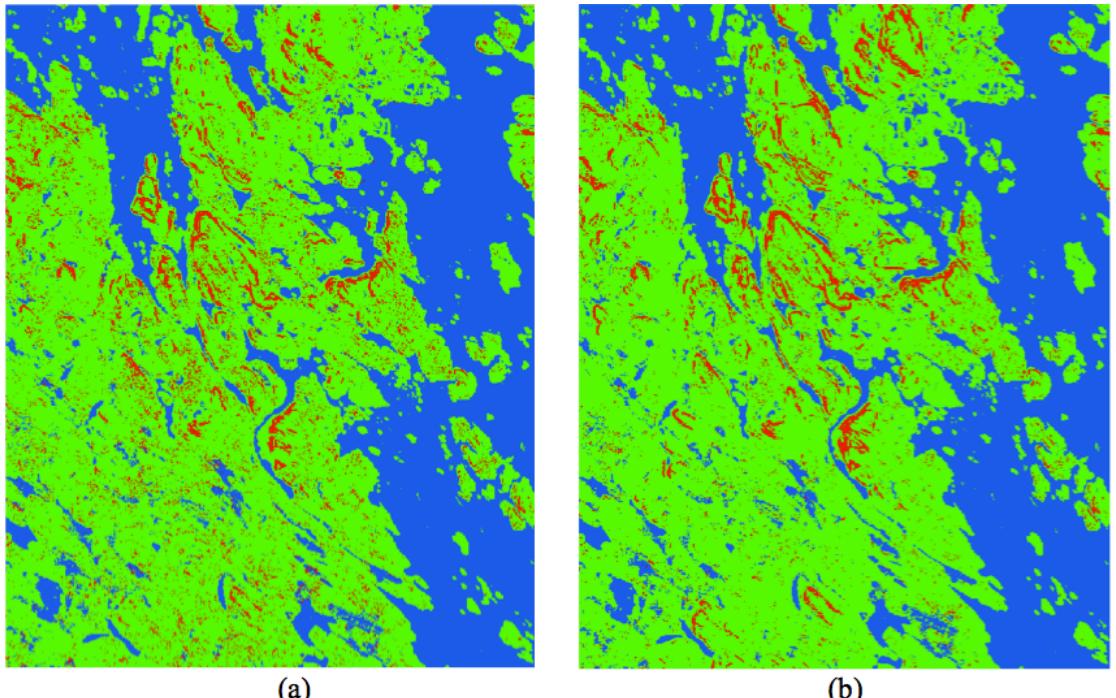


Figure 29 Difference between clusters created using different distance measures. (a) Between squared Euclidean and Mahalanobis distance and (b) Between squared Euclidean and Chebyshev distance.

When a map is created to visualize the difference (figure 29), the areas where difference exists are randomly distributed throughout the map area with red clusters appearing in some particular area. The difference was found in 5.49 % of study area between squared Euclidean and Mahalanobis distance, and in 4.91% of study area between squared Euclidean and Chebyshev distance. Due to aggregation of those differences in particular areas, it is further evaluated by comparing the difference with input layer. From comparison of result of squared Euclidean and Mahalanobis distance with input layers (figure 17), it was found that the areas of difference have particularly steep slope (more than 30 degrees), dense vegetation and rocky soil. Similar to result of Euclidean distance, the difference was attributed to Mahalanobis distance being particularly useful in determining correlation between the three input layers, which could be visible when compared with squared Euclidean distance. Similarly, for Chebyshev distance, when compared with normalized slope layer (figure 18), the higher value of mobility in particular area created the difference in result between squared Euclidean and Chebyshev distance.

	Mahalanobis Distance								
	Clusters	1	2	3	4	5	6	7	8
Manhattan Distance	1	274740	1014	0	0	0	0	625	0
	2	0	28579	0	0	0	0	11184	0
	3	0	0	204077	0	0	0	0	0
	4	0	5062	19498	108631	0	908	0	0
	5	4	1091	0	0	31744	0	0	0
	6	0	797	0	0	0	57158	935	0
	7	0	2664	0	0	0	0	2744	0
	8	0	0	0	0	0	0	0	48545
Kappa									0.9293

