CHAPTER-III

METHODOLOGY

3.1 Introduction

The methodology of the study includes preparation of Aizawl city base map and delineation of boundaries of local councils and municipal wards, determination of sample size, operation of sampling procedure for collection of data and analysis of tabulated data with the help of statistical techniques and graphical methods. The outputs of the analyses were mapped with the help of choropleth mapping techniques wherever appropriate.

3.2 Selection of Sampling Unit

Firstly, a base map of the Aizawl city was prepared with the help of ArcGIS software. Boundaries of Local Council (LC) areas were delineated. The local council is the lowest unit of local administration and lowest unit of enumeration. Each local council is represented by 5 to 7 council members. In local parlance, each local council area is known as veng which may be translated as ‘locality’ or ‘neighbourhood’ where residents are bound by unity and integrity. Normally, veng boundaries coincide with the boundaries of one branch of powerful community organization called Young Mizo Association (YMA). Altogether, there were 82 local councils in the entire city of Aizawl in 2011. These local councils come under 19 municipal wards. Normally, a municipal ward consists of 3 to 8 local councils.

Local councils are considered as suitable units of spatial analysis due to their small size, cohesiveness and presence of sense of belongingness among its residents. From methodological point of view, small areas are appropriate sites of study and it
has been encouraged that studies on intra-urban difference should be taken out at smaller areas since patterns or degrees of inequality are more visible and the criticism of ecological fallacy is diminished at lower levels of aggregation (UNCHS, 2000; UN-HABITAT, 2003).

3.3 Sampling and Sample Size

Collection of precise and reliable data requires careful and judicious selection of sampling technique. Various literatures were surveyed before taking sample to determine sample size and sampling method appropriate for the present study. Das (2008) applied two-stage stratified purposive sampling method. In the first stage, 6 municipal wards were selected out of 60 municipal wards purposively to represent various wards of the city. In the second stage, households were picked selectively from each ward to represent various income groups. In each ward, 3 per cent of the total households were interviewed to make a total of 379 samples. Another sampling methods employed in the previous studies were random stratified sampling (Fakhruddin, 1991; Turkoglu et al., 2011), purposive, stratified and systematic sampling methods (Tesfazghi et al., 2010), multi-stage random sampling method (Oktay and Rustemli, 2011), purposive stratified sampling method (Mridha and Moore, 2011), and stratified random aligned traverse sampling method (Berry and Baker, 1968; Omuta, 1988).

In the present study, random stratified sampling method was employed to generate a sample of households for the entire city of Aizawl consisting of 82 local councils (LCs) belonging to 19 Municipal Wards. Since the sampling unit is very vast, a sample household of 5 per cent from the total household constitutes the sample size for each locality. In the end, a total of 1,600 schedules were supplied to the
respondents and face-to-face interviews were conducted at their residences. Each schedule contained an information sheet mentioning that the identity of the respondents and his/her family should not be revealed and were free to decline answering the questions.

Before taking actual survey, a pilot survey was undertaken first to determine appropriate variables to be included in the study. Those features which were poorly understood, difficult to interpret and were of little or no importance to the public were excluded from further analysis. Thereafter, a schedule was developed based on previous studies but modified whenever necessary to meet the specific purpose of the study. The schedule was designed to collect objective parameters for measuring residential pattern and quality of life as well as to measure residents’ perceptions and their evaluations about aspects of quality of life. Thus, it consists of two parts—the first part contains objective measures and the second part contains subjective measures. The final schedule comprised of 46 questions.

Selection of households was made to represent all sections of population within each local council unit. This includes arbitrary division of each locality into valley, slope and hill top bases on observation. Households were selected from each type of topography in proportionate to the total number of houses. Sometimes, stratification of local council was based on the existing division of Young Mizo Association’s (YMA) branch which coincided with boundary of local council. Each YMA branch or veng is divided into a number of sections. These sections may also be taken suitably as strata in the sampling.

Objective questions provide ratio scale data which could be readily analyzed with any statistics. Regarding subjective questions, responses to each subjective
question were measured on a linear numeric version of a Likert-type scale. Whereas the traditional Likert-type scale produces only ordinal data and is thus inappropriate for parametric statistics, linear-numeric scales lead to equal-interval data that may be analyzed using the most powerful parametric statistics (Alreck and Settle, 1995). For the majority of the questions, respondents were requested to place a tick in one of the five boxes to indicate their level of satisfaction with each item on a five-point linear numeric version of a Likert scale, ‘1’ standing for strong level of dissatisfaction and ‘5’ representing a strong level of satisfaction. For instance, question such as ‘How satisfied are you with the quality of roads in your local council area?’ has to be answered on a 5-point scale from 1 = ‘very dissatisfied’ to 5 = ‘very satisfied’. To get reliable and correct information, only respondents above 18 years of age were selected. The response rate was around 98 per cent since interviews were usually conducted at the respondents’ residents during holidays and at nights. Night-time visit is common in the study area. Only those households without any available person during the visits were skipped.

