# Automated Visual Clustering Functionality for Improved and Effective Mining of GIS Data

Carolyne Wanjiru Kimani
Institute of Computer Science and Information Technology
Jomo Kenyatta University of Agriculture (JKUAT)
Nairobi, Kenya

Joash Kiprotich Bii
Institute of computer Science and Information Technology
Jomo Kenyatta University of Agriculture (JKUAT)
Nairobi, Kenya

Prof. Waweru Mwangi (Phd)
Institute of Computer Science and Information Technology
Jomo Kenyatta University of Agriculture and Technology
Nairobi, Kenya

***Abstract*— The technical progress in computerized data acquisition and storage has resulted in the growth of vast databases for Geographical Information Systems. This has led to continuous increase and accumulation of huge amounts of the computerized data that have far exceeded human ability to completely interpret, analyze and use. In order to understand and make full use of these data repositories, various techniques have been put forward. However, these techniques are not fully reliable as they are not as efficient or of high performance as is expected. This thesis attempts to improve on the efficiency of existing spatial data mining techniques to ensure more efficient and high performance spatial data mining functionality in the present framework and tools used for spatial data mining. This will be done by integrating various techniques with available technologies.**

**The focus of this project is on improving performance and efficiency of spatial clustering, one of the commonly used spatial data mining methods by integrating visualization into clustering with an aim to provide an interactive, efficient and user-friendly approach to this important process for GIS data.**

***Keywords—GIS; spatial; data minig; clustering; visualization***

## I. Introduction

A Geographical Information System (GIS) is a System which involves capturing, storing, processing, manipulating, analyzing, managing, retrieving and displaying data /information which is referenced to the real-world or the earth (i.e. geographically referenced) [13]. Advances in Geographical Information Systems and supporting data collection technologies have resulted in the rapid collection of a huge amount of spatial data. This has resulted in a major problem where GIS are lacking a structured method for organizing this massive amount of data to make it most useful. This also complicates the ability to understand and carry out analysis on this data as it may not show detailed relationships and patterns which are crucial in spatial data. Spatial data in GIS is defined as elements that can be stored in a map, images, graph and tabular forms.

Data mining is the science of extracting useful information from large sets of data[ 9]. It can also be defined as the process of identifying or discovering useful and as yet undiscovered structure in data.

The main goal of data mining is to search for deeply hidden information that can be turned into knowledge for strategic decision making [14].

Spatial Data Mining (SDM) is the extraction of knowledge, spatial relationships or other interesting patterns that are not explicitly stored in databases, which store a large amount of space-related data such as maps, preprocessed remote sensing or image data. Spatial data mining is a process which tries to find patterns in geographic data.

Extracting useful and interesting patterns from massive geo-spatial datasets is important for many application domains, such as regional economics, ecology and environmental management, public safety, transportation, public health, business, and travel and tourism, because space is everywhere [17].

This work focuses on reducing the computational complexity and improving efficiency of spatial clustering, one of the common spatial data mining techniques.

## II. LITERATURE REVIEW

This section introduces some basic principles and concepts of GIS and spatial data mining techniques, mainly clustering that are relevant for this work and that can be used in developing some practical applications for use by groups involved in data mining.

Geographic Information Systems are essentially digital maps linked to database management systems which can be used for the purposes of displaying and querying information, carrying out spatial analysis and assisting in the decision-making process ;

**Components of GIS**

1. Software - these are the computer programs that provide the functions and tools needed to store, analyze, and display geographic information.

Examples include; ArcView, ArcGIS, Quantum GIS, Python and SQL.

2. Hardware - refers to the computer components on which a GIS operates. Hardware components for GIS can be categorized into four major types:

• Input devices, which includes digitizer, scanner, keyboard

• Storage devices includes, hard disc, CD ROM, memory sticks

• Processing devices or processor, and

• Output device includes printers, plotter, and monitor.

3. Data

This includes locations and other characteristics of natural features and human activities on, above and beneath the earth’s surface which are recorded and stored as geographic data for GIS.

A GIS database enables both spatial and non-spatial data to be stored and retrieved. Spatial data have a physical geographical location associated with them. The database is a fully integrated relational database with the additional capability to link with a graphical user-interface (GUI).

