TELKOMNIKA, Vol.16, No.1, February 2018, pp. 334~342
ISSN: 1693-6930, accredited **A** by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v16i1.6985

Analysis of S2 (Spherical) Geometry Library Algorithm for GIS Geocoding Engineering

Risma Ekawati*, Untung Suprihadi
Sekolah Tinggi Teknologi Jakarta, Indonesia
Corresponding author, e-mail: risma@sttj.ac.id, untung@sttj.ac.id

Abstract

Geocoding is a common technique to transform address information into digital latitude/longitude format. One of the engineering conversions can be used is Google Maps based on S2 (Spherical) Geometry Library algorithm. This journal explains the quality analysis of the algorithm using geocoding quality matrix testing from hundreds of address data samples particularly on three cities in Indonesia-Jakarta, Bandung, and Balikpapan. However, the result of this research concludes that completeness of address information will affect its overall fourth matrix quality and the linkages of it such as transform success rate, landmark exactness, the score of accuracy and range of radius in meter.

Keywords*: geocoding, google maps, geographic information system*

1. Introduction

Recently, Geographic Information System (GIS) has become a part of digital life that could not be separated from all of activites. So many GIS development methods had been implemented to produce maps services in order to support operational business or stands to support decision activity for some organizations. For example the availability of GIS and GPS (Global Positioning System) used up combination can improve the added value of business organization extensively [1]. Furthermore, the satellite sensing calibration combination also makes a great and efficient way to monitor the economic modernization [4].

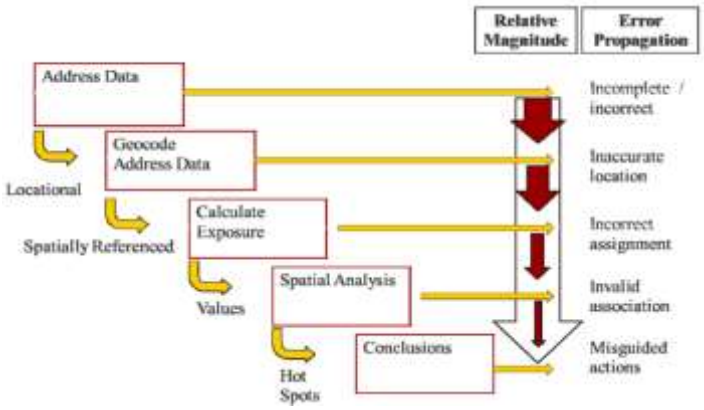

Figure 1. Geocoding conversion method [11]

Based on economic research, 80% data owned by business organization contains geographic location information such as address data, sales area boundary or distribution route [7] or at least there will be a single address column inside customer, distributor or even supplier database that keeps residence information or sales location.

Those address data is valuable organization asset and obtained to be a powerful resource for a relied GIS by converting the address into a digital location. This process itself

Received October 16, 2017; Revised January 4, 2018; Accepted January 12, 2018

called geotagging or geocoding. Geocoding in Figure 1 explains a subsequent conversion process of textual location information (address or place name) into geographic digital data representation [5].

Geocoding engineering process generally built-in from API (Application Programming Interface) facility of online maps service provider such as Google Maps, Bing Maps, OpenStreet Maps, MapQuest, etc. They also provide howto's in order to make GIS system developer easier to use the service. Figure 2 explains that Google Maps is the most favorite utilized online maps service instead of others used by particularly GIS system developer.