3.4 Techniques of Analysis

A number of quantitative techniques and graphical methods are used in the present study. The main techniques are multivariate data analysis techniques like factor analysis (FA), principal components analysis (PCA) and cluster analysis (CA); measures of association like correlation and regression; spatial autocorrelation techniques like global Moran’s I and local indicators of spatial association (LISA). Graphical methods like scatter plots, line graphs, histograms and choropleth maps are also used.
Factor analysis and principal component analysis are powerful multivariate techniques. The goal of research using PCA or FA is to reduce a large number of variables to a smaller number of factors, to concisely describe the relationships among observed variables, or to test theory about underlying processes (Tabachnick and Ridell, 2013). These techniques are data reduction methods that derive a composite, smaller set of correlated but independent variables known as factors or components from a large set of variables. Each of the factors or components may be thought as a ‘super variable’.

PCA can be understood as a special case of factor analysis (FA) and is usually, but not always, the first step in a FA. Both of them are almost similar except in preparation of the observed correlation matrix for extraction and in the underlying theory. Mathematically, the difference between PCA and FA is in the variance that is analyzed. In PCA, all the variances in the observed variables are analyzed. In FA, only shared variance is analyzed and attempts are made to estimate and eliminate variance due to error and variance that is unique to each variable. In PCA, components are a function of the measured variables where as in FA; the measured variables are a function of factors. Second, FA estimates errors which are unique variance while PCA assumes that the measurement is without error.

FA and PCA are the most preferred techniques for measuring urban socio-spatial differentiation (Knox and Pinch, 2010). Specifically, PCA is a preferred method for data reduction while FA is a preferred method for detecting structure (Krishnan, 2010). Tabachnick and Fidell (2013) recommend the use of FA if the researcher is interested in a theoretical solution without error variability or without a unique mathematical solution. However, if the researcher needs an empirical
summary of the data set that explains the maximum variance with a unique mathematical solution, then PCA is preferred.

PCA is, therefore, employed here to develop composite index of various dimensions of quality of life. FA, on the other hand, is employed to detect the underlying structure of residential differentiation or in other words to extract the main axes of urban social differentiation (Robson, 1969).

3.4.1 Principal Component Analysis

In PCA, an original set of variables is transformed into a fewer new set of orthogonal (uncorrelated) variables called principal components. Mathematically, principal components are linear combinations of variables with weights in terms of their eigen vectors. These eigen vectors are derived from the correlation matrix of the variables. Each principal component is a linear combination of Z’s obtained as

\[ Z_1 = \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1q}x_q \]
\[ Z_2 = \alpha_{21}x_1 + \alpha_{22}x_2 + \cdots + \alpha_{2q}x_q \]
\[ \vdots \]
\[ Z_q = \alpha_{q1}x_1 + \alpha_{q2}x_2 + \cdots + \alpha_{qm}x_m \]

Where \( x_1, x_2, \ldots, x_q \) are the variables (indicators), \( q \) the number of variables and \( Z_i \ (i=1,\ldots,q) \) represents the principal components. \( a_{ij} \) are the component loadings which are chosen as weights applied to the variables \( x_j \) in equation(1) so that the principal component \( Z_i \) satisfy the following conditions:

(i) they are uncorrelated (orthogonal);

(ii) the first principal component accounts for the maximum possible proportion of the variance of the set of \( x_s \), the second principal component accounts for the maximum of the remaining variance, and so on until the last of the principal
components absorbs all the remaining variance not accounted for by the preceding components, and
\[ a_{11}^2 + a_{12}^2 + \ldots + a_{1q}^2 = 1 \]
where \( i = 1, 2, \ldots, q \).