Table 8 Misclassification Matrix of Manhattan Distance and Mahalanobis Distance

	Chebyshev Distance								
	Clusters	1	2	3	4	5	6	7	8
Manhattan Distance	1	276368	0	0	0	0	0	11	0
	2	0	39038	0	725	0	0	0	0
	3	0	9801	194276	0	0	0	0	0
	4	0	23440	0	110659	0	0	0	0
	5	650	0	0	0	31993	0	196	0
	6	6	259	0	670	0	57955	0	0
	7	0	0	0	0	0	0	5408	0
	8	0	0	0	0	0	0	2868	45677
Kappa									0.9381

Table 9 Misclassification Matrix of Manhattan Distance and Chebyshev Distance

When Manhattan distance is compared with Mahalanobis and Chebyshev distance (Table 8 and 9), some difference in cluster classification could be observed. The Kappa coefficient of 0.9293 for Mahalanobis distance and 0.9381 for Chebyshev distance represents some differences between the distance measures. Although the difference is small, it is still a subject for further analysis.

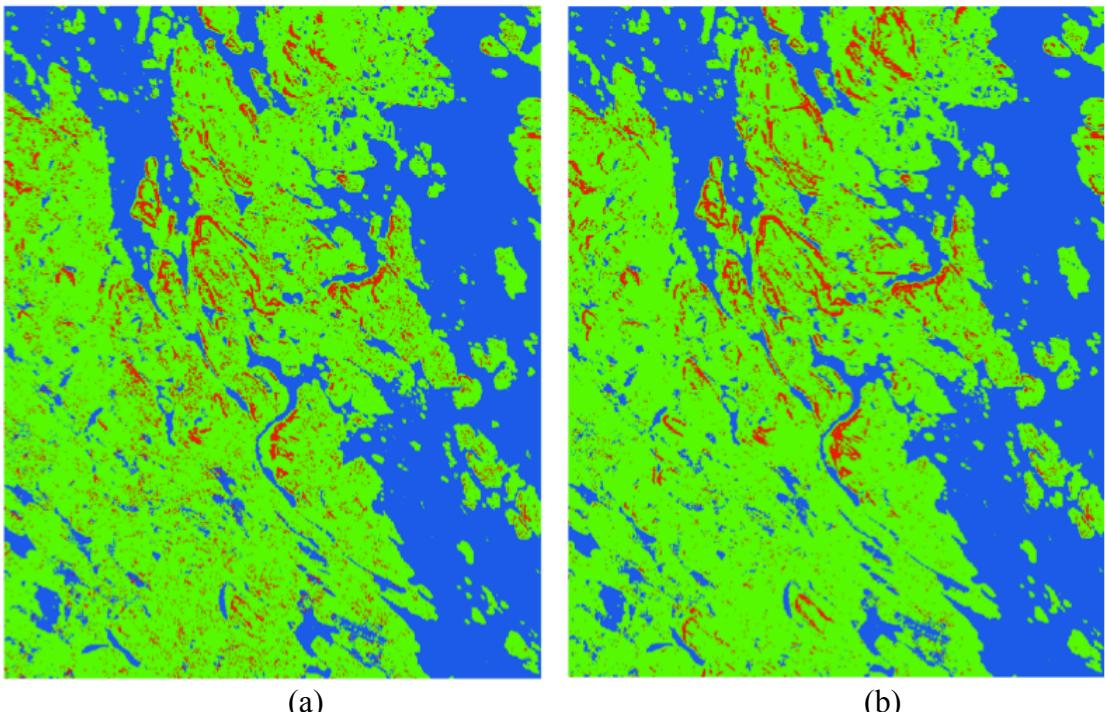


Figure 30 Difference between clusters created using different distance measures. (a) Between Manhattan and Mahalanobis distance and (b) Between Manhattan and Chebyshev distance.

The map is created to visualize the difference (figure 30) again represents that the areas where difference exists (red areas), are randomly distributed throughout the map area with red clusters appearing in some particular area. The difference was found in 5.47% of study area between Manhattan and Mahalanobis distance, and in 4.82% of study area between Manhattan and Chebyshev distance. On further comparison, similar conclusion about the difference is established as that of difference between Euclidean, Mahalanobis and Minkowski distance presented earlier in this chapter.