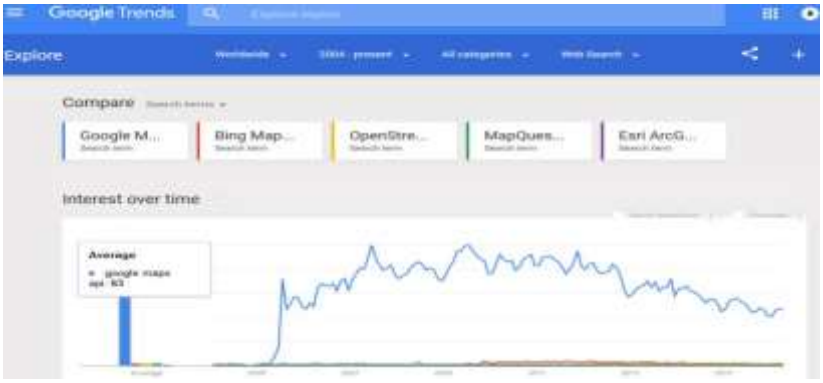

Figure 2. Online maps service utilization comparison trend

Google Maps APIs provides several online maps services such as maps visualization, distance tracker, geocoding, cluster maps and other features. S2 (Spherical) Geometry Library algorithm is powering behind the Google Maps geocoding API extraction process [11]. Geocoding quality result is the main objective focus for GIS developer system [5] because accuracy issues are always to be the most critical point inside geocoding process result-accuracy between the latitude/longitude coordinate and the physical address. This is important since customer maps system in business organization-for example-could not ensure whether how accurate its location compared to geocoding data.

In comparison, a 2014 research about geocoding process based on R algorithm was tested and measured according to geocoding accuracy issues with the result as shown in Table 1 [16]. This current research would be revealed the analysis of Google Maps based S2 (Spherical) Geometry Library geocoding algorithm by testing and measuring the accuracy issues according to geocoding quality matrix from given size of address data sample also conclude the resulting test and addressing some solutions to increase the accuracy of geocoding result.

2. Literature Review

According to Mishra et al, GIS is a system used to gather, integrate, analyze and processing geographic data [8]. Meanwhile, Patil said that GIS is the combination between maps contains digital location data, statistical data analysis and database technology [7]. Her previous research, Risma Ekawati concluded that GIS visualized digital data in maps representation [6]. In business field recently, GIS integrated with marketing programs contains data observatory purpose for sales intelligent system, sales decision support system and sales area coverage analysis [2].

Modern GIS technology support to manage, display and explore business location information, thus evolved from powerful location marked into critical business supporting tool [15]. Instead of business purposes, GIS also used in some other public sectors such as telecommunication, agriculture, health, crime analysis, traffic monitoring, government, research and development, defense system, etc. For example, an effort to decrease air pollution and increase better air quality in Jakarta, Indonesia government supported by GIS system [8].

This research adopted by previous research done by Williams, L. and Wilkins in 2014. While they used R algorithm for the same geocoding process purposed and the result of its quality as shown in Table 1.

Table 1. Accuracy Comparison Hasil Geocoding Result Using R Algorithm [16]

Input Address or Place to be Geocoded	R Function Call	Returned Latitude, Longitude	Returned Formatted Address	Returned Acuracy[a]	Returned Partial Match[b]
400 N Broad St, Philadelphia, PA	gGeoCode("Address", api="maps", accuracy=T, partmatch=T)	39.9599519,-75.1621758	400 North Broad Street, Philadelphia, PA 19132, USA	ROOFTOP	N/A
400 N Broad, Philly	gGeoCode("Address", api="maps", accuracy=T, partmatch=T)	39.9599519,-75.1621758	400 North Broad Street, Philadelphia, PA 19132, USA	ROOFTOP	TRUE
1503RACESTREET PHILADELPHIA[c]	gGeoCode("Address", api="maps", accuracy=T, partmatch=T)	39.9564906,-75.1646132	1503 Race Street Philadelphia, PA 19102, USA	RANGE_INTERPOLATED	N/A
Y-HEP[d]	gGeoCode("Address", api="maps", accuracy=T, partmatch=T)	46.8108711,7.1589661	HEP I, 1700 Fribourg, Switzerland	APPROXIMATE	TRUE
Y-HEP[d]	gGeoCode("Address", api="places")	39.954985,-75.163255	112 North Broad Street, Philadelphia, PA,19134, USA	N/A[e]	N/A[e]
St. Christopher's Hospital for Children	gGeoCode("Address", api="maps", accuracy=T, partmatch=T)	40.00609,-751257141	3601 A Street, Philadelphia, PA, 19134, USA	APPROXIMATE	N/A
St. Christopher's Hospital for Children	gGeoCode("Address", api="places")	40.00609,-75125714	3601 A Street, Philadelphia, PA, 19134, USA	N/A[e]	N/A[e]

[a]Informs whether the geocode was approximate, interpolated, or exact.
[b]Informs whether the address was completely or partially matched.
[c]This address does not actually exist.
[d]Youth Health Empowerment Project (Y-HEP) is a youth outreach organization located in Philadelphia.
[e]Accuracy and partial match options not available for Google Places Application Programming Interfaces.