PCA involves finding the eigen values \( \lambda_j \), where \( j = 1, 2, \ldots, q \), of the sample covariance matrix \( C_M \) as

\[
C_M = \begin{bmatrix}
    c_{m11} & c_{m12} & \ldots & c_{m1q} \\
    c_{m21} & c_{m22} & \ldots & c_{m2q} \\
    \vdots & \vdots & \ddots & \vdots \\
    c_{mq1} & c_{mq2} & \ldots & c_{mqq}
\end{bmatrix}
\]

Where the diagonal element \( cm_{ii} \) is the variance of \( x_i \) and \( cm_{ij} \) is the covariance of variables \( x_i \) and \( x_j \). The eigenvalues of the matrix \( C_M \) are the variances of the principal components and can be found by solving the characteristic equation where \( I \) is the identity matrix and \( \lambda \) is the vectors of eigenvalues.

An important property of the eigenvalues is that they add up to the sum of the diagonal elements of \( C_M \). That is, the sum of the variances of the principal components is equal to the sum of the variances of the original variables:

\[ \lambda_1 + \lambda_2 + \ldots + \lambda_q = cm_{11} + cm_{22} + \ldots + cm_{qq} \]

In order to prevent some variables having undue influences on the principal components, variables are standardized first to have zero means and unit variances at the start of the analysis. The co-variance matrix \( C_M \) then takes the form of the correlation matrix. Given that the correlation matrix rather than the covariance matrix
is used in the PCA, all individual indicators are assigned equal weights in forming the principal components (Chatfield and Collins, 1980)

3.4.2 Factor Analysis

The procedure to construct factor analysis (FA) is similar to PCA. However, while PCA is based simply on linear data combinations, the FA model assumes that the data is based on the underlying factors of the model, and that the data variance can be decomposed into that accounted for by common and unique factors.

The mathematical equation of FA may be given as

\[
\begin{align*}
X_1 &= \alpha_{11}F_1 + \alpha_{12}F_2 + \cdots + \alpha_{1m}F_m + e_1 \\
X_2 &= \alpha_{21}F_1 + \alpha_{22}F_2 + \cdots + \alpha_{2m}F_m + e_2 \\
&\quad \cdots \\
X_q &= \alpha_{q1}F_1 + \alpha_{q2}F_2 + \cdots + \alpha_{qm}F_m + e_q
\end{align*}
\]

Where \(X_i\) \((i = 1, \ldots, q)\) represents the original variables but standardized with zero mean and unit variance; \(\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{im}\) are the factor loadings related to the variable \(X_i\); \(F_1, F_2, \ldots, F_m\) are \(m\) uncorrelated common factors, each with zero mean and unit variance; and \(e_i\) are the \(q\) specific factors supposed independently and identically distributed with zero mean.

The procedures and applications of FA and PCA in geographical analysis are given Figures 3.1 and 3.2.
Figure 3.1 Schematic Diagram of Procedures of Factor Analysis and Principal Component Analysis (Source: Goddard and Kirby, 1976).

Figure 3.2 Applications of Factor Analysis and Principal Component Analysis (modified from Demsar et al., 2001).
3.4.3 Steps in Factor Analysis

According to Ho (2014), there are three basic steps to FA including PCA. They are 1) computation of the correlation matrix for all variables 2) extraction of initial factors and 3) rotation of the extracted factors to a terminal solution.

1) Computation of correlation matrix

Factor analysis is based on correlations between measured variables. Extreme multicollinearity is not permitted to conduct FA as this would cause difficulties in determining the unique contribution of the variables to a factor (Field, 2000:444). A Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy is a statistic for comparing the magnitudes of the observed correlation coefficients. The concept is that the partial correlations should not be very large if distinct factors are expected to emerge from factor analysis (Hutcheson and Sofroniou, 1999).

The Kaiser-Meyer-Olkin (KMO) statistic is computed for each individual indicator, and their sum is the KMO overall statistic. The range of KMO value varies from 0 to 1. A KMO overall should be .60 or higher to proceed with factor analysis (Kaiser and Rice, 1974). Multicollinearity can also be detected via the determinant of the correlation matrix. If the determinant is greater than 0.00001, then there is no multicollinearity (Field 2000).

\[ KMO = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} r_{ij}^2}{\sum_{i=1}^{n} \sum_{j=1}^{n} r_{ij}^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij}^2} \]

where \( r_{ij} \) is the correlation coefficient of variable \( i \) and variable \( j \), and \( s_{ij} \) is the partial correlation coefficient.
The Bartlett’s test of sphericity is used to test the null hypothesis that the individual indicators in a correlation matrix are uncorrelated, i.e. that the correlation matrix is an identity matrix. The statistic is based on a chi-squared transformation of the determinant of the correlation matrix.

