

# Analysis of Spatial Data over Open Street Map Using Latent Semantic Analysis and Genetic Algorithm

Pallavi Tyagi<sup>1</sup>, Dharamveer Singh<sup>2</sup>, Vikas Gupta<sup>3</sup>,

<sup>1,2,3</sup>Deptt. of Computer Science & Engineering, R.D. Engineering College, Ghaziabad, India

**Abstract** - The need for information retrieval about the aforementioned location is increased by the "expansion or improvements of metropolitan areas and rural areas is moderately speedily going on, particularly the metropolis of India like (Delhi, Mumbai, and more), in addition to the said location's transportation infrastructure, goods, hospitality, business opportunities, and agricultural aspects. To remove the barriers or reduce the level of service would be detrimental to the masses' progress and development because the demand for these resources must increase. Latent Semantic Analysis and Genetic Algorithm are two Machine Learning techniques that are incorporated into the proposed approach to enable quick and efficient information retrieval from massive corpora or map repositories. The supervision model, however, is created using a service investigation based on the capacity and volume of data in relation to places and references. With the aid of picture interpretation and measuring vector data in the form of XML from Open Street Map, data on geometrical patterns and land use are obtained. The study's findings show an accuracy level of about 79 percent. The scheme primarily makes use of geometrical data from OSM to effectively extract land use data for information retrieval.

**Keywords:** OpenSteetMap, Machine Learning, Latent Semantic Analysis, Genetic Algorithm, Singular Value Decomposition, X tensible Markup Language.

## I. INTRODUCTION

One formal method of studying items utilizing their topological, geometrical, or geographic features is called spatial analysis or spatial statistics. "Spatial analysis includes a variety of methods, many of which are still in the early stages of development. These methods are applied in areas like botany, forestry, horticulture, and agriculture with the study of plant placement and the use of "places and routes" algorithms to construct intricate cable structures. In a more constrained sense, spatial analysis is a method used to study human-scale systems, particularly when studying geographic data. Spatial analysis raises complex concerns, many of which are not well defined or resolved but serve as the foundation for contemporary study. The issue of locating the entity under study in space is the most fundamental of these. Because there are so many distinct fields of study involved, multiple core approaches can be used, and numerous types of data can be used, categorising spatial analytic techniques is challenging. [1-2] Spatial analysis might be considered [3] has emerged with initial efforts on cartography and surveys but many fields have contributed to its improvement in modern forms. Biology contributes through botanical studies on global plant distribution and location of local plants, studies of animal movement ethnology, ecological landscape studies of vegetation blocks, spatial population dynamics ecological studies, and biogeography studies. Epidemiology contributes with early work on disease mapping, especially John Snow's work on mapping cholera outbreaks, with research on mapping the spread of disease and with site studies for health care delivery. Statistics has made a major contribution through work in spatial statistics. Economics has contributed mainly through spatial econometrics.

## II. LITRETURE REVIEW

Computer tools support the spatial definition of objects as homogeneous and separate elements due to the limited number of database elements and computational structures available, and the ease with which these primitive structures can be created [3-5]. Spatial dependence is a co-variation of property in geographical space: characteristics at proximal locations appear to be correlated, both positively and negatively. Spatial dependence leads to the problem of spatial autocorrelation in statistics because, like temporal autocorrelation, this violates standard statistical techniques that assume independence between observations. For example, a regression analysis that does not compensate for spatial dependence can have unstable parameter estimates and produce significant, unreliable tests. The spatial regression model (see below) captures this relationship and does not suffer from this weakness. It is also appropriate to see spatial dependence as a source of information and not something that needs to be fixed. [5-8]. Spatial "measurement scales are a persistent problem in spatial analysis; more detail is available in the Modification Unit Issue (MAUP) topic entry that can be modified. Landscape ecologists develop a series of invariant scale metrics for fractal ecological aspects. [9] In more general terms, there is no widely agreed-upon method of independent scale analysis for spatial statistics. In order to quantify phenomena that depend on dependence and heterogeneity, spatial sampling entails choosing a variety of sites in geographic space. [10] Dependency demonstrates that we don't require observations at both locations because one site may anticipate the value of the other location. However, heterogeneity demonstrates that this link can alter over geography, thus we cannot rely on the degree of dependence shown outside of potentially tiny regions. Random, group, and systematic sampling techniques are the most fundamental. Bias, distortion, and direct inaccuracies in analysis results are only a few of the issues brought on by the "basic difficulty in spatial analysis. Although there are many connections between these concerns, different efforts have been taken to keep some of them apart. [11]. Open Street Map (OSM) [12-16] is a GIS Web product that can also be operated using web browsers and smartphone so that it can be categorized smart GIS. The use of smart GIS really helps the mapping process because it is more efficient in terms of time, device, and can be taken to any area. Through

the Open Data Commons Open Database License 1.0, OSM contributors can own, modify, and share map data widely. There are various types of digital maps available on the internet, but most have legal and technical limitations. This makes the community, government, researchers and academics, innovators, and many other parties unable to freely use the data available on the map. On the other hand, both the OSM base map and the data available in it can be downloaded for free and open, for later use and redistribution. Thus it is hoped that utilizing OSM will become an alternative source of data, especially for mapping urban areas and rural areas, one of which is related to locations and directions. (Open Street Map, 2019). Open Street map is created by a mutual map community contribute and maintain data on roads, trails, cafes, stations, and many other things throughout the world. Open Street map emphasizes local knowledge. Donors use aerial photography, devices GPS, and field maps to verify OSM accuracy and always updated [12]. Open street map users continue to grow every year. The development of urban areas that is quite rapid, especially in the cities, helped increase the need for transportation facilities and infrastructure. The increasing number of motorized vehicles has an impact on the decreasing level of road services. Road management is based on the results of analysis of the level of road services obtained from comparison of traffic volume and road capacity. The method used to obtain data that is geometric data of roads and land use is done by interpreting images and measuring vector data from Open Street Map [12], as well as field survey activities. The results of the study present the level of accuracy of vector data, especially geometric roads and the ability of imagery available in OSM to tap land use data. Each of the accuracy test results shows that vector and raster data in OSM are suitable as alternative data sources for mapping road services. In general, road conditions in city have a poor service level, which is below class C to F. Road management recommendations given in general are traffic lights, parking, road markings, and road geometry improvements. The Open Street Map project (abbreviated OSM) was launched by Steve Coast in summer 2004 founded in London. The aim was to collect data for a free card. The back then available cards were either expensive or were not under a free license available for any use. So there is a similar one behind Open Street Map Motivation like behind free software - the desire to be able to do what you want with it wants [13]. A large number of volunteers - on November 14, 2016 there were 3 199 742 user accounts [14] - compiled the data. Only some of these users ever have edited an OSM object. In December 2011 there were 505,000 users only 38 percent edited an object [15]. As of July 31, 2019, there were approximately 562,000 of the 2.2 Millions of users (26 percent) who had ever edited an object [16]. The Open Street Map Foundation has been behind the Open Street Map project since 2006, which operates the server and since switching from the Creative Commons license Attribution Share under the same conditions 2.0 to the Open Database License 1.0 (ODb) in September 2012 also holds the rights to the data and it under the Conditions of the ODbL available to everyone [16, 17, 18]. Unlike the Wikipedia, Open Street Map has no relevance criteria, but only records things that are observable on site and not directly personal Data like doorbell signs. Exceptions to this rule only exist for things which are considered important for a card, such as B. Administrative limits Open Street Map started from scratch in many places, only in some areas was early Years of uniform import coverage of data from external sources. In the other areas (especially in Europe) there has been a powerful one Community formed, which records the data manually on site and their focus is located in German-speaking countries (from the start of the project to December 2011, 31 percent of all users had data mainly in Germany, Austria or Switzerland recorded [19]). Bing Maps aerial photos have been available since November 2018.

## 2.1 OPENSTREETMAP DATA MODEL

There are three different types of objects at OSM - nodes, ways and relations. All these Object types can, but do not have to, have tags. OSM data is i. d. Usually transmitted in an OSM-specific XML format, the file extension is usually osm [20]. In addition, a space-saving binary variant is used, which uses the protocol buffers developed by Google [20]. Are protocol buffers a language-independent and platform-independent format for serializing structured data, similar to XML [21]. When processing OSM data, protocol buffers have in XML format has been displaced in recent years because they are faster and easier to parse have less overhead. The data model follows the KISS principle (Keep it simple, stupid) and is special adapted to the needs of a community project. There are very few fixed guidelines on how certain things are to be recorded. The different types of objects (Nodes, ways, relations) are not based on common GIS standards. The Entry hurdle should be as low as possible for new mappers and developers, because without (New) mappers and developers of third-party applications the project has no future [21]. Tags are used to define what the respective OSM object represents. Deals with tags are not the categories that are common with other Web 2.0 services. With OpenStreetMap, tags are key-value pairs that can be freely selected. Key and Value can each be up to 256 characters long [22]. There are no hard and fast rules about which tags to use. However, since countless applications use OSM data and an inconsistent recording of further processing is not beneficial, certain rules have emerged. Common tags are documented in the Map Features list on the OSM Wiki [23]. Nevertheless, the principle Any tags you like applies to OpenStreetMap. The user can define any tags itself [23]. The data model is deliberately kept very flexible at this point, so that early design decisions do not negatively influence the development of the project and the project remains agile.