		Chebyshev Distance								
		Clusters	1	2	3	4	5	6	7	8
Mahalanobis Distance	1	274744	0	0	0	0	0	0	0	0
	2	1660	31605	0	2036	249	797	2860	0	
	3	0	29299	194276	0	0	0	0	0	
	4	0	191	0	108440	0	0	0	0	
	5	0	0	0	0	31744	0	0	0	
	6	0	0	0	908	0	57158	0	0	
	7	620	11443	0	670	0	0	2755	0	
	8	0	0	0	0	0	0	2868	45677	
	Kappa									0.9140

Table 10 Misclassification Matrix of Mahalanobis Distance and Chebyshev Distance

When Mahalanobis distance is compared with Chebyshev distance (Table 10), similar differences as previous could be visible. The Kappa coefficient of 0.9140 for represents rather big differences between two distance measures.

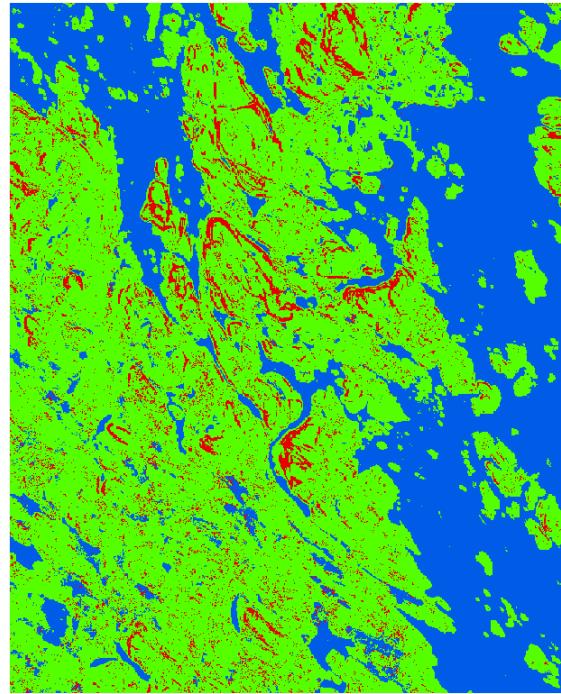


Figure 31 Difference between clusters created by Mahalanobis and Chebyshev distance.

Similarly, when the difference is visualized on a map (figure 31), the areas where difference exists (red areas) are randomly distributed throughout the map area with red clusters in some particular area. The difference was visualized in 6.7 % of study area between Mahalanobis distance and Chebyshev distance. The effect of input layers and its normalization can be viewed in the difference between results of Mahalanobis and Chebyshev distance.

Finally, from the comparison between different distance measures, it was revealed that Euclidean, squared Euclidean and Manhattan distance tend to create similar cluster output. The difference between clusters created by these three distance measures is rather small. On the other hand, for those areas where certain correlation between the input layers was found, like the areas with steep slope, dense vegetation and rocky soil, Mahalanobis distance could reveal better correlation hence creating slightly different clusters than that of previous three distance measures. In addition, Chebyshev distance was mostly affected by the normalization of input layer and tends to create the clusters where one particular layer has higher value of mobility in given area.

4.2.2 Cluster Validation

External indices of cluster validation are used in this study to validate results between different distance measures. The external indices uses misclassification matrix to calculate values of different indices and measures the similarity between different partitions. They take into account distribution of points in different clusters without measuring the quality of distribution. In this study, Jaccard coefficient, Rand Index and Folkes and Mallow index is used to interpret the result between cluster maps created by using different distance measures. Since, no reference dataset is available for comparing the clustering output, for each two different distance measures, three different external indices are calculated that represents how similar the result of two distance measures are (Table 11):

	Indices	Squared Euclidean Distance	Manhattan Distance	Mahalanobis Distance	Chebyshev Distance
Euclidean Distance	Jaccard	0.9953	0.9961	0.8897	0.9132
	Rand	0.9989	0.9991	0.9735	0.9799
	F&M index	0.9977	0.9980	0.9417	0.9548
Squared Euclidean Distance	Jaccard	-	0.9970	0.8923	0.91
	Rand	-	0.9993	0.9742	0.9791
	F&M index	-	0.9985	0.9431	0.9531
Manhattan Distance	Jaccard	-	-	0.8927	0.9104
	Rand	-	-	0.9743	0.9792
	F&M index	-	-	0.9433	0.9533
Mahalanobis Distance	Jaccard	-	-	-	0.8817
	Rand	-	-	-	0.9720
	F&M index	-	-	-	0.9374