2) Extraction of factors

The next step is to extract factors (or components in case of PCA) which are simply aggregates of correlated variables. According to Tabachnick and Fidell (2014), for something to be labeled as a factor it should have at least 3 variables. A factor with 2 variables is only considered reliable when the variables are highly correlated with each another (r > 0.70) but fairly uncorrelated with other variables (Yong and Pearce, 2013).

To extract factors or components, a number of methods are available. The choice of extraction method depends on the nature of research undertaken.

Principal component analysis (PCA) and common factor analysis (CFA) are two basic methods for obtaining factor solutions. Under common factor analysis model, there are six methods of extraction including principal-axis factoring, unweighted least-squares, generalized least-squares, maximum-likelihood, alpha factoring, and image factoring. Among the above methods, Principal Axes solution (or Principal Axis factoring) and PCA are the two most common extraction methods in geography (Clark, Davies and Johnston, 1974).

In the present study, PCA is used to determine composite index of quality of life (QOL). The method is suggested when the purpose of the study is no more than to reduce data in order to obtain the minimum number of factors needed to represent the
original set of data (Ho, 2014). The method is preferred to other methods in the construction of composite index of various dimensions of QOL as all the variances in the observed variables are analyzed.

On the other hand, factor analysis, particularly principal axis factor (PAF) method is employed in the study of residential pattern particularly horizontal pattern of residential differentiation. The PAF method is based on the notion that all variables belong to the first group and when the factor is extracted, a residual matrix is calculated. Factors are then extracted successively until there is a large enough of variance accounted for in the correlation matrix (Tucker and MacCallum, 1997).

3) **Determination of number of factors**

There are two conventional criteria for determining the number of initial unrotated factors to be extracted. These are the eigenvalues criterion and the scree test criterion. The eigenvalues criterion is also called Kaiser’s criterion which suggests retaining all factors that are above the eigenvalue of 1 (Kaiser, 1970). The Scree test rule is based on a visual plot of the eigenvalues against the number of factors in their order of extraction. In a Scree test, factors located above the break (i.e. point of inflexion) are retained. Sometimes, parallel Analysis is also recommended to extract reliable number of factors (Zwick and Velicer, 1986; Streiner, 1998; O’Connor, 2000). In parallel analysis, the eigenvalues derived from the actual data are compared to the eigenvalues derived from the random data sets. Factors are retained as long as the $i^{th}$ eigenvalue from the actual data is greater than the $i^{th}$ eigenvalue from the random data. The Kaiser’s rule is followed in the present analysis.
4) **Rotation methods**

Factors are rotated for better interpretation since unrotated factors are ambiguous. The goal of rotation is to attain an optimal simple structure which attempts to have each variable load on as few factors as possible, but maximizes the number of high loadings on each variable (Rummel, 1970).

The most commonly used methods in factorial ecology are orthogonal (i.e. uncorrelated) and oblique rotation procedures. Orthogonal rotation is when the factors are rotated 90° from each other, and it is assumed that the factors are uncorrelated (Rummel, 1970). On the other hand, oblique rotation is when the factors are not rotated 90° from each other, and the factors are considered to be correlated. The difference between orthogonal and oblique rotation is, however, inconclusive (Giggs and Mather, 1975; Costello and Osborne, 2005).

The choice between orthogonal and oblique rotations depends on the purpose of the study. According to Ho (2014:206) “If the goal of the research is no more than to ‘reduce the data’ to more manageable proportions, regardless of how meaningful the resulting factors may be, and if there is reason to assume that the factors are uncorrelated, then orthogonal rotation should be used. Conversely, if the goal of the research is to discover theoretically meaningful factors, and if there are theoretical reasons to assume that the factors will be correlated, then oblique rotation is appropriate”. Therefore, one orthogonal rotation method ‘varimax’ is used in PCA and ‘direct oblimin’, an oblique rotation method is used in factor analysis.
Factor scores are composite (latent) scores for each subject on each factor (Wells, 1999; Thompson, 2004). They are analogous to the Ŷ scores in the regression equation and are calculated by applying the factor pattern matrix to the measured variables.

The factor score produced with the help of Bartlett estimation method in statistical software called ‘IBM SPSS’ is used for mapping the score of each observation in the factorial analysis of residential pattern. The method uses the least squared procedure to minimize the sums of squares of the factors over the range of variables (Bartlett, 1937). In Bartlett method, only the common factors have an impact on factor scores. The sum of squared components for the ‘error’ factors (i.e., unique factors) across the set of variables is minimized, and resulting factor scores are highly correlated to their corresponding factor and not with other factors. However, the estimated factor scores between different factors may still correlate.