4. Users

These are the skilled people to manage the system and develop plans for applying it to real-world problems as well as all other users of the system. They range from technical specialists who design and maintain the system, to those who use it to help them perform their everyday work.

The user of a GIS system has the capacity to interact with the GIS database. He/she can input data, query the database or update the information stored. ‘What if’ analysis can also be conducted through an appropriate combination of queries to generate useful information [2].

### *A. Data Mining*

Data mining integrates several fields including; Machine Learning, Database Systems, Artificial Intelligence, Pattern Recognition, Data Visualization, Information Theory and Statistics.

The driving factor for this research in spatial data mining is the increase in collection of spatial data through business and geographical database systems [12]. Some of the spatial data collected include remotely sensed images, geographical information with spatial attributes such as location, mobile phone usage data, and medical data. Spatial data mining takes into account spatial and non-spatial data attributes in these data [15].

Common patterns discovered by data mining algorithms include descriptive patterns such as clustering, explanatory patterns such as association rules and predictive patterns such as classification rules and decision trees [10]. Data mining commonly involves four main classes of tasks which are:

- ***Clustering*** - the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- **Classification** – it involves generalizing known structure to apply to new data. For example, an email program might attempt to classify an email as legitimate or spam. Common algorithms include decision tree learning, nearest neighbor, naive Bayesian classification, neural networks and support vector machines.
- **Regression** - Attempts to find a function which models the data with the least error.
- **Association rule learning** - Searches for relationships between variables.

### *B. Spatial Data Mining and Clustering*

Spatial data is data pertaining to the location, shape and relationship among natural and constructed geographical features and boundaries. For instance, natural geographical features and boundaries in a space include oceans, river etc. Constructed geographical features and boundaries include cities, towns and villages. These features are represented by points, lines, and polygons.

Spatial data mining is a process of automating the search for potentially useful patterns.

At present, Geographic Information System (GIS) is a main tool for storing, processing and displaying spatial data [4]. GIS is a good spatial analysis tool, however, with the accumulation of spatial data, the spatial analysis function that GIS offers is not enough. Spatial data mining, which can automatically discover implicit knowledge from spatial data, has recently received wide attention. Lots of spatial data mining methods and systems have been developed. [11]

Several methods that integrate GIS and data mining techniques have been proposed and developed, such as SPIN , S-PLUS for ArcView GIS. [5]

Clustering analysis, also called segmentation analysis or taxonomy analysis, aims to identify homogeneous objects into a set of groups, named clusters, by given criteria.It is a very important technique of knowledge discovery for human beings.

Clustering is considered as an unsupervised classification process. The clustering problem is to partition a dataset into groups (clusters) so that the data elements within a cluster are as similar as possible to elements in the same cluster and as different from elements in different groups/clusters. Cluster analysis is a widely applied technique in data mining. However, most of the existing clustering algorithms are not efficient in dealing with arbitrarily shaped distribution data of extremely large and high-dimensional datasets. On the other hand, statistics-based cluster validation methods incur very high computational cost in cluster analysis which prevents clustering algorithms from being effectively used in practice.

There is no universally applicable clustering technique in discovering the variety of structures display in data sets. Also, a single algorithm or approach is not adequate to solve every clustering problem. There are many methods available, the criteria used differ and hence different classifications may be obtained for the same data. While larger and larger amounts of data are collected and stored in databases, there is increasing the need for efficient and effective analysis methods. Grouping or classification of measurements is the key element in these data analysis procedures. [3]

Modern Examples of Spatial Patterns

1) Cancer clusters to investigate environment health hazards

2) Crime hotspots for planning police patrol routes

3) Bald eagles nest on tall trees near open water

4) Nile virus spreading from north east USA to south and west

5) Unusual warming of Pacific ocean (El Nino) affects weather in USA

The goal of spatial data mining is to automate the discoveries of such patterns which can then be examined by domain experts for validation.

Results of Data Mining Include:

- Forecasting what may happen in the future
- Classifying people or things into groups by recognizing patterns
- Clustering people or things into groups based on their attributes
- Associating what events are likely to occur together
- Sequencing what events are likely to lead to later events

Spatial clustering is a process of grouping a set of spatial objects into groups called clusters. Objects within a cluster show a high degree of similarity, whereas the clusters are as much dissimilar as possible[3]. Clustering does not rely on predefined labels of classes or a priori given number of classes.