Table 11 Comparison of external indices for different distance measures

For all external indices, values close to 1 represents perfect similarity between two clusters compared and values close to 0 represents dissimilar clusters. From Table 11, clusters created using Euclidean distance is most similar to cluster created using squared Euclidean distances with all cluster indices value of >0.99. Similarly, clusters created using Chebyshev distance is most dissimilar to that of Mahalanobis distance as all indices have smallest

values among different combination of distance measures. Table 11 reveals that no two clusters solutions are identical and there exists some difference between clusters produced by different distance measures. However, the overall difference is very small between different distance measures. The difference is localized in certain areas when Mahalanobis and Chebyshev distance is used as explained in section 4.2.1.

In Table 11, if cluster produced from Euclidean distance is considered as reference, clusters produced from squared Euclidean distance are the most similar and cluster produced with Mahalanobis distance are the least similar with it. Similarly, when Squared Euclidean distance is taken as reference, clusters produced from Manhattan distance is the most similar and cluster from Mahalanobis distance is the least similar based on the cluster indices value.

For determining the best distance measures among given five distance measures, each distance measures should be compared with a reference dataset to obtain the value of different cluster indices. In this study, due to unavailability of the reference dataset, to determine the best distance measures was not possible. Although, from the user knowledge and table 11, Euclidean distance could create the most similar cluster with other distance measures, thus, it could be considered better than other distance measures. However, only the comparison with the reference dataset would reveal the truth.

In addition, it could be assumed that for given dataset, where there is no correlation between different data layers, Euclidean distance would already provide good enough result of the analysis. Nevertheless, one should be aware of fact that for different dataset with different normalization condition, such assumption might not hold true. Thus, it is suggested to examine all different distance measures and select the result that best fits the purpose of analysis.

5 Discussion

This study presents collection of different types of clustering methods and its use to reveal important relationship from the spatial data. Particularly, this study analyzes the K-medoids method with different distance measures used to create clusters and its application in spatial analysis process.

K-medoid method used in analyzing different distance method has particular advantage than other clustering method like K-means or density based method, as it diminishes the sensitivity to outliers and could incorporate use of different distance measures. Unlike other methods, the cluster center is the representative object and allows good characterization of all clusters in the dataset. Although the calculation of distance between objects in K-medoid affects the time efficiency, it could significantly become efficient once the distance matrix is computed. The K-medoid method was used in interactive and iterative process during this study where several runs were required to obtain the result.

The main objective of this study is to provide insight to different distance measures that could be used for clustering of dataset. To determine how 'close' or how 'far' the observations are from one another is important for clustering as it affects the final result of clustering. Thus, selection of suitable distance measure is significant aspect to determine proximity between dataset and to reveal important information from the data. The choice of distance measures usually depends on type of data and purpose of analysis. For many types of dense, continuous data, a metric distance measure such as Euclidean distance or squared Euclidean distance is used. For continuous data, proximity is often expressed in terms of differences and distance measure provides a well-defined way of combining these differences into overall proximity measure. For time series data, use of Euclidean distance is justified and if the time series represented different quantities, the shape of time series is of more importance than magnitude, thus Mahalanobis distance is deemed useful in such case. Further, when the dataset consists of different shapes clusters, use of Minkowski distance with different values of 'p' is useful to determine different shape of clusters. In addition, if the dataset

consists both sparsely and dense data, dividing the dataset into subset and applying different distance measure to each subset might be useful.

Typically, in most research projects, Euclidean distance is the sole distance method used. Although it might be tempting and easy to use Euclidean distance, the choice of distance measure should be determined by the distribution of dataset, shape of clusters that could be present in the dataset, type of dataset that is being analyzed and purpose of analysis. Thus, for given dataset, different distance measures may need to be evaluated to determine which one produces the result that best suits the purpose of analysis.

Another particularly important factor is selection of 'k', the optimum number of clusters. As there is no standard procedure for selecting number of cluster, iterating the clustering process with different values of 'k' and evaluating each result individually by expert is useful.

The process used throughout the study is user-controlled and requires user/expert knowledge in different steps throughout the process. The use of expert knowledge started with selection of input layer, data preparation for computational analysis purpose and finally is the key to result interpretation. Particularly, the user knowledge was required during data normalization process to determine normalization parameters for the data. As the result of clustering is affected by the normalization of data, expert knowledge is of great significance. Additionally, in determining number of cluster, user knowledge is required. Finally, interpreting the result and decide whether the result is useful and served the purpose of analysis or not, expert knowledge was the key.