Bartlett factor scores are computed by multiplying the row vector of observed variables, by the inverse of the diagonal matrix of variances of the unique factor scores, and the factor pattern matrix of loadings. Resulting values are then multiplied by the inverse of the matrix product of the matrices of factor loadings and the inverse of the diagonal matrix of variances of the unique factor scores. One advantage of Bartlett factor scores over the other two methods i.e Regression method and Anderson-Rubin method is that this procedure produces unbiased estimates of the true factor scores (Hershberger, 2005).
3.4.4 Construction of Weights using Principal Component Analysis

Principal component analysis and factor analysis can be used to weight and aggregate variables in a composite index. An advantage of these methods is that they require no *a priori* assumptions on the weights of the different dimensions.

One of our main objectives is to construct a composite index of quality of life (QOL) which includes objective and subjective measures. PCA has its own distinction as a weighting technique in the development of composite indices as it has the virtue of simplicity and allows for weights representing the information content of individual indicators (OECD, 2008:69). According to Booysen (2002) PCA and FA are the most frequently used multivariate statistical techniques used in the weighting of composite indices.

Following Greyling (2013) and OECD (2008), the novel method developed by Nicoletti *et al.* (2000) has been applied here as a weighting technique. The method uses PCA to weight the index objectively according to the explained variance in the data. This method considers the factor loadings of the entire extracted components to weight a composite index. The benefits of this method is that a higher proportion of the variance in the data set is explained (Greyling, 2013).

The approach used by Nicoletti *et al.* (2000) is that of

1. Grouping the individual indicators with the highest factors loadings into intermediate composite indicators.

2. The weight of each of the variables in the intermediate composite is derived by squaring the factor loadings of the variables and scaling it to unity sum
within each intermediate composite index. The squared factor loadings represent the proportion of the total variance of the indicator which is explained by the component.

(3) Once the intermediate composite indices have been constructed, they are aggregated by assigning a weight to each of them equal to the proportion of the explained variance of the component in the dataset. In other words, the weights assigned to the intermediate composite indices or weight of respective factor equals the explained variance divided by total variance of each factor. Then, weight score \((W_i)\) is obtained by multiplying the variable weight and weight of respective factor. Finally, the resulting weight or final weight is obtained which is rescaled again to sum up to one to preserve comparability.

(4) After the final weights were obtained, the rank of each local council was obtained by as the product of normalized variable and the final weight.

3.5 Cluster Analysis

Cluster analysis is another multivariate data reduction technique. The difference between cluster analysis and factor analysis is whereas factor analysis groups similar variables (i.e. similarities among columns of data) based on patterns of variation among variables; cluster analysis groups similar observations (i.e. similarities among rows of data) based on similarity among variables (also called distance or proximity of variables). In cluster analysis we seek to reduce the \(n\) original observations into \(g\) groups, where \(1 \leq g \leq n\). The goal of cluster analysis is to minimize the within-group variation and maximize the between-group variation. In other words, the observations within a cluster are more similar than observations between clusters, as measured by the clustering criterion (Wang, 2006).
There are two broad types of cluster analysis- agglomerative or hierarchical method and non-hierarchical or non-agglomerative method (Rogerson, 2001). Hierarchical clustering is more widely used (Wang, 2006). Hierarchical clustering starts with each observation as a separate cluster, i.e. there are as many clusters as cases, and then combines the clusters sequentially, reducing the number of clusters at each step until only one cluster is left.

3.5.1 (Dis)similarity Measure and Clustering Method

Cluster analysis uses dissimilarities or distances between observations to form clusters. There are various measures to express (dis)similarity or distance between pairs of observations. A straightforward and generally accepted way to compute distance or proximity between objects in a multi-dimensional space is by drawing a straight line between them. This type of distance is referred to as Euclidean distance. The Euclidean distance between two points (or locations) \( i \) and \( j \) is the hypotenuse of a triangle ABC. With regard to variable \( X \) and \( Y \) and their coordinates \((X_{1i};X_{2i})\) and \((X_{1j};X_{2j})\), the euclidean distance may be given as

\[
D(i,j) = \sqrt{A^2 + B^2} = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2}
\]

Objects with smaller distances between one another are more similar, whereas objects with larger distances are more dissimilar.

