2012

The application of a visual data mining framework to determine soil, climate and land-use relationships

Yunous Vagh
Edith Cowan University

10.1016/j.proeng.2012.01.1271

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 32 (2012) 299 – 306

Procedia Engineering

www.elsevier.com/locate/procedia

I-SEEC2011

The application of a visual data mining framework to determine soil, climate and land use relationships

Y. Vagh*

[a]*School of Computing and Security Science, Faculty of Computing, Health and Science, Edith Cowan University, Perth, Western Australia*

Elsevier use only: Received 30 September 2011; Revised 10 November 2011; Accepted 25 November 2011.

Abstract

In this research study, the methodology of action research dynamics and a case study was employed in constructing a visual data mining framework for the processing and analysis of geographic land-use data in an agricultural context. The geographic data was made up of a digital elevation model (DEM), soil and land use profiles that were juxtaposed with previously captured climatic data from fixed weather stations in Australia. In this pilot study, monthly rainfall profiles for a selected study area were used to identify areas of soil variability. The rainfall was sampled for the beginning (April) of the rainy season for the known 'drought' year 2002 for the South West of Western Australia. The components of the processing framework were a set of software tools such as ArcGis, QuantumGIS and the Microsoft Access database as part of the pre-processing layer. In addition, the GRASS software package was used for producing the map overlays. Evaluation was carried out using techniques of visual data mining to detect the patterns of soil types found for the cropping land use. This was supported by analysis using WEKA and Microsoft Excel for validation. The results suggest that agriculture in these areas of high soil variability need to be managed differently to the more consistent cropping areas. Although this processing framework was used to analyse soil and rainfall climate data pertaining to agriculture in Western Australia; it is easily applicable to other datasets of a similar attribution in different areas.

Keywords: Visual Data Mining; Processing Framework; ARCgis Grass; QuantumGIS; GIS; Land use; Soil Variability

1. Introduction

In order to understand what a framework is, it is necessary to understand that system data do not exist in isolation but are related to other data by sharing common features [1]. Although the data from different

* Corresponding author. *E-mail address*: vvagh@ecu.edu.au.

doi:10.1016/j.proeng.2012.01.1271

systems may have common features, they appear to be outwardly unrelated, or related in uncharacteristic and undescribed ways [2]. Consequently, a framework is a well-structured and refinable specification that permits the identification and understanding of common properties for the purpose of creating models from common abstractions [3]. Geographic Information Systems (GIS) technology which enables the capture, storage, analysis and display of geo-reference information [16], can be easily integrated into various frameworks. A quality GIS framework such as this which utilises software for a land information system is dependent upon many factors, the most important of which is the input data, followed by the set of data processing functionality [4], especially when decisions are based upon information provided by them.

2. Related Work

Frameworks have been used in a number of different research areas including geographic measurement [5], agricultural policy option comparisons [6] and visual data mining [7]. There have been a number of GIS framework applications to land-use data such as the Army Remote Moisture System (ARMS) that link a land performance modelling system (Land Information System LIS) that provides soil moisture estimates [14] and the application of a hierarchical Bayesian approach to land use data for testing improvements in mapping historical vegetation mapping [15]. There have been no studies done utilizing agricultural, geographic and climate data in the agricultural growing region of Western Australia to determine any relationships between soil, climate and land use. This study specifically used the production of visual geodetic maps for the soil and land use profiles of a selected agricultural region in the South West of Australia. The aim of this paper was to develop a data mining framework suitable for the processing of agricultural geo-spatial data by using DEM, climate, soil and land use data profiles of the South West agricultural region of Western Australia. These generic qualities are related to the fact that GIS data occurs in the form of two digital formats namely, raster and vector data. Raster data are basically numerically coded grid cells or pixel data, while vector data are comprised of coded points, lines or polygons [8]. Although modelling the data statistically, performing clustering and finding association rules are some of the ways to approach data mining problems, it is important to find inter-relationships between data entities involving location when dealing with data that have geographic attributes [9]. This paper deals with these geographic inter-relationships in the context of agricultural land-use. It follows the three-step process typically employed by visual data exploration of overview first, zoom and filter and details on demand or what has sometimes been referred to as Shneiderman's visual information seeking mantra [10]. Furthermore, the study is an exercise in both displaying a solution visually as well as visually finding a solution as espoused by [11]. The focus was on the three main aims of visualisation namely; presentation, confirmatory analysis and exploratory analysis [12]. In this way, the information is visually represented, allowing for direct interaction whereby insights, conclusions and decisions could be gained, drawn and made respectively [13].This work began with the intention of finding a relationship between rainfall, land use and soil substrate and to thereby demonstrate that there was some justification in utilising land for agriculture given the climatic conditions and to also determine future land use depending on the soil type.