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups.

Clustering spatial data is the process of grouping similar objects according to their distance, connectivity, or their relative density in space. Spatial clustering is a process of grouping a set of spatial objects into groups called clusters. Objects within a cluster show a high degree of similarity, whereas the clusters are as much dissimilar as possible.

Spatial Clustering is a data mining technique which discovers interesting and useful patterns in spatial databases by grouping the spatial objects into clusters.

Clustering is a very popular data mining technique for grouping data objects without any prior domain knowledge on data. In machine learning perspective, it is categorized as an unsupervised learning, in which the learning process does not need to know the possible output class. Clustering does not rely on predefined labels of classes or a priori given number of classes.

Clustering algorithms can be separated into four general categories [9]:

1. Partitioning method,
2. Hierarchical method,
3. Density-based method and
4. Grid-based method.

### *C. Visualization*

Geographic Information System (GIS) is a main tool for storing, processing and displaying spatial data. GIS is a good spatial analysis tool, however, with the accumulation of spatial data, the spatial analysis function that GIS offers is not enough. Spatial data mining, which can automatically discover implicit knowledge from spatial data, has recently received wide attention. Lots of spatial data mining methods and systems have been developed.

Visualization is defined as a graphical representation of data or concepts which is either an internal construct of the mind or an external artifact supporting decision making.

Data visualization is essential for understanding the concept of multidimensional spaces. It allows the user to explore the data in different ways at different levels of abstraction to find the right levels of details. Therefore techniques are most useful if they are highly interactive permit direct manipulation and include a rapid response time.

Visualization provides valuable assistance to the users by representing information visually. This assistance may be called cognitive support. Visualization can provide cognitive support through a number of mechanisms such as grouping related information for easy search and access, representing large volumes of data in a small space and imposing structure on data and tasks can reduce time complexity, allowing interactive exploration through manipulation of parameter values.

Visualization techniques could enhance the current knowledge and data discovery methods by increasing the user involvement in the interactive process.

Visual data mining is a step in the knowledge discovery process that utilizes visualization as a communication channel between the computer and the user to produce novel and interpretable patterns. [1].

Visual data mining can be viewed as an integration of data visualization and data mining. Considering visualization as a supporting technology in data mining, four possible approaches can be used

1. The usage of visualization technique to present the results that are obtained from mining the data in the database.

2. Applying the data mining technique to visualization by capturing essential semantics visually.

3. Use visualization techniques to complement the data mining techniques.

4. Use visualization technique to steer mining process.

Both scientific visualization and information visualization create graphical models and visual representations from data that support direct user interaction for interaction for exploring and acquiring insight in to useful information embedded in the underlying data

Visualization used in cluster analysis maps the high-dimensional data to a 2D or 3D space and aids users having an intuitive and easily understood graph/image to reveal the grouping relationship among the data.

Visualization Types

- Standard displays - histograms and pie charts
- Geometrically transformed - parallel coordinates plot m-dimensional space is mapped onto the 2- dimenisions by using m vertical axes.
- Iconic displays - Chernoff faces
- Dense pixel displays - circle segments
- Hierarchical displays - maps the hierarchical structure on a hemisphere[18]

Visual data mining uses visualization as a communication channel between the computer and the user, to produce novel and interpretable patterns. There are three classes of visual data mining;

- Visualization of data mining results. Extracted patterns are visualized to make them more interpretable.
- Visualization of the data mining process. The process of a mining algorithm can be visualized to help the discovery.
- Visualization of the data. Data is visualized before a mining algorithm is applied [7].

The idea to integrate spatial data mining and GIS, will produce mature and practical spatial data mining systems.

### *D. Integration of GIS, Spatial Data Mining and Visualization*

At present, Geographic Information System (GIS) is a main tool for storing, processing and displaying spatial data. GIS is a good spatial analysis tool, however, with the accumulation of spatial data, the spatial analysis function that GIS offers is not enough. Spatial data mining, which can automatically discover implicit knowledge from spatial data, has recently received wide attention. Lots of spatial data mining methods and systems have been developed.