Relations are used for objects and relationships that cannot be easily modeled with nodes, ways and tags. Relations can have tags and an ordered list of members. The members have roles, which are strings (strings with a length of 0 are possible). The type tag indicates the relation type. Common types are multipolygon, boundary, route and turn\_restriction. This list is not exhaustive. Multipolygon relations with the tag type = multipolygon for surfaces that cannot be modeled by a single outer ring. The members of a multipolygon relation are all outer and inner rings, i.e. the ways that represent them. Outer rings have the role of outer, inner rings have the role of inner. A ring can consist of one or more ways. The tags representing the properties of the area describe, hang on the relation object [23, 24]. In total there are about 2.5 million multipolygon relations in OpenStreetMap [24]. The term multipolygon has a different meaning in the OpenStreetMap environment than in the Simple Features standard of the Open Geospatial Consortium (OGC). An OSM multipolygon usually has only one outer ring and is not a collective geometry. In the OGC standard, however, a multipolygon is a MultiSurface, whose elements are of the polygon type. For surfaces that have only one outer ring (and any number of inner rings), the type Polygon exists in the OGC standard [24]. The OSM wiki contains a description of when a multipolygon is valid. In principle, this corresponds to the OGC's Simple Features specification. In

addition, OSM allows two inner rings to not only touch at one point, but to have a common edge [23, 24]. In OpenStreetMap there are a number of old style multipolygons that differ from the specification given above. There are currently around 250,000 multipolygons with the tags hanging on the outer ring. Another 17,000 multipolygons follow an even older version of the OSM multipolygon specification [23, 24]. With them, the tags that describe the area can be found on both the outer and the inner ring. Multipolygons whose members have no role (neither inner nor outer) also originate from this period. Data users are encouraged to consider roles first when designing simple feature geometries from OSM data. If this approach does not lead to a valid geometry, it is recommended to try again without considering the roles. Listing 1.3 shows a multipolygon in the OSM XML representation. Limits Relations with the type = boundary tag are used for administrative polygons, protected areas and restricted areas. They are structured like multipolygon relations. Listing 1.3: Example of a relation of the type multipolygon in the OSM XML representation (abbreviated) [23]. The matrix is a matrix that represents the contents of the entire document, albeit having fewer dimensions. The hallmark of LSA is a technique called Singular Value Decomposition (SVD). SVD is used to perform matrix decomposition after weighting and then measure its similarity with the data to be tested [25]. In 1990 through a journal titled "Indexing by Latent Semantic Analysis" by Scott Deerwester, Susan Dumais, George Furnas, Richard Harshman, Thomas Landauer, Karen Lochbaum and Lynn Streeter, an algorithm was introduced to index words in documents and plot them into a vector base that called Latent Semantic Analysis (LSA)[26]. The LSA method is one of the emerging information retrieval algorithms that can gather several documents in a database and establish associations between them by matching the provided query. The LSA algorithm, more particularly, is a process for creating vector-based words (terms) that are thought to be able to capture the semantics of a document or sentence. By comparing the vector representation of each document, this LSA's primary purpose is to determine how similar two documents are. When creating vector-based word representations, LSA creates a matrix known as a semantic space, which symbolises the relationship between terms and documents. words and documents that are closely associated will be placed close to each other represented by vectors. LSA in its calculations using Singular Value Decomposition (SVD)[27]. SVD represents semantic space in the form of matrices that have smaller orders than the original matrix order, but matrix calculations still produce matrices that are almost the same value. SVD is a linear algebra theorem which is said to be able to break the block of a matrix A into three new matrices, namely an orthogonal matrix U, diagonal matrix S, and Transpose matrix orthogonal V.

The LSA method maps the ways or nodes to a concept space and comparisons are made on this space. The concept space, or more commonly referred to as latent semantic space, is the result of mapping from a high dimensional matrix to a smaller dimension. Although in smaller dimensions, the matrix is a matrix that represents the contents of the whole document. The hallmark of LSA is a technique called Singular Value Decomposition (SVD). SVD is used to perform matrix decomposition after weighting and then measure its similarity with the data to be tested [28].

### III. METHODOLOGY

Under the proposed scheme we are going to extract the geo-spatial data or corpus from Open Street Map using map-extractor facilitation for the specific country, state or region. Thereafter, using the latent semantic analysis we will process the data using singular value decomposition which we remove the noise and will form the un-structured layout with respective entities and attributed for feature detection like relations and indexing. Subsequently, using a genetic algorithm with respect to elements and attributes the effective and evolutionary references will be espionage which will mitigate the retro respective data with its semantic models using crossover and mutation to provide appropriate and accurate information promptly. The below figure depicts the workflow of the proposed scheme for perusal and ready reference.

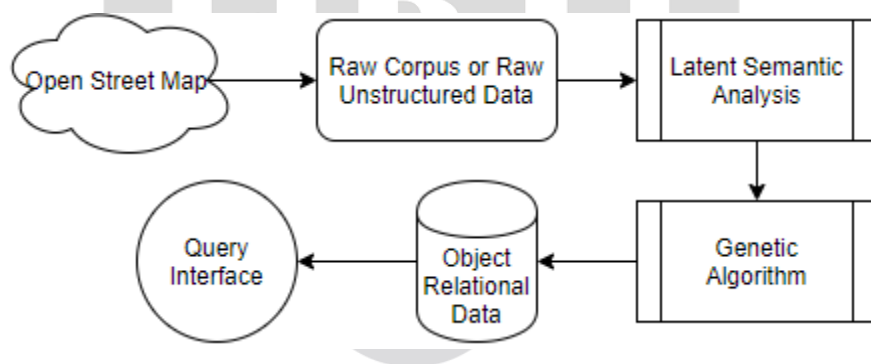


Figure 3.1: Workflow of Proposed Scheme using LSA and GA.

#### 3.1 LATENT SEMANTIC ANALYSIS

LSA is a method based on machine learning in excess of natural language processing, in exacting distributional semantics, of analyzing relationships between a set of credentials/data/corpus/documents and the terms they enclose by fabricating a set of conception related to the data/corpus/documents and conditions. LSA presume that expressions that are close in connotation will transpire incomparable pieces of text (the distributional hypothesis). However, Latent semantic analysis aims to highlight hidden semantic relations among terms and enables protuberance of uncertainty and credentials in the same space defined with semantic dimensions. The preliminary set of documents is accessible in a form of matrix A where one column represents one document. Rows of the matrix are terms that occur in documents. Field  $A_{i,j}$  of the matrix is the occasion of the term i in the document j. If we imagine every row as one dimension of semantic space then every column is a vector that projects the corresponding



document in such space. Matrix A, composed in the described way, has a very large number of rows/dimensions (n). To make this space easier to handle, the reduction of dimensions is necessary. Reduced space is called latent space because hidden (latent) knowledge of co-occurrence of terms is revealed. One of the techniques for dimension reduction is the Singular Value Decomposition (SVD). This decomposition reduces the space in such a way that the difference between projection in original and latent space is minimal. When SVD is applied to matrix A we get three new matrices:  $A = USV^T$  where U and  $V^T$  are orthogonal matrices and S is the diagonal matrix composed of singular values of A matrix. By restricting matrices U, V and S to their first k rows and columns, one can create a projection of matrix A in k-dimensional space ( $k \ll n$ ). Such projection has minimum deviation because singular values in the S matrix are ordered in descending order and they are considered to be weights for the relevance of a particular dimension. If one takes the same approach for representing the queries as semantic vectors, then those queries can be projected in the same semantic space and compared to the vector of particular document. Angle between vectors is usually taken as quantify of relationship. The below diagram depicts the architecture of latent semantic analysis.

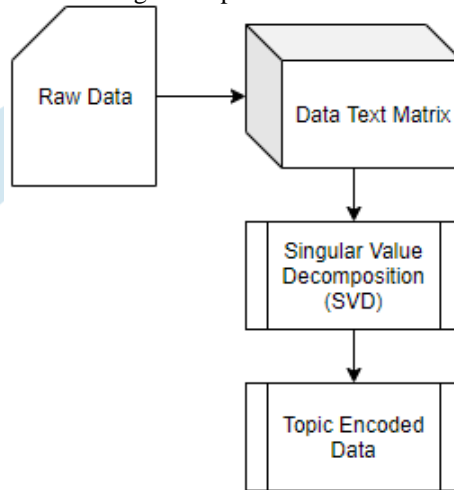


Figure 3.2 Architecture of Latent Semantic Analysis using Singular Value Decomposition

### 3.1.1 LSA Pseudo Code

In the analysis process involves the formation of vectors using XML documents thereafter vector training data documents the process of which is shown as under in pseudo code.

#### Query Vector Training Algorithm

```

01: For i = 1 to p do
02: a) Remove Noise from the Corpus (XML Documents)
03: b) Stem words/Nodes from the training answer Corpus
04: End For
05: Select for the index-term of Vector by matrix document
06: Vectors form of matrix using XML document,  $A_{m \times p}$ 
07: Decompose the  $A_{m \times p}$  matrix using SVD, where the equation will be

$$A_{m \times p} = U_{m \times r} * S_{r \times r} * V_{r \times p}^T$$

08: Truncate / cut / reduce from U, S and VT and make the following equation:

$$A_{k \times k} = U_{m \times k} * S_{k \times k} * V_{k \times p}$$

09: For j = 1 to p Do
10: Form key vector well formed XML Documents as

$$D_j = D_j^T * U_{m \times k} * S_{k \times k}^{-1}$$

11: End For
  
```

The next step is the same as forming a query vector training, the reduced SVD matrix will used to form the query document answer vector students, In this process each XML document node will later formed query vectors that will be compared with queries vector training data answers so that values can be determined the similarity which is the basis for giving value automatically by the system. This process can be demonstrated by equation below along with pseudo code.

$$Q_j = Q_j^T * U_k * S_k^{-1} \text{ (eq. 3.1)}$$

Information:

$Q_j$ : Vector query documents student answers

$Q_j^T$ : Transpose vector query to students' answer documents

$U_k$ : Orthogonal reduction matrix

$S_k^{-1}$ : Singular reduction inverse matrix

#### Algorithm for the Formation of XML Document Vector Queries

```

01: For i = 1 to p do
02: a) Remove Redundant Nodes from essay XML Document  $D'$ 
  
```

03: b) Stem words from Nodes and Attributes

04: EndFor

05: The same matrix query dimensions are form rules using XQuery from matrix documents

06: For j = 1 to p Do

07: Form the answer vector as

$$Q'_j = Q_j^T * U_{mxk} * S_{ksk}^{-1} \text{ (eq. 3.2)}$$

EndFor

### 3.2 GENETIC ALGORITHM

GA was primarily initiated by John Holland in the 1970s (Holland 1975) as a result of examinations into the opportunity of computer programs that will undergo evolutionally in the Darwinian sense. GA is a component of a broader soft computing archetype acknowledged as evolutionary computation. They endeavor to disembark at the finest elucidation during a progression similar to biological evolution. This engages following the principles of survival of the fittest and crossbreeding and mutation to produce better solutions from a pool of obtainable solutions. Genetic algorithms have been seen as equipped for discovering answers for a wide assortment of issues for which no satisfactory algorithmic arrangements exist. The GA strategy is especially appropriate for enhancement and optimization, a critical thinking method wherein at least one generally excellent arrangement are looked for in an answer space comprising of an enormous number of potential arrangements or solutions. GA lessens or reduces the search space by consistently assessing the current generation of candidate solutions, disposing of the ones positioned as poor, and creating another age through crossbreeding and mutating those positioned as great. The positioning of up-and-comer arrangements is finished utilizing some pre-decided proportion of fitness or wellness. The GA evolutionary cycle begins with a randomly selected initial population. The progressions to the population happen through the procedures of choice dependent on fitness, and adjustment utilizing crossover and mutation. The use of choice and adjustment prompts a populace population a higher extent of better arrangements. The evolutionary cycle proceeds until a satisfactory arrangement is found in the present age of populace, or some control parameter, for example, the quantity of ages is surpassed or exceeded the below diagram depicts the genetic algorithm evolutionary cycle.