3. Materials and methods

The data used in this study had different attributions and were made up of five separate but related entities. All of the datasets were in relation to the South West region of Western Australia. The separate datasets were made up of climate, soils, land use and DEM features. The data was extracted at the Department of Agriculture and Food of Western Australia (DAFWA) in Kensington, Perth. The datasets

were extracted using ArcGIS from where the data was first projected into the UTM GDA94 zone 50 reference system in centimeters. Furthermore, the study area was a rectangular grid of 104328 cells of 1000m each stretching from Busselton in the west to Esperance in the east. The ERMapper software was used to create the study area; then used to save the DEM extent as a raster (.ers) file. In addition, all the datasets were fitted specifically to the extraction region of the selected study area. The research methodology was developed after experimentation with a number of software tools that represented an admixture of extraction, pre-processing, analysis, data mining and visualization of climate data sourced from DAFWA. The whole process is depicted in Figure 1.

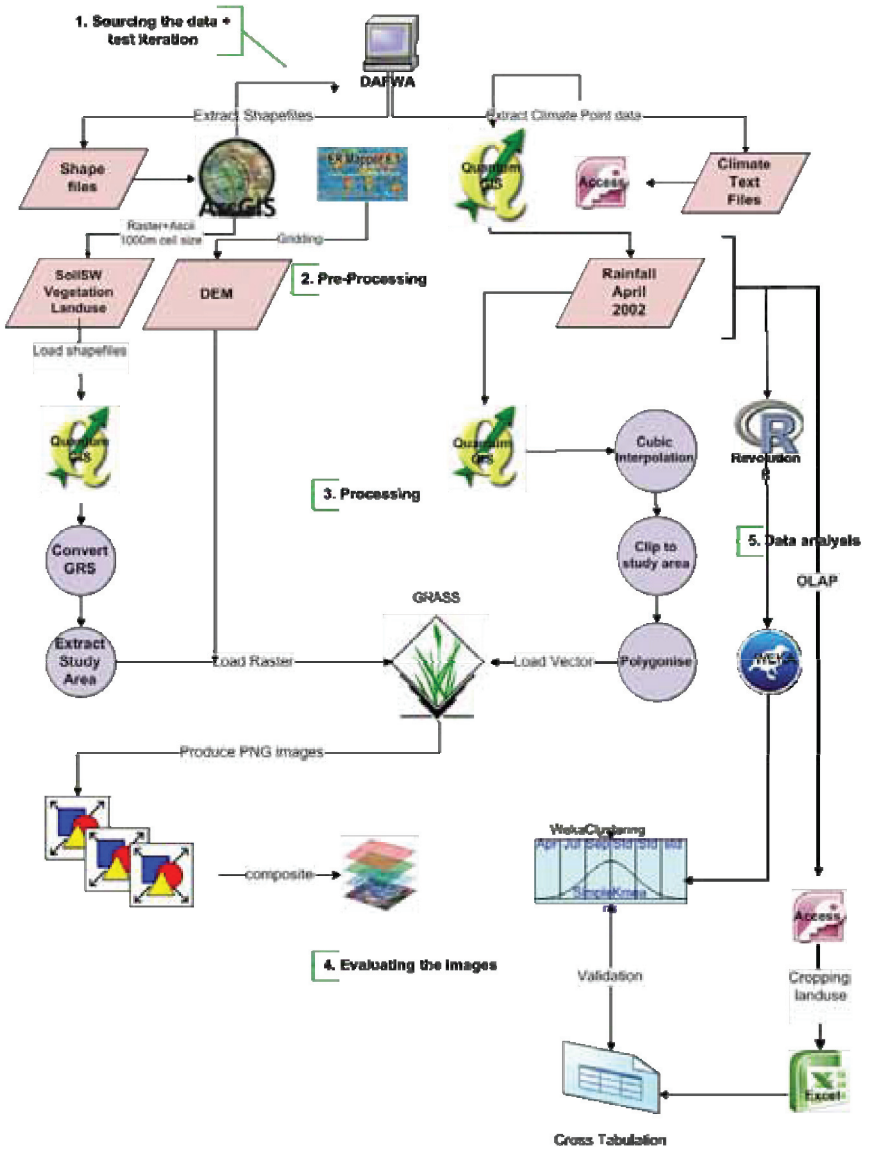

Fig. 1. Process methodology of the visual data mining of climate and geographical data