The result of this study indicates there exists certain level of difference between the clusters created using different distance measures. Although the clusters created by Euclidean, squared Euclidean and Manhattan distance had small differences between them, clusters created by Mahalanobis and Chebyshev distance exhibit some difference with that of other three distance measures. Mahalanobis distance was able to explore the correlation between the dataset in some sub areas; however, its effect in whole dataset was minimal. As correlation between the dataset is very small in such terrain analysis task and virtually the correlation is non-existent, Mahalanobis

distance could be used to visualize the areas with correlation. Similarly, Chebyshev distance favored normalized layer with higher value of mobility among the input layer. For example, when an area of dataset has value of 1, 4 and 5 in each of three input dataset, use of Chebyshev distance would result in creation of cluster with value 5. Thus, Chebyshev distance reveals the dimension, which is most similar to other.

5.1 Challenges

The major challenges faced during the study were attributed to unavailability of reference dataset and lack of proper visualization tool.

The unavailability of reference dataset made the result interpretation particularly ambiguous thus depends on user knowledge for interpretation. As, in many of the spatial analysis process, which usually lacks the reference dataset to determine uncertainty of result produced, alternative method was devised to investigate difference between results of different distance measures. With unavailability of verified source to compare the validity of result, the results were compared within themselves to determine the differences in clustering. Although this process was able to determine difference between clusters, comparison with reference dataset would have provided concrete evidence for visualizing difference differences between the clustering outputs. Further, it would have been possible to rank different distance measures from best to worst based on its comparison with the reference dataset.

Another challenge faced during study was lack of proper visualization tool. For conversion of cluster map to mobility map, there has to be link between different clusters and its corresponding location in the map in order to assign mobility value for a given cluster. To assign the mobility value for a cluster, the cluster should be linked to the topographic map area by the help of PCP. With lack of available tools for such view, it was not possible to assign particular cluster a mobility value and hence not possible to create the mobility map. Since, the application of cluster analysis in this study was to create a mobility map, due to lack of proper tool, the final product of mobility analysis could not be delivered.

5.2 Future Research

The underlying hypothesis for using different distance measures on K-medoids clustering was there exists certain difference between results of clustering with different distance measures. The idea was tested with K-medoid clustering method and the result revealed that there is some difference between clusters created with different distance measures. However, this idea requires to be tested on different dataset and on different region in order to be considered as profound theorem.

Moreover, dividing the dataset into small subset and applying different distance measure on individual subset based on distribution of dataset might reveal some useful information about whole dataset. Through division of dataset into small subsets, each subset could have different shaped clusters or some correlation, which could be revealed by applying different distance measures.

Further, this study was conducted without the availability of reference dataset to which final result could be compared. Thus, testing of the results with reference dataset might reveal additional differences on the result.

In addition, there are other distance measures that could well be incorporated in clustering like Hamming distance, Cosine distance, Jaccard distance etc. Use of such distance measures might provide additional insight to the clustering result.

Furthermore, another important direction of research could be towards development of toolkit that could link the topographic map and PCP of the cluster result. At the moment, the toolkit for such visualization is limited to a prototype or within certain researcher community that is problem specific. Creation of general toolkit, where different visualization methods are linked with each other would reveal additional information about the dataset and aid to result interpretation as well as decision making process.

Finally, the findings of the study accompanied with other research could lead to creation of efficient method to incorporate clustering to mobility analysis of a terrain, visualization method and analysis process that can facilitate knowledge discovery from spatial data.

6 Conclusion

Due to distinct characteristics of spatial data, traditional techniques should be modified and used in different fields in order to reveal useful patterns, associations and other important information in the dataset. In this study, use of different distance method for K-medoids clustering were explored and highlighted with focus on its implementation on mobility analysis. Also, the study attempts to reveal the differences in output of K-medoid clustering using different distance measures.

The research answers the research question "*How the use of different distance measures affects the result of spatial analysis in clustering?*" by comparing different distance measures and validating the result. Although the difference between the results of clusters created using different distance measures was relatively very small, and could be associated with the distribution of data values in the input layers as well as normalization of data layers, this study provides new viewpoint for applying cluster analysis and focuses on fact that use of different distance measures has some affect on the final result unless further research is performed to provide substantial evidence against the statement.