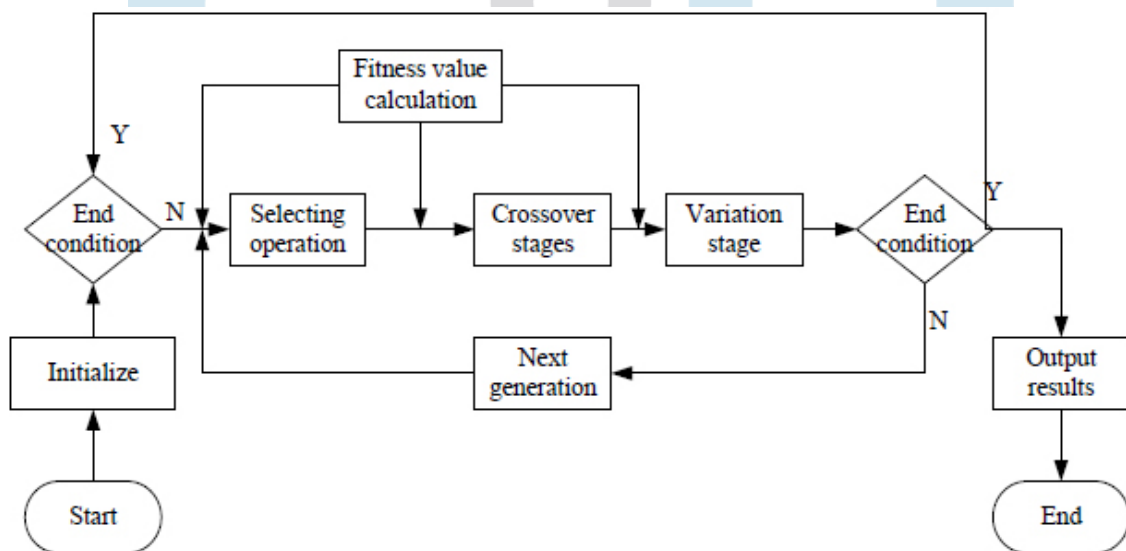


Figure 3.3 Work Flow Model of Genetic Algorithm

The negligible component of a genetic algorithm is known as a *gene*, which symbolize a element of information in the predicament sphere. A sequence of genes, recognized as a *chromosome*, signify one potential solution to the problem. every gene in the chromosome correspond to one element of the solution prototype.

- 1. Selection:** In genetic progression, only the fittest survive and their gene pool contributes to the formation of the next generation. Selection in GA is also pedestal on a comparable procedure. In a familiar form of selection, known as *fitness proportional selection*, each chromosome's likelihood of being preferred as a excellent one is comparative to its fitness value.
- 2. Alteration to improve good solutions:** The amendment steps in the genetic algorithm filter the good solution from the current generation to produce the next generation of candidate solutions. It is carried out by performing crossover and mutation.
- 3. Crossover:** "might be viewed as artificial mating in which chromosomes from two individuals are merged to form the chromosome for the subsequent creation. To do this, two chromosomes from two separate solutions are spliced together at a crossover location and then the spliced pieces are switched. The idea is that some genes with more unique personalities from one chromosome might therefore merge with some beneficial genes from the other chromosome to produce a new chromosome that is more effective.

- 4. Mutation** is a unsystematic modification in the genetic constitution. It is useful for establish new distinctiveness in a population – impressively not accomplish all the way through crossover unaccompanied. Crossover only rearranges accessible characteristics to give new combinations. For example, if the first bit in every chromosome of a generation happens to be a 1, any new chromosome created through crossover will also have 1 as the first bit. The mutation operators revolutionize the present value of a gene to a different one. For bit string chromosome this change amounts to flipping a 0 bit to a 1 or vice versa. Although useful for introducing new traits in the solution pool, mutations can be counterproductive and practical only occasionally and randomly.

Pseudo Code is depicted for ready reference:-

**Step1:** Generate an initial Random Population

While iteration  $\leq$  maxiteration

Iteration = iteration + 1

**Step2:** Calculate the Fitness of each Individual

Select the Individual according to its Fitness

**Step3:** Perform Crossover with probability  $pc$

**Step4:** Perform mutation with probability  $pm$

**Step5:** Population = selected individual after crossover and mutation

End while

**Algorithmic code as under:-**

1. Set  $t := 0$ ;
2. Initialize  $P(t) := \{S_1, \dots, S_N\}$ ,  $S_i \in \{0,1\}^n$ ;
3. evaluate  $P(t) := \{f(S_1), u(S_1)\}, \dots, \{f(S_N), u(S_N)\}$ ;
4. find  $\min_{S \in P(t), u(S)=0} \{f(S)\} \vee \min_{S \in P(t), u(S)>0} \{u(S)\}$ , set  $S' \leftarrow S$ ;
5. while  $(t < t_{max}) \neq \text{true}$  do
6. Select  $\{P_1, P_2\} := \Phi(P(t)) \dot{-} \Phi = \text{matching selection method} *$
7. Crossover  $C := \Omega_f(P_1, P_2) \dot{-} \Omega = \text{uniform crossover operator} *$
8. Mutate  $C \leftarrow \Omega_m(C, m_s, m_a, \epsilon) \dot{-} \Omega_m = \text{static} \wedge \text{adaptive mutation} *$
9.  $C \leftarrow \Omega_{improve}(C) \dot{-} \Omega_{improve} = \text{heuristic improvement Operator} *$
10. If  $C \equiv \text{any } S_i \in P(t) \xrightarrow{\text{then}} C \text{ is redundant} *$
11. Discard  $C$  and go to 6;
12. End if
13. Evaluate  $f(C)$ ,  $U(C)$ ;
14. Find  $aS' \in P(t)$  based on the ranking replacement method  $\wedge$  replace  $S' \leftarrow C$ ;
15. If  $(u(C) = u(S') = 0 \text{ and } f(C) < f(S')) \text{ or } (u(C) > 0, u(S') > 0 \text{ and } u(C) < u(S'))$  then
16.  $S^* \leftarrow C$ ;
17. End if
18.  $t \leftarrow t + 1$
19. end while
20. return  $S^*$ ,  $f(S^*)$  and  $(S^*)$

### 3.3 PROPOSED SCHEME TESTING METHODS

To test the software that has been studied, the authors measure it by calculating accuracy, precision, recall, and F1-Measure. Accuracy is an evaluation measure commonly used in the classification process in machine learning.

$$\text{Accuracy} = \frac{\text{Amount of Correct Data}}{\text{All Amount of Data}} \times 100\% \quad (\text{eq.3.3})$$

Precision is the measure of accuracy between the information requested and the received answers. Precision is number of positive categorized samples classified correctly divided by the total sample classified as positive. Then obtained formula for calculating precision with equation below :-

$$\text{Precision} = \frac{\text{Amount of Correct Data}}{\text{Amount of Data} \in \text{One Category}} \times 100\% \quad (\text{eq.3.4})$$

Recall is a measure of the level of success of the system in rediscovering a information. In other words, recall is used in an information retrieval system (information retrieval) to assess how much success is obtained.

$$\text{Recall} = \frac{\text{Amount of Correct Data}}{\text{Amount of Data Predicted} \in \text{One Category}} \times 100\% \quad (\text{eq.3.5})$$

## IV. IMPLEMENTATION AND RESULTS

### 4.1 IMPLEMENTATION

As per the discussion in Chapter1 and Chapter3 this chapter will emphasis on implementation scenarios along with results and values respectively the modules will be data gathering from openstreetmap.org in form OSM format i.e. XML Compressed format therefore the data had to be uncompressed first, thereafter data cleansing or preprocessing forming the data in well formed manner, subsequently imposing Latent Semantic Analysis for vector and relation formation along with Singular Value Decomposition, thereafter data will be evaluated using Genetic Algorithm for Optimization of relations and nodes. Consequently, data will be bundled in Object Relational Database System for accessing data or information with accuracy and precision.

## 4.2 DOWNLOADING DATA FROM OPENSTREETMAP.ORG

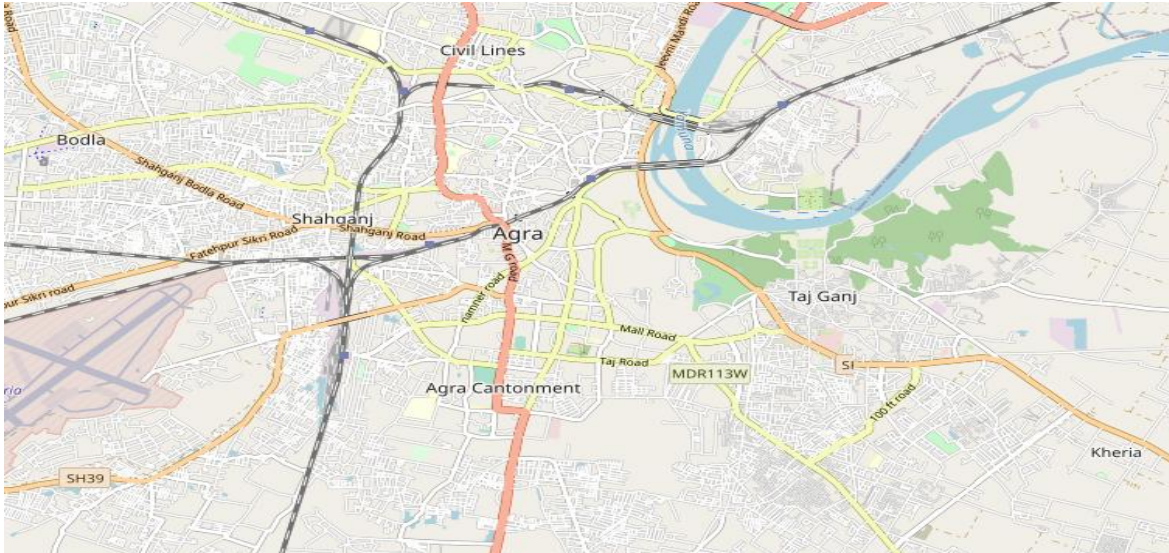


Figure 4.1 Extracting Data from OpenStreetMap.org (Query: Agra)

The proposed scheme is implemented using Linux (Red Hat Enterprise Linux) version 6.5 using virtualization environment underneath the screen depicts the booting process for the same using Grand Unified Boot Loader utility based on VimLinz Image vide Linux kernel 2.6.