Each of the phases denoted in Fig. 1 consisted of a number of steps. The first step (stage 1) was to source the data from DAFWA. This was achieved by extracting shape files of the GIS data for the whole South West agricultural growing region and then doing a subset for the selected study area. A detailed description of the datasets has been provided in Table 1.

4. Processing

Pre-processing was the second stage (2). In this stage, the study area (extent) was loaded into QGIS and then re-projected into the UTM geographic reference system (grs) resulting in coordinates with eastings and northings. The DEM, soil and land use layers were also similarly re-projected and then intersected with the study area for relevance. The rainfall data was imported into QGIS as delimited text and then cubically interpolated to cover the study area in grid cell sizes of 1000m. The next step (stage 3) involved the processing of the data layers within the GRASS environment.

The next series of steps for the third stage (3) was performed in the GRASS software package in order to produce an output that would have some visual semantics and from which a discussion could ensue and subsequent conclusions could be drawn. The first part of the process was to load the raster and vector datasets. Apart from the rainfall data which was in vector format, all of the other datasets were in raster format. Table 1 illustrates the exact nature of the different datasets. In this pilot study only the rainfall for April 2002 was overlaid to determine the correlations.

Table 1. The components and structure of the composite map for Fig. 2

Layer No	Dataset type	Resolution	Feature	Map Opacity	Profile origin	Spatial Type
1	DEM	1000m	elevation	80	ERMapper	raster
2	Soils	1000m	mapping unit	80	ArcMap	raster
3	Land use	1000m	tertiary land use	80	ERMapper	raster
4	Rainfall Apr 2002	1000m	average monthly rainfall	25	QuantumGIS	vector polygons

The first dataset that was loaded in the GRASS workspace was the DEM for the elevation characteristic of the study area. This was overlaid with the rainfall data for April 2002. Rainfall for April was for the start of the 'rainy' season for the 'drought' year 2002.

The second dataset loaded into the display area was the soil type followed by the average monthly rainfall for April 2002. The predominant soil type over the agricultural region was light blue with some patches of light green and yellow as depicted in Fig. 2.

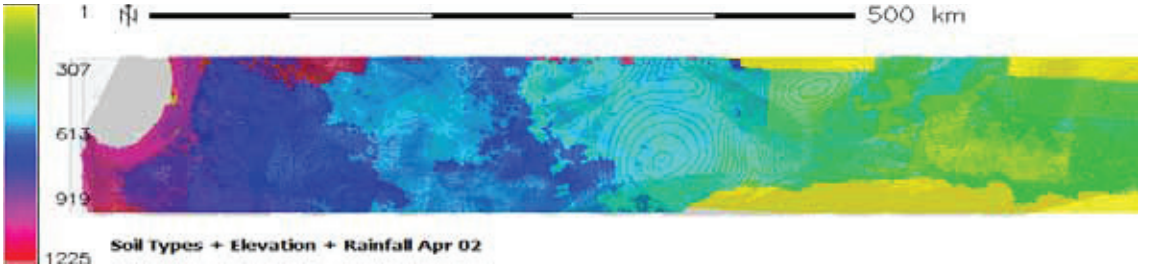

Fig. 2. The visual correlation of the underlying soil types and the average monthly rainfall for April 2002

The third dataset examined was the land use mapping which was again overlaid by the average monthly rainfall for April 2002. The predominant land use across the agricultural growing region was for cropping (340). The other feature values between 341 to 360, represented other farming activities such as cereals, grazing, horticulture, fruit tree plantations and natural parks and reserves.