The idea to integrate spatial data mining and GIS, will produce mature and practical spatial data mining systems. [6]

## III. METHODOLOGY

The approach and methods for this research are summarized into 3 categories:

a) Review of literature on spatial data mining in GIS applications.

b) ‘Desk Research’ methodology– defining the needs for this GIS application in terms of data or information available in spatial databases (deductive research) and

c) Inductive approaches – the implementation of prototype

### *A. Research Design*

The methodology used for this work is Cross Industry Standard Process for Data Mining (CRISP- DM)

The CRISP – DM is an iterative process that typically involves the following phases:

1. *Problem definition* - understanding of the business problem and defining objectives.

2. *Data exploration* - understand the meaning of the metadata. collect, describe, and explore the data. They also identify quality problems of the data.

3. *Data preparation* - build the data model by selecting tables, records, and attributes, are typical tasks in this phase. The meaning of the data is not changed.

4. *Modeling* - select and apply various mining functions

5. *Evaluation* - evaluate the model. If the model does not satisfy their changing its parameters until optimal values are achieved and decide how to use the data mining results.

6. *Deployment*

Use of data mining results and prototype in the domain.

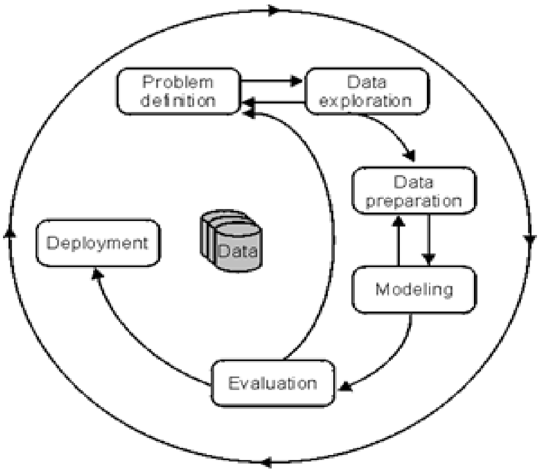


*Figure 1: The Crisp – DM Model*

### B. *Study Area*

The sample area of study used in this study was Kenya. Data was collected for Kenyan health facilities.

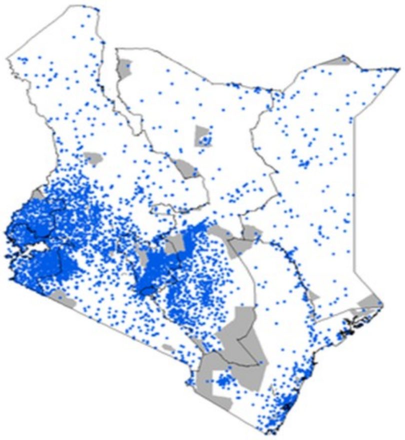


Figure 2: Area of Study – Kenya Map

Text heads organize the topics on a relational, hierarchical basis.

### C. *Data Sources, Software and Hardware*

TABLE I

| Item | Source | Remarks |
|---|---|---|
| **PostgreSQL V8.4** | http://www.postgresql.org/ | Open Source |
| **PostGIS V2.0** | http://www.postgresql.org/ | Open Source |
| **Quantum GIS Tethys** | http://www.qgis.org/ | Open Source |
| **Python V2** | http://www.python.org | Open Source |
| **Health Facilities Data** | DEPHA,ILRI,WHO,KEMRI,opendata.go.ke | No Data Fees |
| **Geoserver** | http://www.geoserver.org/ | Open Source |
| **Open Layers** | http://www.openlayers.org/ | Open Source |

Table 1 : Data Sources, Software and Hardware

### Data Description

The data collected for this project is mainly on Kenya Health facilities as shown in fig 2. It is a coverage showing health service providers for Kenya compiled by Kenya Medical Research Institute (KEMRI/Welcome Trust collaborative group). It shows the relative location of health service providers for Kenya categorized by type and supporting agency.

A number of methods were used to provide a longitude and latitude for each health service provider identified. These included the use of global positioning systems (GPS) by various NGO and research groups; extraction and triangulation of coordinates from hand drawn maps against GIS data on administrative boundaries and roads through a process of on-screen digitizing using Arcview GIS (Version 3.2); the use of 1:50,000 topographical maps; matching names of facilities to digital databases of village names and market centers created by the International Livestock Research Institute, Kenya; and finally matching facility names to fifth-level administrative boundary units where these units were small by extracting a centroid position.