After determining measure of (dis)similarity, the next procedure is to choose clustering algorithm or the way in which clusters should be joined at each stage. A number of methods are again available. In the present study, Ward’s method has been employed. In this method all possible pairs of clusters are combined and the sum of the squared distances within each cluster is calculated. The combination that gives the
lowest sum of squares is chosen. Ward’s distance between clusters $C_i$ and $C_j$ is the difference between the total within cluster sum of squares for the two clusters separately, and the within cluster sum of squares resulting from merging the two clusters in cluster $C_{ij}$.

$$D_w(C_i, C_j) = \sum_{x \in C_i} (x - r_i)^2 + \sum_{x \in C_j} (x - r_j)^2 - \sum_{x \in C_{ij}} (x - r_{ij})^2$$

Where as $r_i$ is centroid of $C_i$, $r_j$ is centroid of $C_j$ and $r_{ij}$ is centroid of $C_{ij}$.

### 3.5.2 Dendrogram - Selection of Number of Cluster

A tree-structured graph called Dendrogram is a graphical representation of the resulting clustering. It is used to illustrate the arrangement of the clusters produced by hierarchical clustering and also used to determine the number of clusters. Dendogram consists of vertical line and horizontal line. The vertical line denotes the relative similarities or dissimilarities between observations or clusters, while the horizontal line represents the observations or clusters. Clusters are formed along the horizontal line by joining individual observations or existing clusters at nodes.

In a dendrogram, the clusters are linked at increasing levels of dissimilarity. The height of the node (or joining point of observations/clusters) can be thought of as the distance value between the right and left sub-branch clusters. Any sudden increase in the difference between adjacent steps will indicate an appropriate number of clusters to consider. But whatever cut-off point is used, cluster analysis does not necessarily establish an optimal or unique solution (Mather, 1969).
3.6 Correlation Analysis

Pearson’s product moment correlation is employed to measure association between selected variables. The technique is one of the most popular methods in quantitative geography. It measures the magnitude and direction of association between two or more variables. The formula of product moment correlation for $x$ and $y$ variables is

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}[n(\Sigma y^2) - (\Sigma y)^2]}$$

The value of correlation coefficient ranges between -1 and +1. A value of -1 refers to perfect negative correlation while a value of +1 refers to perfect positive correlation. If the value is 0, it implies no relationship.

3.7 Measurement of Spatial Autocorrelation

One of the main objectives of the study is to locate the incidence of similar and dissimilar pattern of various indices of quality of life (QOL). The composite indices of QOL were analyzed with spatial autocorrelation statistics like global Moran’s I and its local equivalent, called Local Indicator of Spatial Association (LISA). These spatial statistics enable the measurement of spatial clustering and identification of spatial clusters or axes and spatial outliers in the studied data set (Goodchild, 1987). Spatial autocorrelation is very similar with correlation. However, the difference between them is “whereas correlation shows relationships between or among variables, spatial autocorrelation shows the correlation within variables across georeferenced space” (Getis, 2008: 298). The statistic is based on Tobler’s first law of geography which says that “everything is related to everything else, but near things are more related than distant things” (Tobler, 1979: 236).
3.7.1 Global Moran’s I

Developed by P. Moran in 1948, Global Moran’s I spatial autocorrelation identify clusters of variables with similar spatial patterns. Moran’s I is a test for spatial randomness; rejection of the null hypothesis implies with a certain degree of certainty that spatial autocorrelation exists. Moran’s I has an expected value of $-\frac{1}{n-1}$, that is, the value that would be obtained if there was no spatial pattern to the data. If the value I exceed $-\frac{1}{n-1}$, it indicates positive spatial autocorrelation while if the value of I is less than the expected value, it indicates negative spatial autocorrelation.

Like Pearson’s correlation coefficient, values of Moran’s I range from +1 indicating a strong positive spatial autocorrelation (high values tend to be located near one another and low values tend to be located near one another) to −1 meaning a strong negative spatial autocorrelation, wherein 0 indicates a random pattern or absence of spatial pattern. However, it may be noted that Moran’s I differs from Pearson’s product moment correlation in the sense that space is included by means of a $W$ matrix and instead of finding the correlation between two variables, the goal is to find the correlation of one variable with itself vis-à-vis a spatial weights matrix (Getis, 2010).