Search Results

Results from [OpenStreetMap](#)  
[Nominatim](#)

City Agra, Uttar Pradesh, 280001, India

County Boundary Agra, Uttar Pradesh, India

Region Boundary Agra, Uttar Pradesh, India

Export

27.3590

77.6610 78.3586

26.9912

Manually select a different area

**Complete OSM Data**

[Latest Weekly Planet XML File](#)  
90 GB, created 33 hours ago.  
md5: 89a27174171960a3178ce1c26e2861dc

[Latest Weekly Changesets](#)  
3.4 GB, created 33 hours ago.  
md5: 98a4109737729ccb14149f2c895d0567.

[Latest Weekly Planet PBF File](#)  
51 GB, created 33 hours ago.  
md5: cf0beaa7d0accf75753e00854ac94695.

Each week, a new and complete copy of all data in OpenStreetMap is made available as both a compressed XML file and a custom PBF format file. Also available is the ['history'](#) file which contains not only up-to-date data but also older versions of data and deleted data items.

Figure 4.2: Downloading OSM file to Eco-System

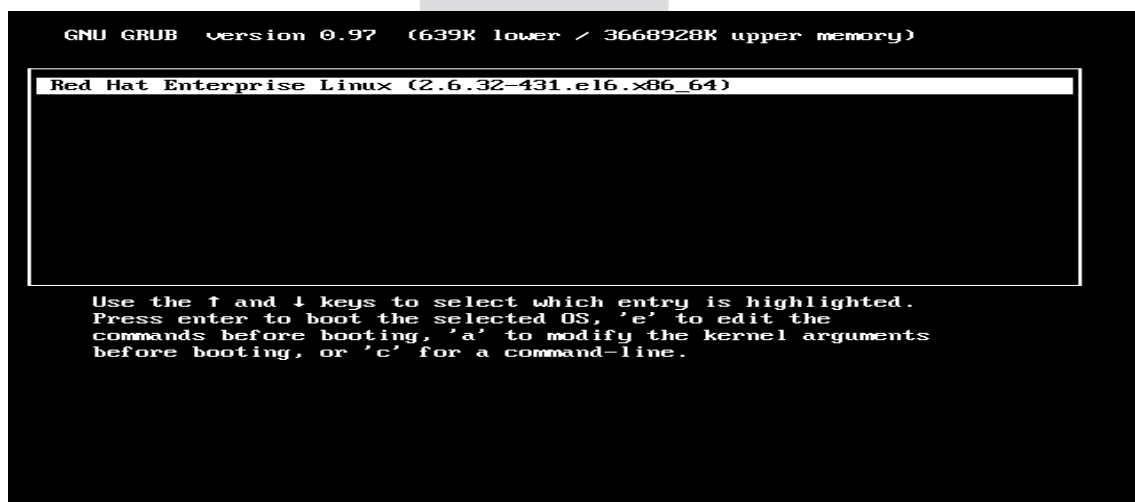




Figure 4.3: Linux (Red Hat Enterprise Linux) Version 6.5 using Virtualization Environment

The below figure depicts the natty-gritty of the open street map file of region comprising of size i.e. 712 MB on the file system however as per the properties the volume is unknown that means the file is beyond the scope of Operating System formats or the POSIX (portable operating system for UNIX family) generic categories.



Figure 4.4 Description of Open Street Map File for Extraction (Query: Agra)

#### 4.3 VISUALIZING THE OPEN STREET DATA IN XML FORMAT

The below figure depict the data downloaded from openstreetmap.org of Agra Region based on universal transformation format defined by ECMA (European Computer Manufacturer Association and W3C (World Wide Consortium) for electronic exchange for heterogeneous software and hardware resources based on standards defined by ISO (International Standard Organization). Below figure comprise properties like longitude, latitude, timestamp, location change-set, nodes, bounds, ways, direction etc.

```
<?xml version='1.0' encoding='UTF-8'?>
<osm version="0.6" generator="osmconvert 0.8.5" timestamp="2017-02-03T15:00:02Z">
  <bounds minlat="27.106" minlon="77.886" maxlat="27.259" maxlon="78.151"/>
  <node id="245756314" lat="27.1353758" lon="78.0948156" version="3" timestamp="2015-05-19T11:56:46Z"
  changeset="31281402" uid="1306" user="PlaneMad">
    <tag k="source" v="AND"/>
  </node>
  <node id="245756428" lat="27.1474758" lon="78.0788411" version="2" timestamp="2008-09-27T16:10:12Z"
  changeset="706576" uid="23057" user="prolineserver">
    <tag k="source" v="AND"/>
  </node>
  <node id="245756552" lat="27.1577626" lon="78.0499624" version="4" timestamp="2015-05-21T10:42:13Z"
  changeset="31339469" uid="510836" user="Rub21">
    <tag k="source" v="AND"/>
  </node>
  <node id="245756878" lat="27.2028236" lon="78.0430076" version="4" timestamp="2015-05-19T13:06:58Z"
  changeset="31283275" uid="1051550" user="shravan91">
    <tag k="source" v="AND"/>
  </node>
```

Figure 4.5 Sample OSM File Comprising Properties like Longitude, Latitude, Timestamp, Location Change-Set, Nodes, Bounds, Ways and Directions etc.

#### 4.4 DERIVING GENERIC INFORMATION USING LATENT SEMANTIC ANALYSIS

Below figure depicts the call of libraries used using python like xml.etree for XML iteration, print for standard output, re for regular expression, nltk for natural language processing, math for mathematical formulations, the scheme will iterate from OSM file using LSA and extract the information like tags, ways and other important numpy for scientific computing and scipy for machine learning models therefore using the below code snippet information.



```

import xml.etree.cElementTree as ET
import pprint
import re
import nltk
import math
import numpy
from nltk.corpus import stopwords
from nltk.stem.wordnet import WordNetLemmatizer
import scipy.io
from scipy import linalg
OSMFILE = "new_delhi.osm"
def matrix_reduce_sigma(matrix, dimensions=1):
    """This calculates the SVD of the matrix, reduces it and
    creates a reduced matrix.

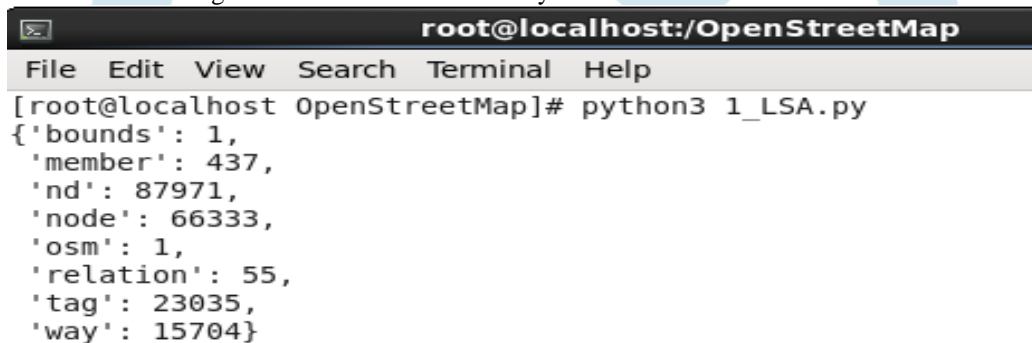
    @params matrix the matrix to reduce
    @params dimensions dimensions to reduce.

    @return matrix The reduced matrix
    """
    uu, sigma, vt = linalg.svd(matrix)
    rows = sigma.shape[0]
    cols = sigma.shape[1]

    #delete n-k smallest singular values
    #delete ie settings to zero
    smallerBound = min(rows, cols)
    for index in xrange(smallerBound - dimensions, rows):
        sigma[index] = 0

```

Figure 4.6 Latent Semantic Analyses for Information Retrieval



```

root@localhost:/OpenStreetMap
File Edit View Search Terminal Help
[root@localhost OpenStreetMap]# python3 1_LSA.py
{'bounds': 1,
 'member': 437,
 'nd': 87971,
 'node': 66333,
 'osm': 1,
 'relation': 55,
 'tag': 23035,
 'way': 15704}

```

Figure 4.7 Result Derived using LSA depicts the Information and Counts of Nodes, Relation, Tag and Ways in OSM file.

#### 4.5 DATA CLEANSING

Subsequently after deriving the results as per figure 4.7 the next step is iterate inside the nodes and remove the noise which exists in Open Street Map files below code snippet the data cleansing technique to form the well firmness inside the Corpus.

```

freq_list = [0] * len(documents)
for index, document in enumerate(documents):
    #freqs_doc =
    #nltk.FreqDist(tokenize_and_lemmatize(document)) # TODO: optimize this

    for word in tokenize_and_lemmatize(document):
        if word == keyword:
            freq_list[index] += 1#freqs_doc.freq(keyword)

    return freq_list

def key_type(element, keys):
    if element.tag == "tag":
        for tag in element.getiterator('tag'):
            k = tag.get('k')
            if lower.search(element.attrib['k']):
                keys['lower'] = keys['lower'] + 1
            elif lower_colon.search(element.attrib['k']):
                keys['lower_colon'] = keys['lower_colon'] + 1
            elif problemchars.search(element.attrib['k']):
                keys['problemchars'] = keys['problemchars'] + 1
            else:
                keys['other'] = keys['other'] + 1

    return keys

def process_map(filename):
    keys = {"lower": 0, "lower_colon": 0, "problemchars": 0, "other": 0}
    for _, element in ET.iterparse(filename):
        keys = key_type(element, keys)
    return keys

[root@localhost OpenStreetMap]# python3 2_Semantics.py
{'lower': 22529, 'lower_colon': 484, 'other': 22, 'problemchars': 0}