According to Venkatesh, GIS consists of 5 elements: software, hardware, geographic data, public data and organization [15]. Meanwhile, Mishra et al concluded that GIS contain 5 elements: software, hardware, data, people and method [8]. In recent times, the elements combination brought significant changes for GIS development technology [6]. For the last decades, Google, OpenStreetMap, and Bing made improvements to their free online maps services. But paid maps service companies like ArcGIS or QGIS also joining the battle competition of GIS technology and creating their modern customized GIS features to get more particular GIS system developer as the customer.

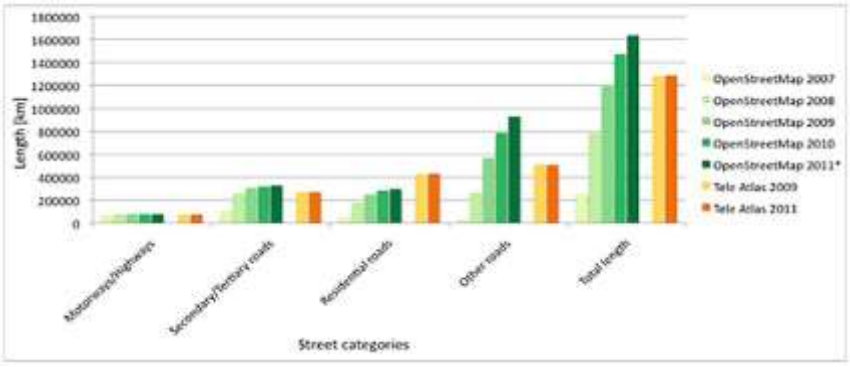

Figure 3. 2007-2011 OpenStreetMap road maps length comparison [9]

In fact, Google produced hundreds powerful API's to their maps customers [6]. From period 2005 to 2012, at least there were 800.000 peoples uses Google Maps API for their GIS purposes [3]. OpenStreetMap for another example: a great significant amount of road maps length in Germany increased by the community during 2007-2011 in Figure 3 [9]. Therefore as a conclusion, GIS utilization trend would increase year by year and would prepare for future scientific geographic data transformation, including business needs purpose.

In another phrase, geotagging (or geocoding) described as identification process and extraction of data entity and transform it into geographic content such as people, organization, and location [10]. Basically, geocoding conversion processed based from 3 references layer of geographic object data such as single location point, road segmented data (address, city, postcode) and areal unit group (geographic polygon object) [5]. As the output result, geocoding conversion process from a single address given will produce a group of numbers as known as latitude/longitude data. Google Maps as the most favorite free online maps service provider also provides geocoding extraction process based on S2 (Spherical) Geometry Library algorithm [11].

Maps visualization has the ability to shows a basic quality level of geocoding process. Both models of map visualization used most for analytical study and quality control of multiple addresses geocoding process are heat maps and cluster maps. Heat maps visualization model as shown in Figure 4 represents graph from individual data in different colors [12].

(a) (b)

Figure 4. Maps visualization model (a) Cluster maps and (b) Heat maps [12]

Heat maps and cluster maps are often used to visualize a group of latitude/longitude data results from geocoding process [3]. Both maps model (Figure 5)using the same statistical pattern recognition as visual identification methodology but with different approach such as: (a) Unsupervised Classification (cluster), used when there is no more data available to process (less of relationship between points or nothing at all); and (b) Supervised Classification (heat), used when there are more data available to process, by using discriminant function g(x), where x shows points under n-dimension Euclidean metric that complied gi(x) > gj(x) for each points relationship [4].

(a) (b)

Figure 5. Statistical pattern recognition used in (a) Unsupervised classification (cluster) and (b) Supervised classification (heat) [4]

Statistical pattern recognition by cluster approach requires a minimum basic system but still able to show relevant visualization [4]. However, Table 2 explains that heat maps visualization model is more intuitive and accurate in visualizing data representation [14]. Heat maps API from Google Maps is based on geocode clustering data algorithm that similar to K-means clustering but faster to process than Fuzzy C-Means clustering algorithm [13].

Tabel 2. Comparison of Cluster and Heat Accuracy Level [3]

Map	Minsk		Tiraspol	
Type	Correct (%)	Incorrect (%)	Correct (%)	Incorrect (%)
Cluster	47.50	52.50	37.50	62.50
Heat	60.00	40.00	55.00	45.00

*) Minsk is under Belarusian area and Tiraspol is under Moldova area

Figure 6 shows scatter graph from 3 clusters K-Means proves that geocode clustering data algorithm is faster to process heat maps visualization. K-Means clustering basic principle are points grouping, data initialization, data classification, centroid calculation and convergent criteria [13].