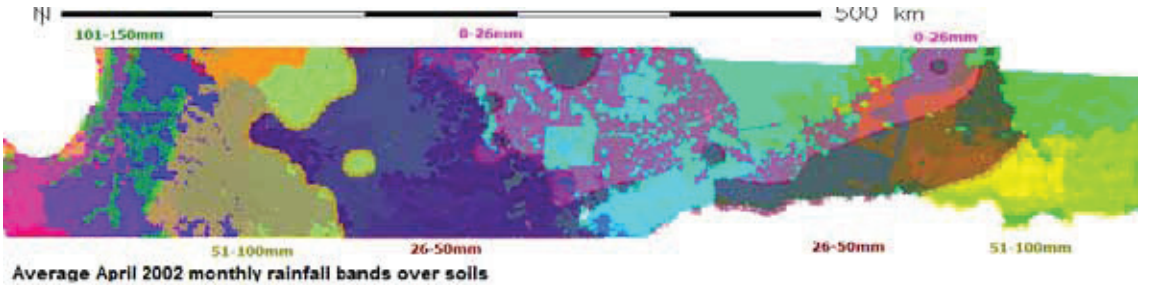

Fig. 3. The visual correlation of the underlying soil types and the bands of average monthly rainfall for April 2002

The map in Fig. 3 shows the distribution of the different rainfall bands over the cropping areas. Generally the coastal areas receive higher rainfall as represented by the green (101-150mm) and brown (51-100mm) patches, whilst the majority of the agricultural area receives rainfall in the 2 bands of 0-25mm and 26-50mm rainfall as represented by the pink and grey patches. Each of the separate rainfall bands featured a predominant soil type associated with the amount of rainfall received. This breakdown of soil types to rainfall band is shown in Table 3.

Table 3. Visual analysis of the composite maps

Rainfall band	Predominant soil type
0-25mm	Grey deep sandy duplex, Yellow/brown deep sandy duplex & Duplex sandy gravel
26-50mm	Grey deep sandy duplex
51-100mm	Grey deep sandy duplex
101-150mm	Brown loamy earth, Brown deep loamy duplex & Friable red/brown loamy earth
151-250mm	Wet soil & Semi-wet soil

5. Experimental results and analysis

The underlying raster image layers formed the dependant variables for the attributes of soil type and the rainfall data was the independent variable. The map image displayed in Fig. 2 visually denoted the visible correlation between the rainfall and soil type. Rainfall during the start of the 'rainy' season in the south-west region clearly falls at a greater concentration along the coast and then tapers off as it progresses inland. There were some large pockets of high rainfall which matched well with the underlying soil type layers (green and light green colouring) that corresponded to the actual agricultural growing region in Fig. 2. The results were collated into two tables for the analysis of stage 5 of Fig. 1. Table 4 represents the observations and visual analysis from the GRASS image of Fig. 2, whereas Table 5 represents the Weka cluster analysis of the average rainfall for the month of April 2002.

Table 4. Visual analysis of the composite maps

Map	Predominant Colour	Salient Feature ex Weka
DEM	Light blue	Height of 100-200 metres
Soil type	Light blue, light green, yellow	Loamy, deep sandy, wet
Land use	Blue, light blue, green	Cropping, cereals, hay & silage, seasonal horticulture, irrigated tree fruits

The rainfall figures obtained from the cubic interpolation and prediction done in the Revolution R statistical package, were validated by cluster analysis done in the Weka software package. The cluster method used was simple K-means clustering with a selection of 10 clusters. The results were captured in Table 5 for the rainfall for the month of April 2002. The clusters 1, 5 and 8 recorded the highest rainfall figures for April 2002 and these were concentrated on the coastal region to the west of the study region. The centroids for the moderate and high rainfall areas were all located in the agricultural growing region in the middle to right of the study area in Fig. 2.

Table 5. Weka simpleKmeans cluster result of the April 2002 rainfall

Weka Cluster No	Centroid Northings	Centroid Northings	No of instances in cluster	Percentage of total instances in cluster	Actual April 02 Rainfall (centroid)
0	999397.68	6259963.22	10238	10	56.44
1	521305.50	6281938.90	8072	8	139.17
2	882156.79	6275105.18	11737	11	21.67
3	384458.62	6268890.72	9394	9	44.26
4	495964.35	6330428.90	11150	11	32.19
5	682647.66	6259607.85	11044	11	89.68
6	553535.22	6235644.05	11539	11	27.50
7	923405.75	6320754.49	11913	11	27.32
8	661851.52	6338114.87	8838	8	64.03
9	682961.36	6303379.62	10403	10	51.97

6. Discussion

The numbers 1 to 27 represented the 27 different soil types in the agricultural cropping region after the reduction process from the original 682 soil types. The graph in Fig. 5 showed that most of the rainfall falling in the agricultural cropping region was mainly within the first three bands of 0-25mm (9745 instances), 26-50mm (13695 instances) and 51-100mm (7398 instances) for the average monthly rainfall of April 2002. These were the blue, red and green bars in Fig. 4 respectively.