```

Figure 4.8 Code Block using LSA for Data Cleansing

#### 4.6 SEM ANTIC RELATION BINDING

After removing the noise from the corpus (agra\_india\_sample.osm) as mentioned in figure 4.8 the Latent Semantic Analysis will form the relational among the nodes or will define the among them. The relation is based upon the nodes and the ways exists in open street map files which will segregate the complexity and form the new open street map (agra\_india\_sample.osm) file with respective nodes which contains information and relations. Therefore, the Latent Semantic Analysis will parse only those nodes which contains the information in itself or in attributes and relation to another or other nodes.

```
import xml.etree.ElementTree as ET
import pprint
import re
import nltk
import math
import numpy
from nltk.corpus import stopwords
from nltk.stem.wordnet import WordNetLemmatizer
import scipy.io
from scipy import linalg

OSM_FILE = "agra_india.osm"
SAMPLE_FILE = "agra_india_sample.osm"

k = 100 # Parameter: take every k-th top level element

def get_element(osm_file, tags=('node', 'way', 'relation')):
    """Yield element if it is the right type of tag

    Reference:
    http://stackoverflow.com/questions/3095434/inserting-newlines-in-xml-file-generated-via-xml-etree-elementtree-in-python
    """
    context = iter(ET.iterparse(osm_file, events=('start', 'end')))
    _, root = next(context)
    for event, elem in context:
        if event == 'end' and elem.tag in tags:
            yield elem
            root.clear()

with open(SAMPLE_FILE, 'wb') as output:
    output.write('<?xml version="1.0" encoding="UTF-8"?>\n')
    output.write('<osm>\n ')
```

Existing Corpus

New Corpus After Semantic Relations

Figure 4.9 Espionage the Relevant information from Raw Corpus (Existing Corpus) to New Well Formed Corpus using Latent Semantic Analysis

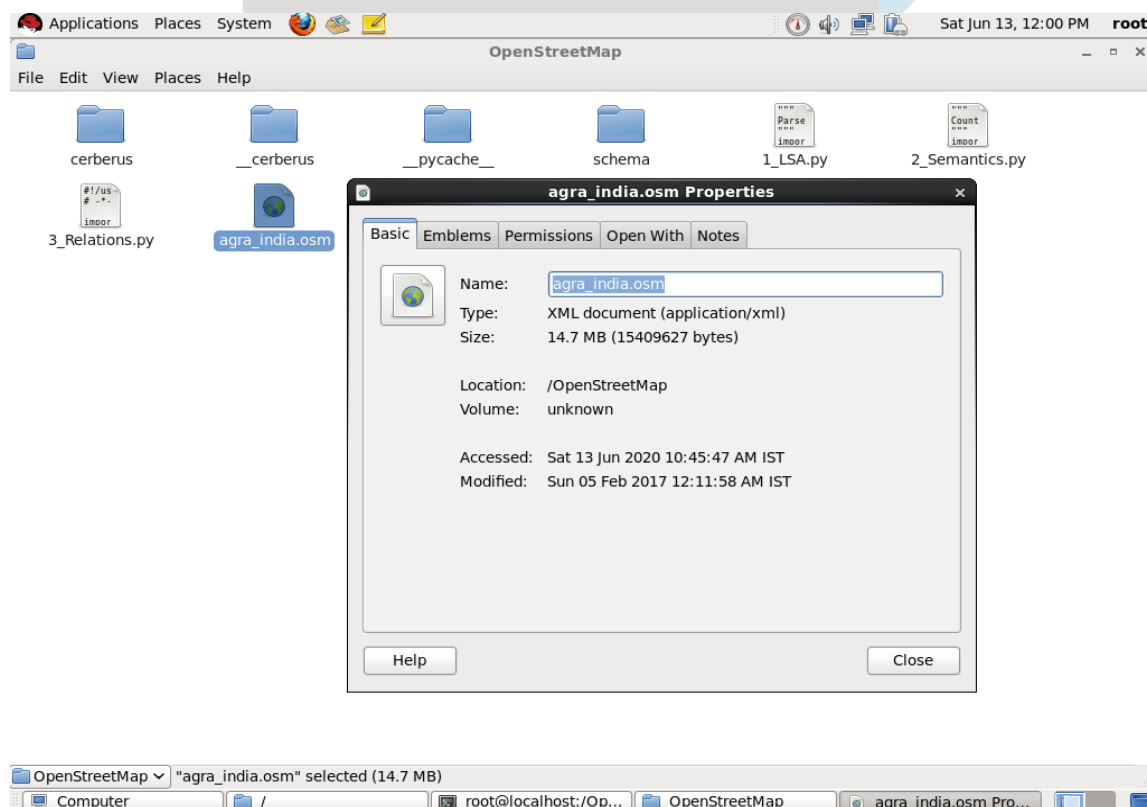


Figure 4.10 above Screen Depicts the Size approx 14.7 MB of Raw Corpus (agra\_india.osm) downloaded from openstreetmap.org

```
from nltk.corpus import stopwords
from nltk.stem.wordnet import WordNetLemmatizer
import scipy.io
from scipy import linalg
OSM_FILE = "agra_india.osm"
SAMPLE_FILE = "agra_india_sample.osm"

k = 100 # Parameter: take every k-th top level element

def get_element(osm_file, tags=('node', 'way', 'relation')):
    context = iter(ET.iterparse(osm_file, events=('start', 'end')))
    _, root = next(context)
    for event, elem in context:
        if event == 'end' and elem.tag in tags:
            yield elem
            root.clear()
```

Figure 4.11 Forming Nodes Relation using Ways

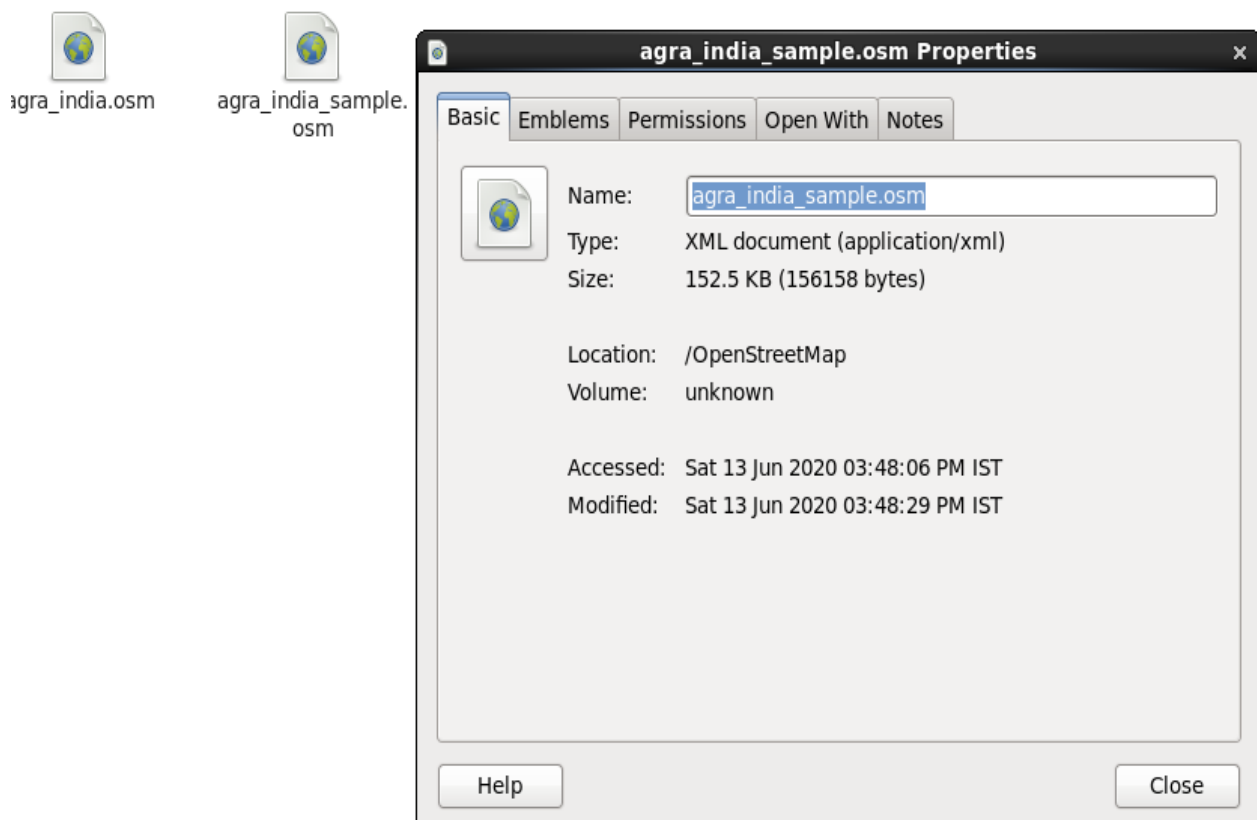
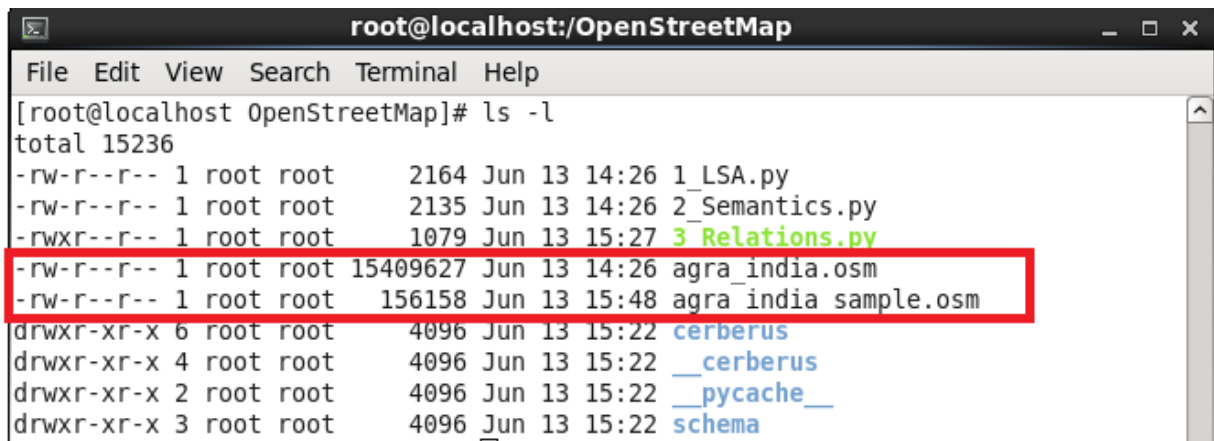