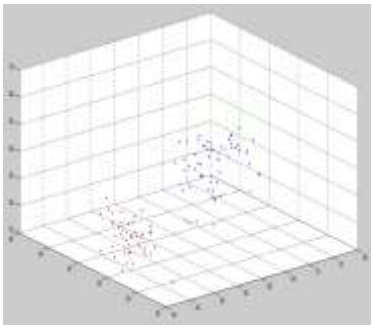

Figure 6. K-Means scatter graph using 3 cluster [13]

3. Methodology

This research methodology starts from a customized-build software of address geocoding converter prototype based on Google Maps S2 (Spherical) Geometry Library algorithm to produce digital latitude/longitude format and shows the result into heat maps model visualization. Geocoding data result also tested using GPS (Global Positioning System) hardware based on fourth geocoding result matrix quality [5]: 1) Transform success rate, shows the proportional success value of the geocoding conversion result; 2) Landmark exactness, shows exactness percentage between geocode data and physical landmark (eg: street, post code, building, etc); 3) Score of accuracy, shows similarity level between location and geographic reference; and 4) Range of radius, shows distance accuracy level between geocoding location and real exact location.

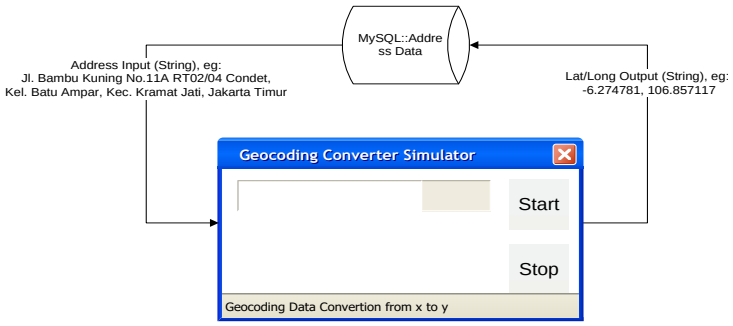

Figure 7. Prototype of Geocoding conversion system

As shown in Figure 7, MySQL database used to keep both input address data and output geocoding data (latitude/longitude). The addresses data are stored in MySQL database with information table and data type as follow:

Table 3. Address Data Table

Field No	Field Name	Data Type	Information
1	no	Integer (5)	Sequential data number (primary key)
2	name	Varchar (50)	Name of Address Owner
3	prov	Varchar (50)	Province of address
4	address	Varchar (200)	Address in string
5	latlong	Varchar (50)	Address in latitude/longitude format
6	sts	Integer (1)	Status 0: not processed and 1: processed

Figure 8 shows heat maps visualization based on geocoding data result in order to support the final objective research test.

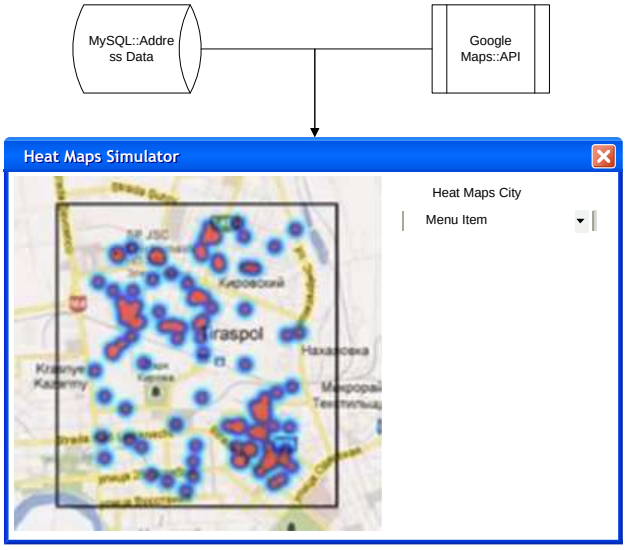

Figure 8. Prototype of heat maps visualization based on google maps API

4. Result and Discussion

There are 6.188 of data addresses population from 3 major cities in Indonesia involved in this research (Balikpapan city: 310 addresses, Bandung city: 1.646 addresses and Jakarta city: 4.232 addresses).