Another research question to be answered by this study was "*Can a distance measure provide proper insight about similarity of the cluster and reveal useful information?*". To analyze the answer of above question, one must understand that the result of clustering depends on type of data, normalization of dataset, selection of optimum number of clusters and expert knowledge for interpreting the output. Apart from that, understanding different distance measure is essential factor for a clustering method to reveal useful information. For example, Mahalanobis distance is affected by the correlation between the data. So, for data with certain degree of correlation, Mahalanobis distance produces better clusters than that of Euclidean distance. Thus, it is essential to understand what kind of cluster each distance measure is able to produce in order to use it for the cluster analysis. With good knowledge about each distance measures and with understanding of different kind of applicable dataset, good cluster result could be produce, which in turn is able to reveal useful information.

Further, the interpretation of clustering output is subject of expert evaluation for revealing useful information. The role of user is essential for understanding overall process and to create meaningful representation from the output. Although, due to lack of proper tool and reference dataset, the clusters could not be assigned a mobility value to create the mobility map, there exists a theoretical potential to obtain useful information from the data. Finally, this study also points out most important steps of cluster analysis from data processing, data analysis, result interpretation and validation so as to enable the user the most suitable approach for cluster analysis and its further implementation.

References

- Abonyi, J. & Feil, B., 2000. *Cluster Analysis for Data Mining and System Identification*. Basel: Birkhäuser.
- Agarwal, R. & Sahoo, L., 2008. *Vlsi Technology And Design*. Pune: Technical Publications.
- Alarcon-Aquino, V. et al., 2014. Biometric Cryptosystem based on Keystroke Dynamics and K-medoids. *IETE Journal of Research*, 57(4), pp.384-94.
- Aldenderfer, M., 1991. *Cluster Analysis*. London: Sage.
- Backer, E., 1995. *Computer Assisted Reasoning in Cluster Analysis*. Hertfordshire: Prentice Hall.
- Borruso, G., 2008. Network Density Estimation: A GIS Approach for Analysing Point Patterns in a Network Space. *Transactions in GIS*, 13(3), pp.377-402.
- Camila, M. et al., 2008. Accelerating k-medoid-based algorithms through metric access methods. *The Journal of Systems and Software*, 81, pp.343-55.
- Cichosz, P., 2015. *Data Mining Algorithms: Explained Using R*. Chichester: Wiley.
- Dalfo, C., Comellas, F. & Fiol, M.A., 2007. The Multidimensional Manhattan Network. *Electronic Notes in Discrete Mathematics*, 29, pp.383-87.
- Davis, J. et al., 1991. Testing of soil moisture prediction model for army land managers. *Journal of Irrigation and Drainage Engineering*, 117(4), pp.476-89.
- Demaj, D., 2013. Geovisualizing spatio-temporal patterns in tennis: An alternative approach to post-match analysis. Available at: http://gamesetmap.com/?page_id=2.
- Desgraupes, B., 2013. *Clustering Indices*. [Online] Available at: <ftp://apache.cs.uu.nl/mirror/CRAN/web/packages/clusterCrit/vignettes/clusterCrit.pdf> [Accessed 29 February 2015].
- Dillmann, R., Beyerer, J., Hanebeck, J.D. & Schultz, T., 2010. KI 2010: Advances in Artificial Intelligence. In *33rd Annual German Conference on AI*. Karlsruhe, 2010. Springer.
- Ding, W. et al., 2009. Discovery of feature-based hot spots using supervised clustering. *Computers and Geosciences*, 35, pp.1508-16.
- Dong, S., Zhou, D., Ding, W. & Gong, J., 2013. Flow Cluster Algorithm Based on Improved K-means Method. *IETE JOURNAL OF RESEARCH*, 59(4), pp.326-33.
- Everitt, B. & Hothorn, T., 2011. *An Introduction to Applied Multivariate Analysis with R*. New York: Springer.
- Everitt, B., 2011. *Cluster Analysis*. Chichester, West Sussex, U.K.: Wiley.
- Fabrikant, S.I. & Montello, D.R., 2008. The effect of instructions on distance and similarity judgements in information spatializations. *International Journal of Geographical Information Science*, 22(4), pp.463-78.
- Filipe, J. & Cordeiro, J., 2011. Enterprise Information Systems. In *12th International Conference, ICEIS 2010*. London, 2011. Springer.
- Gore Jr., P.A., 2000. Cluster Analysis. In *Handbook of Applied Multivariate Statistics and Mathematical Modelling*. Academic press.