Figure 4.12 Latent Semantics Forming the Refined Corpus from Existing Corpus Size 152.5 KB (agara\_inda\_sample.osm) from 14.7 MB Raw Corpus (agra\_india.osm) using Semantic Relations





```

root@localhost:/OpenStreetMap
File Edit View Search Terminal Help
[root@localhost OpenStreetMap]# ls -l
total 15236
-rw-r--r-- 1 root root    2164 Jun 13 14:26 1_LSA.py
-rw-r--r-- 1 root root    2135 Jun 13 14:26 2_Semantics.py
-rwxr--r-- 1 root root    1079 Jun 13 15:27 3_Relations.py
-rw-r--r-- 1 root root 15409627 Jun 13 14:26 agra_india.osm
-rw-r--r-- 1 root root   156158 Jun 13 15:48 agra_india_sample.osm
drwxr-xr-x 6 root root    4096 Jun 13 15:22 cerberus
drwxr-xr-x 4 root root    4096 Jun 13 15:22 __cerberus__
drwxr-xr-x 2 root root    4096 Jun 13 15:22 __pycache__
drwxr-xr-x 3 root root    4096 Jun 13 15:22 schema

```

Figure 4.13 Results on File System Raw Corpus (agra\_india.osm of 15MB Approx) and Corpus With Semantic Relations (agra\_india\_sample.osm of 15 KB Approx)

#### 4.7 OPTIMIZATION OF CORPUS (OSM) USING GENETIC ALGORITHM

Under the scheme, genetic algorithms begin by initializing a set of randomly generated solutions. This set of solutions is called population. Each individual in the population is called a chromosome which describes a solution of the problem to be solved. A chromosome can be expressed in a symbol string, for example a collection of nodes. Chromosomes can change continuously called regeneration. For each generation, chromosomes are evaluated using a measuring instrument called the fitness function (level of fitness). To make the next generation, new chromosomes are called node formed by combining two chromosomes from the current generation using operators crossover / crossing or changing a chromosome by using a mutation operator. A new generation is formed by means of a selection made against parents and node based on the fitness value and eliminating the others. The more suitable chromosomes have a probability of being chosen. After several generations, this algorithm will converge towards the shape of the chromosome best, hoping to get it states the optimal solution of the problem being solved.

##### 4.7.1 General Structure of Genetic Algorithms

If  $P(t)$  and  $C(t)$  are the parent and node of  $t$  generation, the general structure of the genetic algorithm is as follows: Genetic algorithm procedure:

```

begin
  t ? 0;
  initialization P (t); P (t)
  evaluation;
  while (termination conditions not met) do recombination P (t) to produce
  children
  C (t); C (t)
  evaluation;
  selection P (t + 1) from P (t) and C (t); t ? t + 1;
end

```

##### 4.7.2 Operator and Evaluation Function

Usually, initialization is assumed randomly. Recombination involves crossover and mutations to produce node. In fact, there are only two types of operations in genetic algorithms, namely genetic operations ( crossover / crosses and mutations) and evolutionary operations ( selection). In the theory of evolution, this mutation is a chromosome operator that allows living things to adapt to their environment even though the new environment is incompatible with the original parent environment. The biggest factor in the theory of evolution that causes a chromosome to survive, extinct, make a cross or mutation is the environment. In genetic algorithms, environmental factors are played by the evaluation function. The evaluation function uses chromosomes as input and produces certain numbers that indicate the performance of the problem being solved. In the optimization problem, the evaluation function is the objective function ( objective function). The value of the evaluation function is called the suitability value ( fitness value). This value will determine whether a string will appear in the next generation or die. Below is the simulation depicts the above scenario using python code blocks with results.

##### 4.7.3 Selection

The selection will determine which individuals will be selected for recombination and how node formed from selected individuals. The first step done in this selection is the search for fitness values. There are several selection methods, including: Roulette Wheel Selection : The roulette wheel selection method is the simplest method, and is often also known by name stochastic sampling with replacement. As the name implies, this method mimics the roulette-wheel game in which each chromosome occupies a circle piece on the roulette wheel proportionally according to its fitness value. Chromosomes that have a greater fitness value occupy a larger circle than a chromosome with a low fitness value. Table 4.1 illustrates an example of using the roulette wheel method.

Fitness Value	Chromosomes	Probability
K1	1	0.25
K2	2	0.5
K3	0.5	0.125
K4	0.5	0.125
Amount	4	

Table 4.1 Example of using the Roulette Wheel Selection Method

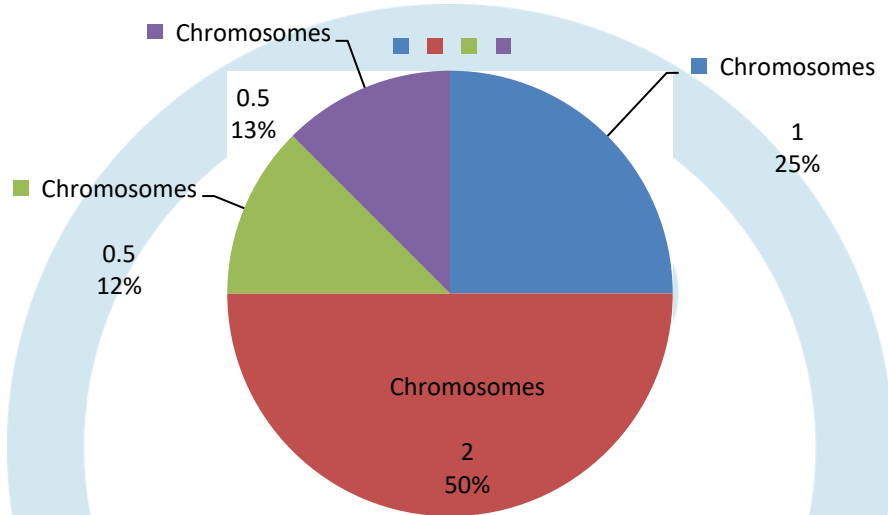


Figure 4.14 Graph Representation of above Table Using the Roulette Wheel Selection Method

```

root@localhost:/OpenStreetMap
File Edit View Search Terminal Help
[root@localhost OpenStreetMap]# python 4_Selection.py
['church': set(['church road'])]
church road => Church Road
[root@localhost OpenStreetMap]#

```

Figure 4.15 Selection Results using Genetic Algorithm

```

import xml.etree.cElementTree as ET
from collections import defaultdict
import re
import pprint
import random
import sys
import time

def _generate_parent(length, geneSet, get_fitness):
    genes = []
    while len(genes) < length:
        sampleSize = min(length - len(genes), len(geneSet))
        genes.extend(random.sample(geneSet, sampleSize))
    genes = ''.join(genes)
    fitness = get_fitness(genes)
    return Chromosome(genes, fitness)

def _mutate(parent, geneSet, get_fitness):
    index = random.randrange(0, len(parent.Genes))
    childGenes = list(parent.Genes)
    newGene, alternate = random.sample(geneSet, 2)
    childGenes[index] = alternate if newGene == childGenes[index] else newGene
    genes = ''.join(childGenes)
    fitness = get_fitness(genes)
    return Chromosome(genes, fitness)

OSMFILE_sample = "agra_india_sample.osm"
regex = re.compile(r'\b\S+\.?', re.IGNORECASE)

expected = ["Agra", "Agra Cant", "Hospital", "Cinema", "Power", "Tower", "Bus Stop", "Station", "Convenience", "Road", "NCR",
            "Fuel", "Building", "Gandhi", "Bridge", "Society"] #expected names in the dataset

mapping = {"agra": "Agra",

```

Figure 4.16 Selection using Genetic Algorithm

#### 4.7.4 Crossover

One of the most important components in genetic algorithms is crossing or crossover. Cross or crossover function combines two different parent strings into two different descendant strings its parent. A chromosome that leads to a good solution can be obtained from the crossing of two chromosomes. Crosses can also be bad if the population size is very small. In a very small population, a chromosome with genes that lead to a solution will spread very quickly to other chromosomes. To overcome this problem with a certain probability That is, the crossing can be done only if a random number [0,1) is raised less than contextual range which is determined. From the results of studies that have been done by genetic algorithm practitioners, it is suggested that the probability of crosses is high enough, which is 80% to 95% to give good results. For some specific problems the 60% crossing probability gives better results. As soon as the parent for the crossbreed is chosen, a crossbreeding operation is used to form a data into comma separated values from Corpus (agra\_india\_sample.osm).