The soil types which corresponded to the highest rainfall concentrations of over 600 instances in the agricultural cropping region are shown in Table 6. It is evident from the table that all soil types except numbers 4 and 6 are sandy soils. In addition, the graph in Fig. 4 indicates that different soil compositions pertain to each rainfall band. For example, the dominant soil types which receive the lowest rainfall of between 0-25mm are the numbers 2, 3, 5, 9, 12 and 15 and, the dominant soil types receiving rainfall between 26-50mm are the numbers 1, 2, 3, 4, 7, 8 10, 11 and 14. The soil types receiving the high rainfall averages are the numbers 1, 2, 4 and 6. The combination of the three bands therefore shows a predominance of the first 5 soil type numbers. The results also suggest that higher than 100mm of average monthly rainfall does occur within the agricultural cropping region for some of the 27 soil types such as

soil type number 6, 18 and 22. However, these instances only represent 3.2% of the total for the other three selected rainfall bands. Nevertheless, these areas with high rainfall variability may indicate that different soil management techniques may be needed for them due to the rainfall received.

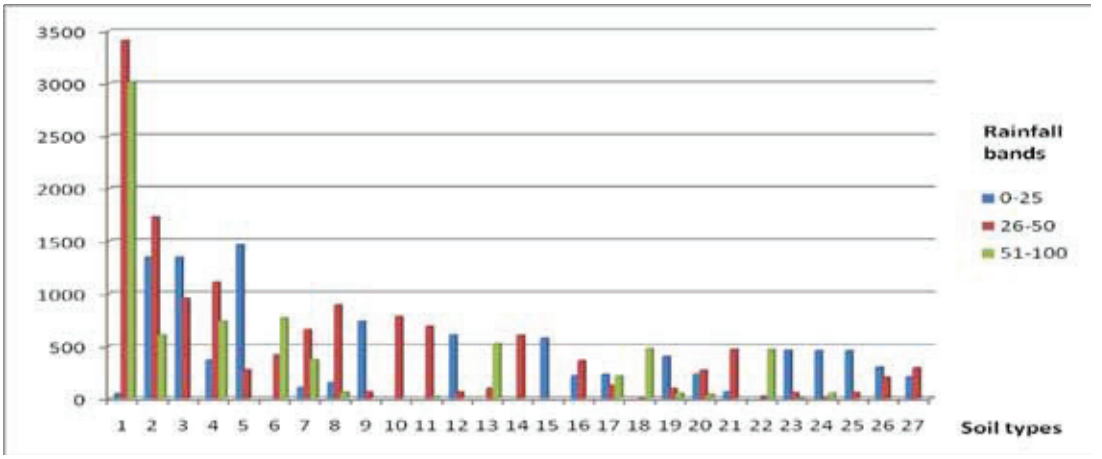

Fig. 4. The graph of average monthly rainfall (April 2002) versus soil types

In addition, for the purposes of validation and support, the data mining (DM) technique of classification using the J48 algorithm was performed on the dataset containing only the 27 soil types of interest, in order to determine a relationship between the results obtained from the two methods. Although the accuracy of the classification was only 18.86%, an examination of the area under the ROC curve showed some correlation between the two results, for example the blue coloured rows in Table 6.

Table 6. The soil type descriptions with the highest rainfall instances from the cross tabulation and WEKA classification

No	Soil complex description	Rainfall Instances	% of Total	ROC Area Ex WEKA
1	Grey deep sandy duplex	6468	20.97	0.70
2	Alkaline grey shallow sandy duplex	3685	11.95	0.70
4	Saline wet soil	2205	7.15	0.62
6	Pale deep sand	1179	3.82	0.75
7	Grey shallow sandy duplex	1132	3.67	0.69
13	Grey deep sandy duplex, Bare rock & Duplex sandy gravel	620	2.01	0.86
18	Duplex sandy gravel	483	1.57	0.84
22	Duplex sandy gravel, Loamy gravel & Deep sandy gravel	494	1.60	0.84