- Gorsevski, P.V., Jankowski, P. & Gessler, P.E., 2005. Spatial Prediction of Landslide Hazard Using Fuzzy k -means and Dempster-Shafer Theory. *Transactions in GIS*, 9(4), pp.455-74.
- Groenen, P.J.F. & Jajuga, K., 2001. Fuzzy clustering with squared Minkowski distances. *Fuzzy Sets and Systems*, 120, pp.227-37.
- Hair, J.F. et al., 2006. *Multivariate Data Analysis*. New Jersey: Pearson Prentice Hall.
- Halkidi, M., Batistakis, Y. & Vazirgiannis, M., 2001. On Cluster Validation Techniques. *Journal of Intelligent Information Systems*, pp.107-45.
- Hall, A., Ahonen-Rainio, P. & Virrantaus, K., 2014. Knowledge and Reasoning in Spatial Analysis. *Transactions in GIS*, 18(3), pp.464-67.
- Hand, D., Mannila, H. & Smyth, P., 2001. *Principles of Data Mining*. Cambridge: MIT Press.
- Jean-Daniel, F., 2004. The InfoVis Toolkit. In *10th IEEE Symposium on Information Visualization*., 2004. IEEE Press.
- Jung, A., Fenwick, D.H. & Caers, J., 2013. Training image-based scenario modeling of fractured reservoirs for flow uncertainty quantification. *Computational Geoscience*, 17, pp.1015–31. DOI 10.1007/s10596-013-9372-0.
- Kaufman, L. & Rousseeuw, P.J., 1990. *Finding Groups in Data*. New Jersey, USA: Wiley.
- Maesschalck, R.D., Jouan-Rimbaud, D. & Massart, D.L., 2000. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, pp.1-18.
- Mahalanobis, P.C., 1936. *Digital Library of India*. [Online] Available at: http://www.new.dli.ernet.in/rawdataupload/upload/insa/INSA_1/20006193_49.pdf [Accessed 27 February 2015].
- Manchester Metropoliton University, . *Dendrogram*. [Online] Available at: <http://www.alanfielding.co.uk/multivar/dend.htm> [Accessed 28 February 2015].
- Manly, B.F.J., 1986. *Multivariate Statistical Method*. London: Chapman and Hall. Available at: <http://matlabdatamining.blogspot.fi/2006/11/mahalanobis-distance.html> [accessed 27 February 2015].
- Mcintosh, J. & Yuan, M., 2005. Assessing Similarity of Geographic Processes and Events. *Transactions in GIS*, 9(2), pp.223-45.
- Miller, H.J. & Han, J., 2001. *Geographic Data Mining and Knowledge Discovery: an overview*. London: Taylor and Francis.
- Montello, D.R., Fabrikant, S.I., Ruocco, M. & Middleton, R.S., 2003. Testing the first law of cognitive geography on point-spatialization displays. In *Spatial Information Theory: Foundations of Geographic Information Science, Conference on Spatial Information Theory (COSIT)*. Berlin, 2003. Springer.
- Morales-Esteban , A., Martínez-Álvarez , F. & Scitovski, S., 2014. A fast partitioning algorithm using adaptive Mahalanobis clustering with application to seismic zoning. *Computers & Geosciences* , 73, pp.132-41.
- Nikander, J., 2012. *Interaction and Visualization Methods in Teaching Spatial Algorithms and Analyzing Spatial Data*. Espoo: Aalto University.