This data was towards an initiative to developing a framework for equitable and effective resource allocation for health that is dependent on knowledge of service providers and their location in relation to the population they should serve.

The data acquired is appropriate for this project however, there needs to be a better mechanism to carry out analysis as well as provide users with only appropriate information, as presenting all these data to the user may overwhelm them and make analysis and interpretation of this data a complex task and an experience the user may not like much as they need to perform certain functions.

Properties of collected data include their formats which are as follows:

1. Attribute Data
2. Maps
3. Spatial Data
4. Shape files

Fields used to categorize the data were as follows:

- Location
- Facility Type
- Owner
- Services
- Operational Status
- County
- Services Offered

## IV. DESIGN AND IMPLEMENTATION

### *A. Proposed Prototype Design*

The architectural design for the automated visual clustering functionality is as shown below;

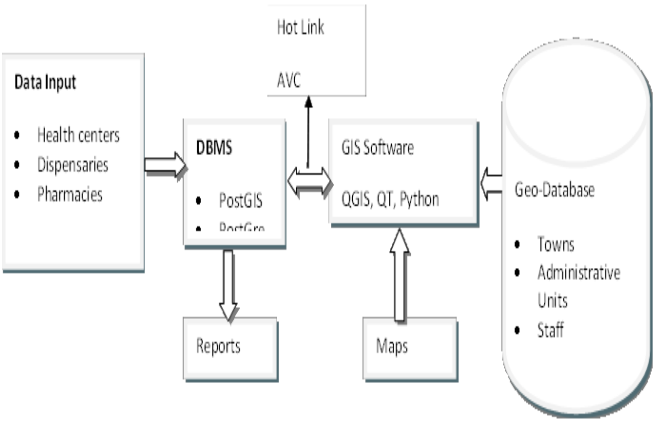


Figure 2 : Architectural Design

### ***Database Design***

Entity relationship Diagram

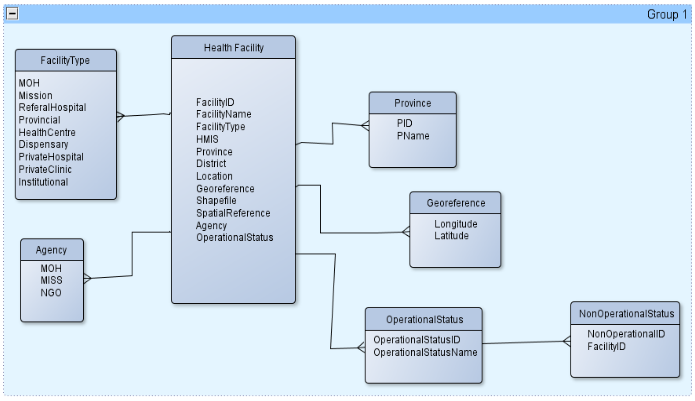


Figure 3: Entity relationship Diagram

### ***Sample Attribute information of the database:***

<u>Facility type</u>

F_NAME: Name of the institution

HMIS: Unknown

COUNT: Name of county in which facility is located

DIST: Name of district in which facility is located

DIVISION: Name of division in which facility is located

LOCATION: Name of location in which facility is located

SUBLOCATI: Name of sub-location in which facility is located

LONG: Longitude

LAT: Latitude

SPATIAL_REF: Method that was used to capture the position of the facility

F_TYPE: Code representing type of facility

1) Hospital MoH and Mission Districts, sub-districts

2) Referral Hospital and Provincial Hospitals

3) Health Centre

4) Dispensary

5) Private Hospital

6) Private Clinics and Medical centres

7) Nursing Homes and Maternity Hospitals

8) Special Treatment Hospitals

9) Institutions Health Facilities - schools, Universities, Employer, Police, Prisons, Other Ministries, Airport & Port Authorities, Armed forces

AGENCY: Agency running the facility

Layer Characteristics

| | |
|---|---|
| Projection: | Geographic |
| Theme: | Infrastructure |
| Layer type: | Arc View shape |
| Feature type: | Point |