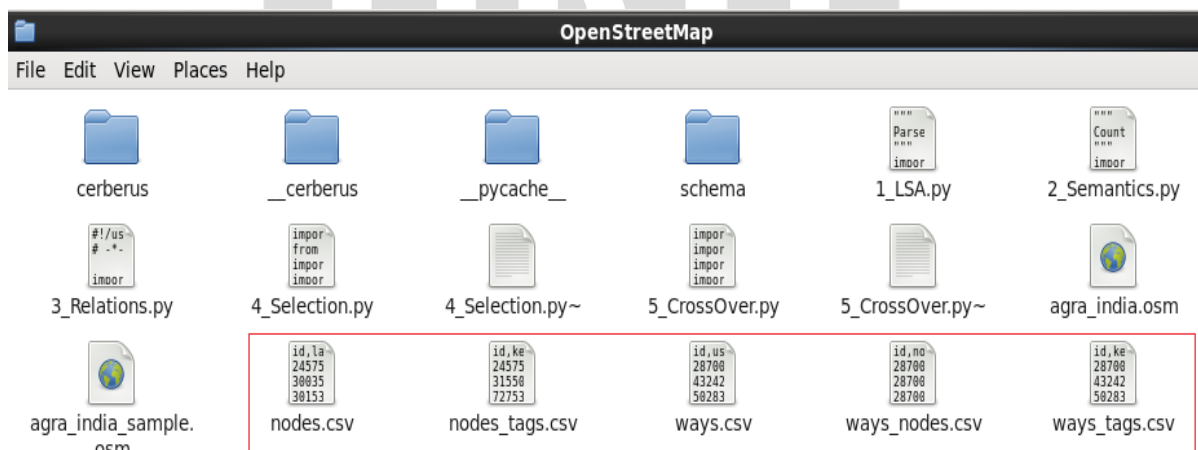


Figure 4.17: Crossover using Genetic Algorithm forming Comma Separated



```

id,lat,lon,user,uid,version,changeset,timestamp
245756314,27.1353758,78.0948156,PlaneMad,1306,3,31281402,2015-05-19T11:56:46Z
300351107,27.2002017,78.0565884,prolineserver,23057,1,706576,2008-09-27T16:30:21Z
301537598,27.1884708,77.9963677,jaimemd,404532,5,34240406,2015-09-25T08:06:47Z
301546843,27.1528344,77.9681801,prolineserver,23057,1,26253,2008-10-02T12:00:45Z
301548197,27.1484506,77.9406456,prolineserver,23057,1,26253,2008-10-02T12:13:13Z
301599830,27.1582821,77.9652701,prolineserver,23057,1,28184,2008-10-02T17:33:28Z
315503073,27.1975936,78.0084803,pratikyadav,2905914,13,32044875,2015-06-18T07:34:32Z
315505755,27.19252,78.0009043,chandusekharreddy,439726,44,36215922,2015-12-28T10:21:41Z
315507531,27.1734342,78.035218,matvey_kiev_ua,371387,2,11605255,2012-05-15T14:00:19Z
315509169,27.1611872,78.0583475,Rub21,510836,3,31338227,2015-05-21T09:36:16Z
315514760,27.1944547,78.0075078,pratikyadav,2905914,12,32045140,2015-06-18T07:52:29Z
315525981,27.1599996,78.000356,Divjo,63375,3,811638,2008-11-30T07:57:41Z
315722789,27.1758859,78.0053908,Geeta parmar,2328970,6,25794879,2014-10-01T17:50:22Z
315725049,27.1590963,77.9717603,Divjo,63375,3,811638,2008-11-30T07:52:18Z
315726537,27.1560874,78.0186754,Divjo,63375,1,811638,2008-11-30T08:04:10Z
315727702,27.2099911,77.9507221,ruthmaben,2554698,2,31253808,2015-05-18T12:30:12Z
316057454,27.211782,78.02989,pratikyadav,2905914,3,32004690,2015-06-16T11:50:58Z
316058322,27.1982055,78.0150953,Divjo,63375,1,10118,2008-12-01T15:25:17Z
316181726,27.1772414,78.0084061,jaimemd,404532,3,34948135,2015-10-29T12:54:45Z
316183909,27.1880624,78.0164604,ramyaragupathy,2823295,2,31340582,2015-05-21T11:33:14Z
366987624,27.1841792,78.0539332,MichaelCollinson,308,2,867419,2009-03-29T18:43:45Z
366989869,27.2558336,78.0972348,MichaelCollinson,308,1,867419,2009-03-29T18:47:50Z
366990089,27.2219508,78.0728159,MichaelCollinson,308,1,867419,2009-03-29T18:48:01Z
542663636,27.1892298,78.0305335,Sven L,105255,1,2969571,2009-10-27T22:09:04Z
542790438,27.1905649,78.0346836,Sven L,105255,1,2970186,2009-10-27T23:51:41Z
542790656,27.1979628,78.0427067,Sven L,105255,1,2970186,2009-10-27T23:51:46Z
542790941,27.2042932,78.0567241,Sven L,105255,1,2970186,2009-10-27T23:51:53Z
542834071,27.1876709,78.0183277,ramyaragupathy,2823295,2,31340984,2015-05-21T11:53:03Z

```

Figure 4.18: Nodes Extracted from Corpus using Crossover (GA)

```

id,key,value,type
245756314,source,AND,regular
315505755,name,st.johns crossing.,regular
727538589,ele,168,regular
727538589,iata,AGR,regular
727538589,icao,VIAG,regular
727538589,name,Agra Kheria Airport,regular
727538589,name_1,Kheria Airport,regular
727538589,source,wikipedia,regular
727538589,aeroway,aerodrome,regular
727538589,landuse,military,regular
727538589,en,Agra Airport,name
727538589,military,airfield,regular
727538589,operator,Airports Authority of India,regular
727538589,wikipedia,en:Agra Airport,regular
727538589,city_served,Agra,regular
727538589,country,India,is_in
727538589,type,military/public,aerodrome
1282866909,power,tower,regular
1282867872,power,tower,regular

```

Figure 4.19: Nodes with Tags Extracted from Corpus using Crossover (GA)

```

id,user,uid,version,changeset,timestamp
28700996,PlaneMad,1306,7,31340004,2015-05-21T11:11:03Z
43242766,sowjanya,2901480,6,39548603,2016-05-25T05:05:18Z
50283495,Oberaffe,56597,2,26277281,2014-10-23T12:56:56Z
92881405,PlaneMad,1306,8,31482143,2015-05-26T18:09:52Z
158316138,PlaneMad,1306,1,11210173,2012-04-07T07:25:13Z
175663885,NoelB,236361,2,12710718,2012-08-13T08:37:28Z
234852488,bdiscoe,402624,1,17479328,2013-08-24T04:57:53Z
242854404,PlaneMad,1306,3,31338475,2015-05-21T09:48:31Z
268635858,keepright!_ler,1731253,1,21287952,2014-03-24T16:09:25Z
300602926,Petr Dlouhý,17615,1,25067584,2014-08-28T07:29:56Z
300603028,Petr Dlouhý,17615,1,25067584,2014-08-28T07:30:03Z
303424013,Rajesh Diwakar,2329016,3,25471472,2014-09-16T05:59:06Z
303576776,pratikyadav,2905914,5,32004690,2015-06-16T11:50:55Z
303581159,pratikyadav,2905914,8,31999608,2015-06-16T07:27:21Z
303656308,Oberaffe,56597,4,25762701,2014-09-30T08:42:54Z
304486525,sowjanya,2901480,4,39468678,2016-05-21T12:07:00Z
304922251,stjohnscollege,3308183,4,34646322,2015-10-15T05:25:50Z
305104828,sowjanya,2901480,5,39618171,2016-05-28T05:56:10Z
305709106,veekesh yadav,2354635,1,25767537,2014-09-30T13:21:21Z
310304726,Warin61,1830192,1,26425921,2014-10-29T23:43:59Z
311885110,Warin61,1830192,2,26680855,2014-11-10T06:09:44Z
312538019,Warin61,1830192,1,26748941,2014-11-13T05:22:56Z
313016464,Warin61,1830192,1,26812105,2014-11-15T23:51:56Z
345944555,sowjanya,2901480,2,39683784,2016-05-31T09:03:22Z
345946915,sowjanya,2901480,2,39683784,2016-05-31T09:03:22Z
345954171,Chetan Gowda,2644101,1,31252512,2015-05-18T11:41:43Z
345955618,PlaneMad,1306,1,31252580,2015-05-18T11:44:38Z

```

Figure 4.20: Ways Extracted from Corpus using Crossover (GA)

#### 4.7.5 Mutation

After the crossing process is complete, then a mutation process is imposed on the corpus. Mutation is the process of changing the value of one or several genes in 1 chromosome. Mutations function in making changes that are not caused by crossing. If the process of selecting chromosomes tends to continue on good chromosomes, it is very easy for early convergence to occur, which is to reach the optimum local solution. To avoid early convergence and to maintain differences in chromosomes in the population, in addition to taking a more efficient selective approach, mutation operations can also be carried out. This mutation process is random, so it does not always guarantee that after the mutation process a better fitness chromosome will be obtained, but in the presence of this mutation it is hoped that the chromosomes obtained will have better fitness than before the mutation surgery.

Table 4.21: Example of the mutation process

Chromosomes before mutations	1 1 0 0 1 1
Chromosomes after mutations	1 1 0 1 0 0

However, mutation has a controversy in its application in genetic algorithms because of its random nature so that it can interfere with chromosomes with the best fitness that has been obtained. Sometimes mutation is still used with a very small probability that is  $P_m < 1$ . So the possibility of chromosomes undergoing changes due to mutations is very small. However, the code block below depicts the scenarios as per the scheme.

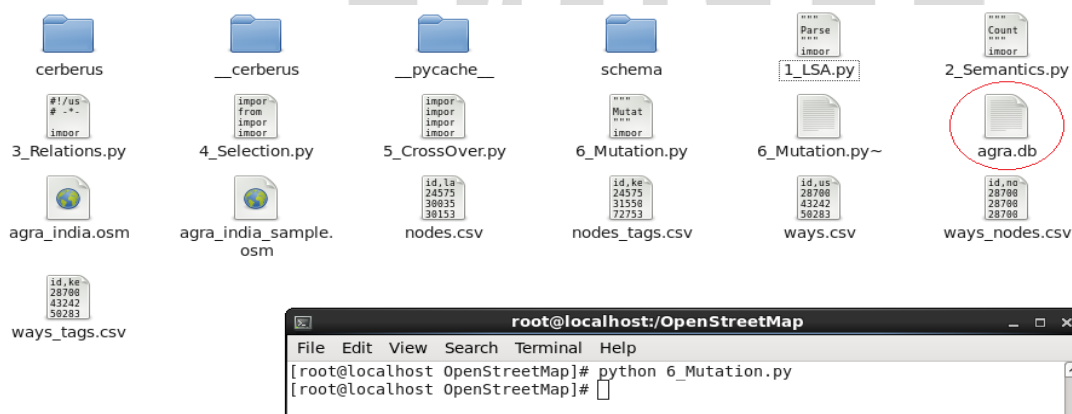


Figure 4.21: Process Mutation using GA forming Object Relational Model above Marked in Circle as agra.db

```

"""
Mutation of the CSV files with the repective table names.
"""

import csv, sqlite3
import random
import statistics
import sys
import time
from bisect import bisect_left
from enum import Enum
from math import exp
con = sqlite3.connect("agra.db")
con.text_factory = str
cur = con.cursor()
class Chromosome:
    def __init__(self, genes, fitness, strategy):
        self.Genes = genes
        self.Fitness = fitness
        self.Strategy = strategy
        self.Age = 0

# create nodes table
cur.execute("CREATE TABLE nodes (id, lat, lon, user, uid, version, changeset, timestamp);")
with open('nodes.csv','rb') as fin:
    dr = csv.DictReader(fin)
    to_db = [(i['id'], i['lat'], i['lon'], i['user'], i['uid'], i['version'], i['changeset'], i['timestamp']) \
              for i in dr]

cur.executemany("INSERT INTO nodes (id, lat, lon, user, uid, version, changeset, timestamp) \
VALUES (?, ?, ?, ?, ?, ?, ?, ?);", to_db)
con.commit()

```

Figure 4.22: Mutation of Comma Separated Value into Sqlite Schema using GA

## V. CONCLUSION AND FUTURE SCOPE

### 5.1 CONCLUSION

From the results using Latent Semantic Analysis and Genetic Algorithm research on Open Street Map based Corpus, There are several conclusions that can be drawn, namely:

1. Latent Semantic Model is implemented to remove noise from huge corpus and bind the relation among the nodes to form the ways and direction for espionage the best data model.
2. Genetic algorithms designed and implemented generally provide solutions that are near optimal. After testing and obtaining the results, it can be said that the Genetic Algorithm works well to get the optimal solution (minimizing completion time) with accuracy of 78.86%.
2. After conducting the experiment, it was suspected to solve and to provide the accurate information with the probability of the right crossover and mutation using Object Relation Model (Schema Based Model) from the Corpus. This result is not absolute because the solution using Genetic Algorithm in principle uses the rules of random selection. Therefore, it may vary time to time as per the desired information to be retrieved from the corpus or Open Street map file.