The Moran’s I statistic for spatial autocorrelation is given as:

$$I = \frac{N \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (x_i - \bar{x}) (x_j - \bar{x})}{(\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}) \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

whereas $N$ is the number of observations (points or polygons), $\bar{x}$ is the mean of the variable, $X_i$ is the variable value at a particular location, $X_j$ is the variable value at
another location and $w_{ij}$ is a weight indexing location of $i$ relative to $j$. If the variables are transformed into $Z$-scores $Z = \frac{X - \bar{X}}{\sigma}$ where $\bar{X}$ and $\sigma$ are mean of variable $X$ and its standard deviation respectively, Moran’s I may be given as

$$I = \frac{n \sum_i \sum_j w_{ij} z_i z_j}{(n - 1) \sum_i \sum_j w_{ij}}$$

The spatial weight matrix defines the structure of spatial relationships in the study region. It delimits the extent of clustering that the clustering technique is able to detect. The choice of $W$, therefore, should be considered carefully in clustering analysis. (Aldstadt, 2010:281).

There are several groups of commonly applied methods such as contiguity, inverse distance, $k$-nearest neighbours and distance band methods. The simplest and one of the most commonly used set of spatial weights is the binary contiguity matrix. Here, $W_{ij}$ is equal to one if units $i$ and $j$ share a common boundary and zero otherwise. Contiguity can be defined as ‘linear contiguity’ when regions which share a border with the region of interest are immediately on its left or right, ‘rook contiguity’ when regions share a common side with the region of interest, ‘bishop contiguity’ when regions share a vertex with the region of interest, and ‘queen contiguity’ when regions share a common side or a vertex with the region of interest (LeSage, 1999).

![Figure 3.3 Spatial contiguity weights. (a) Queen (b) Rook (c) Bishop](image)

73
The spatial weight matrix employed in the present study is a simple binary contiguity W matrix based on the concept of Queen contiguity in GeoDa software. Queen contiguity is given as polygon contiguity (first order) in ArcGIS. If a district \( i \) shares a border or a vertex with another district \( j \), they are considered as neighbours, and \( W_{ij} \) takes the value 1 and 0 otherwise. This matrix is also zero along its diagonal implying that a district cannot be a neighbour to itself.

The weights matrix used in cluster analysis is standardized so that the elements of each row sum to one (row standardization). This procedure serves to equalize the weight given each observation in the analysis with respect to its number of neighbors. The elements of this standardized matrix are calculated as

\[
W^*_ij = \frac{w_{ij}}{\sum_j w_{ij}}
\]

Another important component of spatial autocorrelation is visualization of spatial autocorrelation by means of a Moran Scatter plot (Anselin, 1995; 1996). This is a specialized scatter plot with the spatially lagged transformation of a variable on the y-axis and the original variable on the x-axis. A spatial lag of a variable is defined as a weighted average of observations on the variable over neighbouring units (Drukker et al., 2013). For example, a locality with three neighbouring tracts that had a value of 2, 4 and 6 would have a spatial lag of 4; that is \((2+4+6/3=4)\). A row-standardized spatial weight matrix is used in the construction of the spatial lag and simply represents the average rate of each neighbouring tract. Thus, the spatial lag is the spatially weighted average of the values at neighbouring units, and is calculated as

\[
lag_i = \frac{\sum_{j=1}^n w_{ij}x_j}{\sum_{j=1}^n w_{ij}}
\]
The Moran I scatter plot (Figure 3.4 below) places the unit values \( (x_i) \) on the horizontal axis and the spatial lag \( (\text{lag}_i) \) for the same variable on the vertical axis. The axes of the plot are drawn so that they cross at the average value of \( x_i \) and \( \text{lag}_i \), respectively. The four quadrants of the plot separate the spatial association into four components. Units that fall into the quadrants labelled High-High (HH) and Low-Low (LL) represent clustering of high and low values respectively. The remaining two quadrants High-Low (HL) and Low-High (LH) indicate spatial randomness. Units that fall into these quadrants have negative association with their neighbours and can be considered as spatial outliers. A spatial outlier may arise from a cluster consisting of just one unit. The Moran scatter plot is a useful visualization tool for assessing spatial pattern and spatial clustering.

![Figure 3.4 Moran Scatter plot](image)

Moran’s I is not only a descriptive statistic but also an inferential statistic and the result of the analysis have to be interpreted within the context of null hypothesis. In other words, statistical inference analysis is required to statistically confirm Moran’s I value against the null hypothesis of spatial randomness.