Figure 9 explains geocoding process of those 6.188 addresses data from customized protoype software based on API Google Maps geocoding S2 (Spherical) Geometry Library built using Borland Delphi 6.0 compiler. The following Delphi code shows geocoding routines call to the API:

Figure 9. Address geocoding software prototype

```
$url="https://maps.googleapis.com/maps/api/geocode/json?key=AIzaSyBhsZZKU5Xwxp7jwterBGpX9zFpukkIGGo&address=".urlencode("PONDOK KARYA AGUNG TC 56-57, Gunung Bahagia, Balikpapan Selatan, Balikpapan, Kalimantan Timur");
$json=file_get_contents($url);
$data=json_decode($json, TRUE);
echo $data["results"][0]["geometry"]["location"]["lat"] . ", " .
$data["results"][0]["geometry"]["location"]["lng"];
```

Figure 10 describes heat maps visualization as the result of geocoding process. It shows 4.232 addresses in Jakarta and 1.646 in Bandung. This heat maps model helps to visualize the distribution of the addresses on this research in each different colors: red, yellow and green representing the density level of latitude/longitude coordinate points. In order to test the accuracy level of the sample addresses, the research uses GPS Garmin eTrex Touch 25 particularly model.

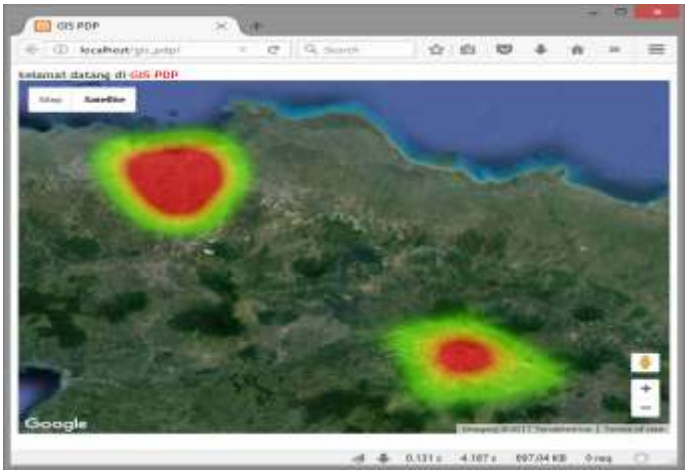

Figure 10. Heat maps visualization software prototype

While sample data sizes from the population are calculated based on Slovin formula:

$$n = \frac{N}{1+Ne^2}$$

N means population with 5% error rate in each city. The result of the GPS test (based on fourth geocoding matrix quality [5]) as shown on the following Table 4.

Tabel 4. Combination of GPS Test Result with Fourth Geocoding Matrix Quality

City	Addresses Population	Sampling (Slovin)	Transform success rate	Landmark exactness	Score of accuracy	Range of radius (m)
Balikpapan	310	175	100%	44,000%	25,143%	82,114
Bandung	1.646	322	100%	34,783%	16,460%	85,848
Jakarta	4.232	365	100%	52,600%	35,068%	71,759
Total	6.188	862	100%	43,795%	25,557%	79,907

Based on Slovin formula, there are 862 sample addresses to test from the total of 6.188 addresses population. Table 4 describes that 100% of 862 sample addresses are successful to transformed (geocoding process) using Google Maps based on S2 (Spherical) Geometry Library algorithm.

The algorithm has the ability to identify overall landmark exactness by street name identification at average 43,795% percentage rate: Jakarta 52,600%, Balikpapan 44,000%, and Bandung 34,783%. From GPS test, similarity rate between location and geographic references reach at average number of 25,557% percentage score: Jakarta 35,068%, Balikpapan 25,143% and Bandung 16,460%. While the overall score of radius range (in meters) shows average 79,907 m: Jakarta 71,759m, Balikpapan 82,114m, and Bandung 85,848m. From the test result, those last third matrix elements (Landmark exactness, Score of accuracy and Range of radius) are actually related to each other. It proves that low percentage rate of landmark exactness followed by lower percentage rate score of accuracy and higher range of radius between latitude/longitude and real position.