### 5.2 FUTURE SCOPE

From the results of the study Of Latent Semantic Analysis and Genetic Algorithm on OSM based corpus, there are several suggestions that can be taken, namely:

1. The comparison or relation can be formed using cosine similarity index along with Singular Value Decomposition which can speed up the process.
2. For further research, chromosome representation can be used in other forms, such as Job based representation, Preference-list based representation, Job-pair-relation-based representation, Priority-rule-based representation, Disjunctive graph based representation, Completion-time-based representation, Machine based representation or Random key representation for constant accuracy.
3. For further research, a cross operator / crossover other methods, for example job-based order crossover, partial mapped crossover, or other and mutation operations can also be done by other methods, for example inversion, insertion, or reciprocal exchange mutation.
4. Techniques like Adaptive Boost can be inculcated for swift and quick results.

## References

- [1] Michael P. Bishop, Brennan W. Young, Da Huo, Zhaohui Chi, Spatial Analysis and Modeling in Geomorphology, Reference Module in Earth Systems and Environmental Sciences, Elsevier, 2020, ISBN 9780124095489, <https://doi.org/10.1016/B978-0-12-409548-9.12429-7>.



- [2] Bankston Cotton, Environmental Psychology: Principles and Practices, Scientific e-Resources, Mar 4, 2019, ISBN 9781839474088 <https://learn.arcgis.com/en/arcgis-book>
- [3] Kathryn Keranen, Instructional Guide for The ArcGIS Imagery Book Lyn Malone, Esri Press, 380 New York Street, Redlands, California 92373-8100 Copyright © 2017, Esri, <https://downloads.esri.com/LearnArcGIS/pdf/instructional-guide-for-the-arcgis-imagery-book.pdf>
- [4] Murayama, Yuji, Thapa, Rajesh Bahadur, Spatial Analysis and Modeling in Geographical Transformation Process, <https://www.springer.com/gp/book/9789400706705>
- [5] Thapa, Rajesh & Murayama, Yuji. (2011). Spatial Analysis and Modeling in Geographical Transformation Process: GIS-based Applications. 10.1007/978-94-007-0671-2.
- [6] Andreano, Maria & Benedetti, Roberto & Piersimoni, Federica. (2019). A Distance Correlation Index of Spatial Dependence for Compositional Data. Papers in Regional Science. 10.1111/pirs.12451.
- [7] Congdon, Peter. (2019). Representing Spatial Dependence. 10.1201/9780429113352-6.
- [8] Şen, Zekai. (2016). Spatial Dependence Measures. 10.1007/978-3-319-41758-5\_5.
- [9] Manley D. (2014) Scale, Aggregation, and the Modifiable Areal Unit Problem. In: Fischer M., Nijkamp P. (eds) Handbook of Regional Science. Springer, Berlin, Heidelberg
- [10] Degbelo, A., Kuhn, W. Spatial and temporal resolution of geographic information: an observation-based theory. Open geospatial data, softw. stand. 3, 12 (2018)
- [11] Ndehedehe, Christopher & A, Ekpa & O, Okwuashi & Simeon, Ogunlade. (2013). UNDERSTANDING ERRORS AND THEIR MEASUREMENT IN GEOINFORMATION. Journal of Environmental Sciences and Resources Management. Volume 5. Pp. 74 - 87.
- [12] Jokar Arsanjani, Jamal & Zipf, Alexander & Mooney, Peter & Helbich, Marco. (2015). An Introduction to OpenStreetMap in Geographic Information Science: Experiences, Research, and Applications. 10.1007/978-3-319-14280-7\_1.
- [13] Mooney, Peter & Minghini, Marco. (2017). A review of OpenStreetMap data. 10.5334/bbf.c.
- [14] Mocnik, F., Mobasheri, A. & Zipf, A. Open source data mining infrastructure for exploring and analysing OpenStreetMap. Open geospatial data, softw. stand. 3, 7 (2018). <https://doi.org/10.1186/s40965-018-0047-6>
- [15] S. S. Sehra, J. Singh and H. S. Rai, "A Systematic Study of OpenStreetMap Data Quality Assessment," 2014 11th International Conference on Information Technology: New Generations, Las Vegas, NV, 2014, pp. 377-381, doi: 10.1109/ITNG.2014.115.
- [16] <https://en.wikipedia.org/wiki/OpenStreetMap>
- [17] Jokar Arsanjani, Jamal & Mooney, Peter & Zipf, Alexander & Helbich, Marco. (2015). An introduction to OpenStreetMap in GIScience: Experiences, Research, Applications.
- [18] Zhang L, Pfoser D (2019) Using OpenStreetMap point-of-interest data to model urban change—A feasibility study. PLoS ONE 14(2): e0212606. <https://doi.org/10.1371/journal.pone.0212606>
- [19] Mooney, P and Minghini, M. 2017. A Review of OpenStreetMap Data. In: Foody, G, See, L, Fritz, S, Mooney, P, Olteanu-Raimond, A-M, Fonte, C C and Antoniou, V. (eds.) Mapping and the Citizen Sensor. Pp. 37–59. London: Ubiquity Press. DOI: <https://doi.org/10.5334/bbf.c>. License: CC-BY 4.0
- [20] Seto, T.; Kanasugi, H.; Nishimura, Y. Quality Verification of Volunteered Geographic Information Using OSM Notes Data in a Global Context. ISPRS Int. J. Geo-Inf. 2020, 9, 372.
- [21] Jonathan Bright, Stefano De Sabbata, Sumin Lee, Bharath Ganesh, David K. Humphreys, OpenStreetMap data for alcohol research: Reliability assessment and quality indicators, Health & Place, Volume 50, 2018, Pages 130-136, ISSN 1353-8292, <https://doi.org/10.1016/j.healthplace.2018.01.009>.
- [22] Nurin Swasti, Taufik Hery Purwanto, Application of OpenStreetMap (OSM) to Support the Mapping Village in Indonesia, IOP Conference Series: Earth and Environmental Science, Published 1 November 2016
- [23] <https://www.openstreetmap.org/help>
- [24] [https://wiki.openstreetmap.org/wiki/Beginners%27\\_guide](https://wiki.openstreetmap.org/wiki/Beginners%27_guide)
- [25] P. Kherwa and P. Bansal, "Latent Semantic Analysis: An Approach to Understand Semantic of Text," 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), Mysore, 2017, pp. 870-874, doi: 10.1109/CTCEEC.2017.8455018.
- [26] [https://en.wikipedia.org/wiki/Latent\\_semantic\\_analysis](https://en.wikipedia.org/wiki/Latent_semantic_analysis)
- [27] Baker, Kirk. (2013). Singular Value Decomposition Tutorial. 2005.
- [28] Yongchang Wang and L. Zhu, "Research and implementation of SVD in machine learning," 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), Wuhan, 2017, pp. 471-475, doi: 10.1109/ICIS.2017.7960038.
- [29] Stewart, Sepideh & Thomas, Michael. (2006). Student thinking about eigenvalues and eigenvectors: Formal, symbolic and embodied notions. 487-495.
- [30] Lee, Shyi-Long & Yeh, Yeong-nan. (1993). On Eigenvalues and Eigenvectors of Graphs. Journal of Mathematical Chemistry. 12. 121-135. 10.1007/BF01164630.
- [31] Gene H. Golub, Henk A. van der Vorst, Eigenvalue computation in the 20th century, Journal of Computational and Applied Mathematics, Volume 123, Issues 1–2, 2000, Pages 35-65, ISSN 0377-0427,
- [32] Vasuki, A. (2020). Genetic Algorithm. 10.1201/9780429289071-4.
- [33] John H. Holland, Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence, <https://ieeexplore.ieee.org/book/6267401>
- [34] Kumar, Kaushik & Zindani, Divya & Davim, J.. (2019). Genetic Algorithm. 10.1201/9781351049580-2.

- [35] Chambers, Lance & Taylor, Michael. (2019). Genetic Algorithms. 10.4324/9780429437625-7.
- [36] Judson, Richard. (2008). Genetic algorithms Genetic Algorithms. 10.1007/978-0-387-74759-0\_218.
- [37] Patil, Lalit. (2019). Lecture 10 : Genetic Algorithms.
- [38] Alam, Tanweer & Dixit, Amit & Benaida, Mohamed. (2020). Genetic Algorithm: Reviews, Implementations, and Applications. 10.20944/preprints202006.0028.v1.
- [39] Kumar, Sandeep & Sharma, Harish & Jain, Er. (2018). Genetic Algorithms. 10.1201/9780429445927-2.
- [40] Awange, Joseph & Palancz, Bela & Lewis, Robert & Volgyesi, Lajos. (2018). Genetic Algorithms. 10.1007/978-3-319-67371-4\_5.
- [41] Schuppert, A. & Ohrenberg, A.. (2020). data mining. 10.1002/9783527809080.cataz04524.
- [42] Chowdhary, Prof. (2020). Data Mining. 10.1007/978-81-322-3972-7\_17.
- [43] Bramer, Max. (2020). Data for Data Mining. 10.1007/978-1-4471-7493-6\_2.
- [44] Olagoke, Lukman & Topcu, Ahmet. (2020). Data Mining (Preprint). 10.2196/preprints.20930.
- [45] Sigera, Suresh. (2019). Data Mining.
- [46] Guarascio, Massimo & Manco, Giuseppe & Ritacco, Ettore. (2018). Knowledge Discovery in Databases. 10.1016/B978-0-12-809633-8.20456-1.
- [47] Moro, Gianluca. (2020). La Metodologia CRoss-Industry Standard Process for Data Mining.
- [48] Nodeh, Mohsen & CALP, M. Hanefi & şahin, Ismail. (2020). Analyzing and Processing of Supplier Database Based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) Algorithm. 10.1007/978-3-030-36178-5\_44.
- [49] Lashonda, Dr & Warne, Tara. (2020). Utilizing the Cross-Industry Standard Process for Data Mining to Detect Unique Retention Patterns of First-Time Freshmen.
- [50] Dewa, P & Mulyanti, Budi & Widiaty, I. (2020). Geographic information system in education. IOP Conference Series: Materials Science and Engineering. 830. 042097. 10.1088/1757-899X/830/4/042097.
- [51] Piovan, Silvia. (2020). Geographic Information Systems. 10.1007/978-3-030-42439-8\_6.



**IJRTI**