- Nikander, J., Kantola, T. & Virrantaus, K., 2012. Exploratory vs. Model-Based Mobility Analysis. *Nordic Journal of Surveying and Real Estate Research*, 9(1), pp.7-29.
- Park, H.-S. & Jun, C.-H., 2009. A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, 36, pp.3336–41.
- Patentdocs, 2011. *Grid based Clustering method*. [Online] Available at: http://www.faqs.org/patents/imgfull/20110040758_03 [Accessed 28 February 2015].
- Pollard, D., 1981. STRONG CONSISTENCY OF K-MEANS CLUSTERING'. *The Annals of Statistics*, 9, pp.135-40.
- Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, pp.53-65.
- Sage Publications, 2008. *Cluster Analysis*. [Online] Available at: [20Cluster%20Analysis.pdf](http://www.uk.sagepub.com/burns/website%20material/Chapter%2023%20-%20Cluster%20Analysis.pdf) <http://www.uk.sagepub.com/burns/website%20material/Chapter%2023%20-%20Cluster%20Analysis.pdf> [Accessed 07 February 2015].
- Sood, M. & Bansal, S., 2013. K-Medoids Clustering Technique using Bat Algorithm. *International Journal of Applied Information Systems (IJAIS)*, 5(8).
- Tan, P.-N., Steinbacch, M. & Kumar, V., 2006. *Introduction to Data Mining*. Boston: Pearson Education Inc.
- Theodoridis, S. & Koutroumbas , K., 2003. *Pattern Recognition*. Amsterdam: Elsevier/Academic Press.
- Tobler, W.R., 1970. A computer movie simulating population growth in the Detroit region. *Economic Geography*, 42, pp.234-40.
- University of Florida, 2015. *Knowledge Discovery in Databases*. [Online] Available at: <http://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/KDD3.htm> [Accessed 12 February 2015].
- University of Texas, 2000. *Cluster Analysis*. [Online] Available at: <http://www.uta.edu/faculty/sawasthi/Statistics/stcluan.html> [Accessed 27 June 2015].
- Webb, A., 1999. *Statistical Pattern Recognition*. London: Arnold Publishers.
- Wikia, 2013. *Computer Vision*. [Online] Available at: http://computervision.wikia.com/wiki/Manhattan_distance [Accessed 07 February 2015].
- Wiktionary, 2013. *Manhattan Distance*. [Online] Available at: http://en.wiktionary.org/wiki/Manhattan_distance [Accessed 27 February 2015].
- Wilmer, W.J. et al., 2008. The influence of multiple dispersal mechanisms and landscape structure on population clustering and connectivity in fragmented artesian spring snail populations. *Molecular Ecology*, 17, pp.3733-51.
- Wire, 2012. Data Mining Knowledge Discovery. WIRE. pp.209–25.
- Worboys, M.F., 2001. Nearness relations in environmental space. *International journal of geographical information science*, 15(7), pp.633-51.

Zadegan, S.M.R., Mirzaie, M. & Sadoughi, F., 2013. Ranked k-medoids: A fast and accurate rank-based partitioning algorithm for clustering large datasets. *Knowledge-Based Systems*, 39, pp.133-43.

Zhai, Q., Yang, J., Xie, M. & Zhao, Y., 2014. Generalized moment-independent importance measures based on Minkowski distance. *European Journal of Operational Research*, 239, pp.449-55.

Appendix 1

Pseudo Code of K-means Algorithm

Input:

- k : number of cluster
- D : a dataset containing n objects

Output: A set of K clusters

Method:

- Arbitrarily choose k objects from D as the initial cluster centers;
- Repeat;
- (Re) assign each object to the cluster to which the object is the most similar based on the distance between the object and the cluster mean;
- Update the cluster mean by calculating the mean value of the objects for each cluster;
- Until no change;

Appendix 2

Pseudo Code of PAM algorithm

Input:

- k : number of cluster
- D : a dataset containing n objects

Output: A set of K clusters

Method:

- Arbitrarily choose k objects from D as the initial cluster centers;
- Repeat;
- For each non-representative object O_h do;
- For each representative object O_i do;
- Calculate the total cost TC_{ih} of swapping between O_i and O_h ;
- Find i and h where TC_{ih} is the smallest;
- If $TC_{ih} < 0$ then replace O_i with O_h ;
- Until $TC_{ih} \geq 0$;
- Assign each non-representative object to the cluster with the nearest representative object;

Appendix 3

Normalization of slope data

Slope Value	Classified Values
0-10	10
11-20	7
21-30	3
31-41	1

Normalization of Vegetation data

Vegetation diameter value	Classified Value
0	10
1	8
2	6
3	4
4	3
5	2
6	2
7 - 9	1
10 - 40	0

Normalization of Soil Type

Soil Type	Classified Value
3	0
4	3
5	1
6	0
11	0
12	3
13	5
14	5
15	1
16	8
17	4
18	5
19	6
20	8
21	0
22	0