During research test, some parameters known affect the result test such as address data completeness, different numbering between I and 1 inside the address information and presumption that Google Maps unable to identify some of apartement name, housing name and building name from the 3 cities as the object of the research including block and number information from the address. In order to increase the quality of geocoding result, it is recommended to complete the address parameters since Google Maps doesn't have housing number data.

5. Conclusion and Suggestion

The research has demonstrated the analysis of S2 (Spherical) geometry library algorithm from given 862 sample addresses of total 6.188 population based on Slovin statistical formula. As the conclusion of this research, population and sample are known would not affect the fourth geocoding quality matrix significantly. Also, there is a direct proportional of relationships between last third matrix elements (landmark exactness, the score of accuracy and range of radius) that depends on address data information completeness. Heat maps visualization model made easier to analyze cluster distributed addresses as the result of geocoding process. This research shall continue to compare addresses aside from Indonesia to examine the consistency result of this analysis of S2 (Spherical) geometry library algorithm.

References

[1] A. G. Matani, M. S. Tripathi, & Pallavi M. Information Technology Tools Improving Supply Chain Management Productivity in Food Processing Industries. *International Journal of Advanced Engineering Technology*, 2012; III(1): 237-238.

[2] A. Quist-Aphetsi Kester. An Integrated Geographic Information System and Marketing Information System Model. *International Journal of Advanced Technology & Engineering Research (IJATER)*, 2012; 2(6): ISSN No: 2250-3536.

[3] Craig A, Richard T, Edward R. The Usability of Online Data Maps: An Ongoing Web Based Questionnaire Investigation into User's Understanding and Preference for Geo-Spatial Visualisations. *Research Councils UK Digital Economy Programme University of Southampton*. 2013.

[4] Dana Klimešová, Eva Ocelíková. Knowledge Management Improvement Using GIS. *International Journal of Mathematics and Computers in Biology, Business and Acoustics*. 2011. ISBN: 978-960-474-293-6.

[5] Daniel WG, Myles GC. The Effect of Administrative Boundaries and Geocoding Error on Cancer Rates in California. *International Journal of National Institutes of Health.* 2013; 3(1): 39–54.

[6] Ekawati R, Suharjito. Thematic Mapping Visualization of Geographic Information System using API Fusion Tables and Google Maps Integration. Proceedings the 10th International Conference on Knowledge Information and Creativity Support Systems (KICSS 2015). Phuket. 2015.
[7] Kavita KM, Gouri Patil. Geographic Information System (GIS) – for Business Analytics. *International Journal of Scientific & Engineering Research*. 2011; 2(11). ISSN 2229-5518.
[8] Mishra S, P Chandekar. Study of Geographical Information System and its Applications. *International Journal of Environmental Engineering and Management*. 2013; 4(5): 451-456.
[9] Pascal N, Dennis Z, A Zipf. The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. *International Journal of Future Internet*. 2012; 4: 1-21. (ISSN: 1999-5903).
[10] Rongjian L, Marco DA, Hanan S. Spatio-Temporal Disease Tracking Using News Articles. *International Workshop on Use of GIS in Public Health* (ACM SIGSPATIAL Pubs). 2014.
[11] Shaw B. Learning to Rank for Spatiotemporal Search. *Journal of International Conference on Web Search and Data Mining.* 2013. ACM: 978-1-4503-1869-3.
[12] Shilin Z. Advanced Heat Map and Clustering Analysis Using Heatmap3. *BioMed Research International* (Hindawi Publishing Corporation). 2014. Article ID 986048.
[13] Soumi G, Sanjay KD. Comparative Analysis of K-Means and Fuzzy C-Means Algorithms. *(IJACSA) International Journal of Advanced Computer Science and Applications.* 2013; 4(4): 35.
[14] Tinashe SC. Crime Prediction: Integration of Data Mining Techniques (Clustering and Classification) to Enhance Crime Prevention Through Analysis of the Relationships Between Type of Crimes, Locations, Times and Weather Patterns. Dissertation: National College of Ireland. 2014.
[15] Venkatesh J, SP, Aarthy C. GIS in Indian Retail Industry-A Deliberate Tool. IRACST - *International Journal of Computer Science and Information Technology & Security (IJCSITS)*. 2012; 2(3). ISSN: 2249-9555.
[16] Williams L, Wilkins. A No-Cost Geocoding Strategy Using R. *International Journal of Epidemiology*. 2014; 25(2). ISSN: 1044-3983.