### ***Proposed Graphical User Interface (GUI) Design***

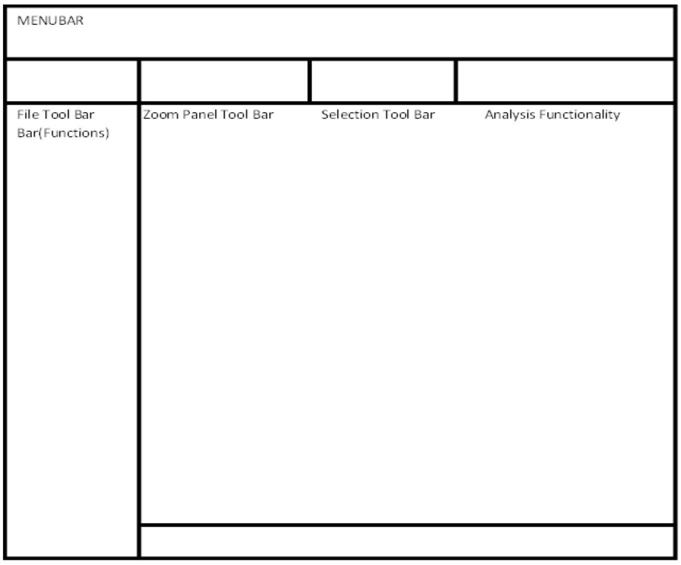
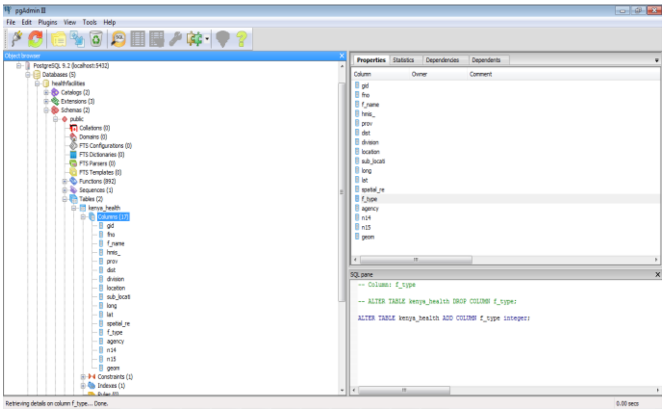


Figure 4 : Graphical User Interface Design

## V. RESULTS AND DISCUSSION

The objective of the study was to develop a prototype that extracts as much GIS data as possible and carry out automated visualized clustering.

The Database (health facilities) was developed using postgres as shown in the screenshot below:



| Column | Owner | Comment |
|---|---|---|
| gid | | |
| fno | | |
| f_name | | |
| hmis_ | | |
| prov | | |
| dist | | |
| division | | |
| location | | |
| sub_locati | | |
| long | | |
| lat | | |
| spatial_re | | |
| f_type | | |
| agency | | |
| n14 | | |
| n15 | | |
| geom | | |

```
-- Column: f_type

-- ALTER TABLE kenya_health DROP COLUMN f_type;

ALTER TABLE kenya_health ADD COLUMN f_type integer;
```

Retrieving details on column f_type... Done.
0.00 secs

Figure 6 : Kenyan Health Facilities

The objective of the study was to develop a prototype that extracts as much GIS data as possible and carry out automated visualized clustering. This was achieved using the hardware and software tools described in the methodology. This led to development of the Automated Visual Clustering Functionality Prototype. Below are screenshots of the prototype that demonstrate the end product of this research which incorporates automated clustering, visualization and improved and efficient mining of GIS data. This meets the objective of this research as is shown in the figures of screenshots;

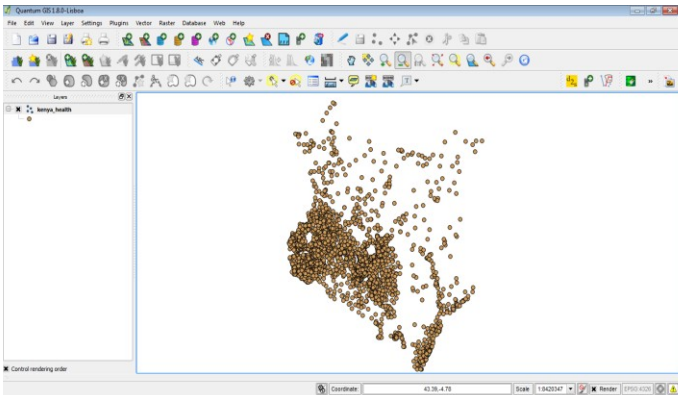


Figure 7 : Kenyan Health Facilities

<u>Exclusive reports:</u>

Examples Include:

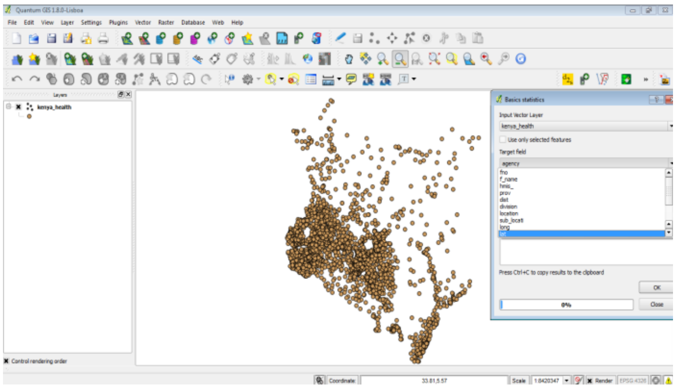
Summary by county, health facility type, service, ownership/agency, operational status

My work includes a search engine to enable optimal use of the data set where users can search the data using very specific fields:

Facility Name,

Location : county, constituency, location,

The objective of the study was to extract as much data as possible, mine it and visualize it.



Basics statistics

Input Vector Layer: kenya_health

□ Use only selected features

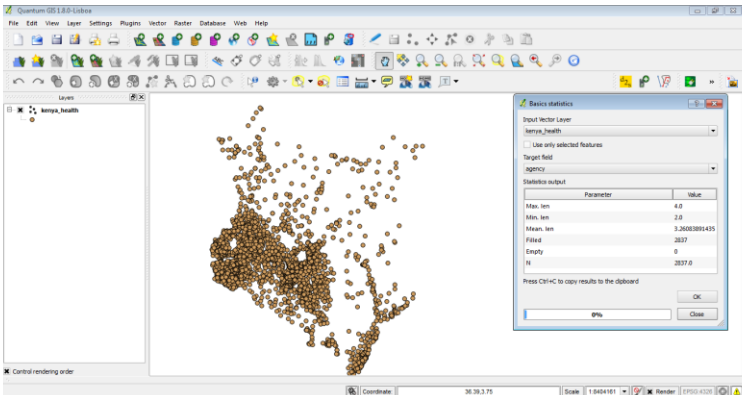
Target field: agency

fno
f_name
hmis_
prov
dist
division
location
sub_locati
long
lat

Press Ctrl+C to copy results to the clipboard

OK
0%
Close

Figure 8 : Basic Statistics Querying Capability



Basics statistics

Input Vector Layer: kenya_health

□ Use only selected features

Target field: agency

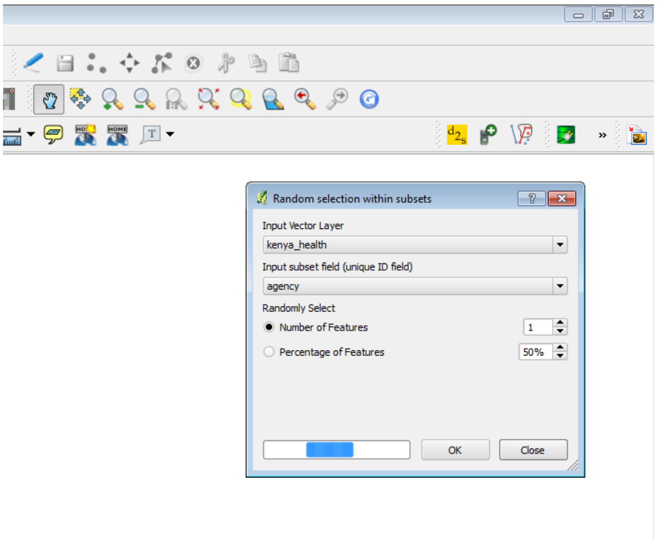
Statistics output

| Parameter | Value |
|---|---|
| Max. len | 4.0 |
| Min. len | 2.0 |
| Mean. len | 3.26083891435 |
| Filled | 2837 |
| Empty | 0 |
| N | 2837.0 |