7. Conclusion

The use of a case-study of geographical data of soil profiles combined with average monthly rainfall data for the month of April and the year 2002, was a demonstration of the complexity and challenges of working with diverse and multi-source datasets that effectively covered the various dimensions of raster, vector and text data. The results do indicate a relationship between the soil type and the average monthly

rainfall as was evident from Figure 4. However, this relationship needs to be investigated further using non-categorical data. Nevertheless, this pilot study helped to uncover a foundation methodology comprised of the different levels of software, data and analytics that was useful in devising a framework for the analysis of complex data with a geospatial dimension. The validation methods used in this study involved the production of a portable network graphics image that served as a composite picture for visual inspection and analysis. In addition, the components of human intuition and background knowledge were used as part of the visual analytics to both find solutions and to display them graphically. The results indicate particular soil types associated with consistent rainfall. Farmers, agricultural experts and consultants may find these initial results a bedrock upon which to develop further strategies for soil management in agricultural areas. Furthermore, they could make use of this method to obtain an exact soil composition of their own growing regions given the geographic coordinates or shire name. This study may be extended to firstly examine the effect of rainfall on vegetation and secondly, to expanding the analysis to include time-series rainfall data covering different years and months as well as increasing the attribution to cover temperature in order to uncover or confirm further relationships.

Acknowledgements

The author wishes to especially acknowledge Mr. Phil Goulding from the DAFWA organisation for his invaluable assistance in providing the data as well as his advice and suggestions on sampling the data.

References

[1] D. Garlin and D. Notkin. Formalising design spaces: Implicit innovation mechanisms. *Presented at the Formal Software Development Methods:* Int Symposium VDM Europe*,* 1991.
[2] M. Luck and M. D'Inverno. A Conceptual Framework for Agent Definition and Development. *The Computer Journal*, 2001;**44**:1-20.
[3] M. D'Inverno, et al. A formal framework for specifying design methodologies. *Software Process Improvement Pract.,* 1996;**2:**181-195.
[4] R. F. Tomlinson, *Thinking about GIS*. Redlands. California: Ingram Publishers, 2007.
[5] J. Miller and J. Han. *Geographic data mining and knowledge discovery.* 2nd ed. New York: Taylor & Francis Inc; 2009.
[6] M. K. van Ittersum, et al. Integrated assessment of agricultural systems – A component-based framework for the European Union (SEAMLESS). *Agricultural Systems*. 2008;**96**:150-165.
[7] H. J. Schulz, et al. A Framework for Visual Data Mining of Structures. *Presented at the Twenty-Ninth Australasian Computer Science Conference (ACSC2006)*, Hobart, Tasmania : Australia, 2006.
[8] R. G. Congalton. Exploring and Evaluating the Consequences of Vector-to-Raster and Raster-toVector Conversion. *American Society for Photogrammetry and Remote Sensing.*
[1] Van der Geer J, Hanraads JAJ, Lupton RA. The art of writing a scientific article. *J Sci Commun* 2000;**163**:51–9.
[10] B. Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. *Proceedings of the IEEE Symposium on Visual Languages.* 1996:336–343.
[11] B. Kovalerchuk. *Decision Process and its Visual Aspects*. Visual and spatial analysis: advances in data mining. K. B and J. Schwing, Eds., ed Netherlands: Springer, 2004.
[12] R. E. Garcia, et al. Visual Analysis of Data from Empirical Studies. *Presented at the International Workshop on Mining Software Repositories*, 2004.
[13] D. A. Keim, et al. Visual Analytics: Scope and Challenges. *Visual data mining: theory, techniques and tools for visual analytics* S. J. Simoff, et al., Eds., ed Berlin, Heidelberg: Springer-Verlag. 2008:76-90.
[14] M. Tischler, et al. A GIS framework for surface-layer soil moisture estimation combining satellite radar measurements and land surface modeling with soil physical property estimation. *Environmental Modelling and Software*. 2007;**22** 891-898.
[15] H. S. De, et al. Mapping pre-European settlement vegetation at fine resolutions using a hierarchical Bayesian model and GIS. *Plant Ecology.*2007;**1**:85-94.
[16] T.P. Avasthi. An Introduction to GIS. *The Third Pole.* 2009;**5-7**:76-78.