The null hypothesis \( (H_0) \) and the alternative hypothesis \( (H_1) \) may be stated as

\[
H_0: I = 0 \text{ or } H_0: \text{ spatial randomness}
\]
H₁: \( I \neq 0 \) or H₁: spatial dependence (clustering or dispersion)

The null hypothesis can be accepted or rejected by comparing the nominal significance level \( \alpha \) with the pseudo \( p \)-value. Since the \( p \)-value depends on the number of permutations, it is often called pseudo \( p \)-value. The null hypothesis (\( H_0 \)) is rejected when the pseudo \( p \)-value is less than the specified significance level (\( \alpha \)) which is normally 0.05.

### 3.7.2 Local Indicators of Spatial Association (LISA)

While global Moran’s I quantifies the spatial autocorrelation as a whole, the local indicators of spatial association (LISA) measures the degree of spatial autocorrelation at each specific location (Anselin, 1995) by using local Moran’s I. It is a measure to identify patterns of spatial clustering and spatial outliers (Harries, 2006) and the formula is given by Levine (2004) as

\[
I_i = \frac{z_i - \bar{z}}{\sigma^2} \sum_{j=1, j \neq 1}^{n} W_{ij}(z_j - \bar{z})
\]

where \( \bar{z} \) is the mean value of \( z \) with the sample number of \( n \), \( z_i \) is the value of the variable at location \( i \); \( z_j \) is the value at other locations (where \( j \neq i \)); \( \sigma^2 \) is the variance of \( z \); and \( W_{ij} \) is a distance weighting between \( z_i \) and \( z_j \).

### 3.8 Choropleth Map

Choropleth or choroplethic maps are planimetric representations of volumetric statistical distributions, and are normally symbolized by patterns which divide the area into sub-regions (Jenks and Coulson, 1963). A number of choropleth maps were prepared and used as an important analytical tool.
A widely discussed and debatable issue in preparation of choropleth map is determination of class interval which is an intrinsic component of choropleth map. Selection of appropriate class interval method helps in more accurate generalization and effectiveness of the map. The problem of selecting classes is, therefore, the most important phase of constructing a statistical map since in this step the map-maker controls map interpretation (Jenks and Coulson, 1963).

Different methods to determine class interval for choropleth map have been suggested. One of the most popular methods is ‘equal interval’ method which place boundaries between classes at regular (equal) intervals. This is the simplest and the most appropriate method for variables with a rectangular frequency distribution (Monmonier, 1972). Equal interval method does not account for data distribution, and may result in most data values falling into one or two classes, or classes with no values. Another method is ‘quantile method’ in which each class contains an equal number of observations or data. A quantile classification is well suited to linearly distributed data. The problem with quantile schemes is that they often place similar values in different classes or very different values in the same class. Another popular method is standard deviation method (Armstrong, 1969). In this method, data are classified into groups according to the variances from the data’s mean value by subtracting or adding the calculated standard deviation from the mean of the data set. The main constraint with standard deviation method is that it works well only for data that exhibit normal (Gaussian) distribution. No method may be considered as the best method since each has its advantages and limitations.

In the present study, natural break data classification method developed by Jenks and his associates has been adopted for determining class interval. This method is known as ‘Jenks Natural Breaks’ method or ‘Jenks Optimization method’ is an
algorithm used to classify features using natural breaks in data values. The method partitions statistical data into classes using an algorithm which calculates groupings of data values based on the data distribution (Jenks, 1967). The method is designed to determine the best arrangement of values into different classes by minimizing variance within groups and maximizing variance between groups. In other words, Natural-breaks schemes minimize differences between values within classes and maximize differences between values in different classes. However, manual calculation is almost impractical as there are an overwhelming number of different ways to set map ranges. As such, Jenks method is calculated using ArcGIS software that automatically figures the natural breaks. The number of classes for all the choropleth maps were specified at five since fewer classes often result in distinct patterns while more classes often result in complex patterns.

3.9 Limitations of the Study

1. Both objective and subjective indicators identified in the present study are not definitive. A number of indicators were dropped to conform to the technical specificities required by the statistical techniques employed.

2. The impartiality of the respondents is invalidated. They may provide faulty answers depending on their personal integrity, aspirations and level of knowledge.

3. The scope of the study is also limited by the impossibility of presenting temporal analysis due to unavailability of time-series data.

4. One of the limitations of factor analysis and principal component analysis is that naming the factors/components is purely subjective and can be problematic. Factor names may not accurately reflect the variables within the factor.
6. Some variables are difficult to interpret because they may load onto more than one factor which is known as split loadings. These variables may correlate with each another to produce a factor despite having little underlying meaning for the factor.