Press Ctrl+C to copy results to the clipboard

OK
0%
Close

Figure 9 : Results of Basic Statistics



Random selection within subsets

Input Vector Layer: kenya_health

Input subset field (unique ID field): agency

Randomly Select

● Number of Features 1

○ Percentage of Features 50%

OK
Close

Figure 10 : Random Selection within subsets

## REFERENCES

[1] Ankerst, M. 2000. Visual Data Mining. PhD Thesis, Ludwig-Maximilians-Universität, München, Germany.

[2] Anselin L, Syabri I, Kho Y (2009), GeoDa: an introduction to spatial data analysis. In Fischer MM, Getis A (eds) Handbook of applied spatial analysis. Springer, Berlin, Heidelberg and New York, pp.73-89

[3] Diansheng, G. (2002), ,"Spatial Cluster Ordering and Encoding for High-Dimensional Geographic Knowledge Discovery", UCGIS2, Summer, 2002

[4] ESRI, (2000). Challenges for GIS in Emergency Preparedness and Response. Retrieved Feb. 23, 2014 from: http://www.esri.com/library/whitepapers/pdfs/challenges.pdf

[5] ESRI inc., Environmental systems research institute [Online] cited August 28 2013. Available http://www.esri.com/what-is-gis/index.html

[6] Gahegan, M (2001), Gahegan, M., M. Harrover, T. M. Rhyne and M. Wachowicz (2001).

[7] 'The integration of Geographic Visualization with Databases, Data mining, Knowledge Discovery Construction and Geocomputation', Cartography and Geographic Information Science, 28, 29–44.

[8] Gupta, G. K., J. Ghosh. 2001, Detecting seasonal trends and cluster motion visualization for very high dimensional transactional data. Proc. First Siam Conf. On Data Mining, (SDM2001). 115–129

[9] Gething PW, Noor AM, Goodman CA, Gikandi P, Hay SI, Sharif SK, Atkinson P, Snow RW: Information for decision making from imperfect national data: tracking major changes in health care use in Kenya using geostatistics.

[10] (Hand et al 2001), Han J., Kamber M. 2001, Data Mining. Concepts and Techniques. Morgan Kaufmann .

[11] Hsu, J. 2003. Critical and Future Trends in Data Mining: A Review of Key Data Mining Technologies/Applications, in: Wang, J. (ed.) Data Mining, Opportunities and Challenges, Idea Group Inc., Palo Alto, California, USA.

[12] Karimi, H. A., & Blais, J. A. R. (1997). Current and future direction in GISs. Computer, Environment and Urban Systems, 20(2), 85–97.

[13] Lo, C. P., and Yeung, A. K. W. (2002). Concepts and Techniques of Geographic Information Systems, Prentice Hall, New Jersey.

[14] Longley et al. (2005), Longley P., Goodchild M., Maguire D., Rhind D., (2005). Geographical Information Systems and Science. 2nd edition, Willey, England.

[15] Miller and Han, 2001 Miller H. J. and Han J., Geographic data mining and knowledge discovery, An overview, in Geographic data mining and knowledge discovery, Miller H. J. and Han J., Taylor & Francis

[16] Montoya L., (2003). Geo-data acquisition through mobile GIS and digital video: an urban disaster management perspective. Environmental Modeling & Software 18, 869–876.

[17] Noor AM, Gikandi PW, Hay SI, Muga RO, Snow RW: Creating spatially defined databases for health service planning in resource poor countries: The example of Kenya.

[18] Peuquet, D., (1999) . Davis, J. R., & Cuddy, S., (1999), Geographic information systems and environmental modelling. In A. J. Jakeman, M. B. Beck, & M. J. McAleer, Modelling change in environmental systems (pp. 543–556). New York, USA: John Wiley & Sons.

[19] Shekhar, S. , (2002). Shekhar, S., Lu, C. T., Zhang, P. and Liu, R. (2002) "Data mining for selective visualization of large spatial datasets," Proceedings of 14th IEE International Conference on Tools with Artificial Intelligence (ICTAI '02), IEEE Press.