Taylor & Francis
Taylor & Francis Group

Human and Ecological Risk Assessment: An International Journal

ISSN: 1080-7039 (Print) 1549-7860 (Online) Journal homepage: http://www.tandfonline.com/loi/bher20

Spatial data analysis with R programming for environment

Efdal Kaya, Muge Agca, Fatih Adiguzel & Mehmet Cetin

To cite this article: Efdal Kaya, Muge Agca, Fatih Adiguzel & Mehmet Cetin (2018): Spatial data analysis with R programming for environment, Human and Ecological Risk Assessment: An International Journal, DOI: 10.1080/10807039.2018.1470896

To link to this article: https://doi.org/10.1080/10807039.2018.1470896

Published online: 22 May 2018.

HUMAN AND ECOLOGICAL RISK ASSESSMENT
https://doi.org/10.1080/10807039.2018.1470896

Taylor & Francis
Taylor & Francis Group

Check for updates

Spatial data analysis with R programming for environment

Efdal Kaya[a], Muge Agca[b], Fatih Adiguzel[c], and Mehmet Cetin [d]

[a]Department of Geomatics Engineering, Natural and Applied Science, Kocaeli University, Kocaeli, Turkey; [b]Department of Geomatics Engineering, Izmir, Katip Celebi University, Izmir, Turkey; [c]Nevsehir Haci Bektas Veli University, Urgup Sebahat and Erol Toksoz Vocational School, Programs of Cultural Heritage and Tourism, Nevsehir, Nevsehir, Turkey; [d]Department of Landscape Architecture, Faculty of Engineering and Architecture, Kastamonu University, Kastamonu, Turkey

ABSTRACT

The use of open source software, which has been constantly evolving since the mid-2000s, has affected every research discipline. Disciplines using geographic information systems (GIS) and remote sensing (RS) data have been heavily affected owing to this evolution of technology. Researchers working on these data sets have begun to use open source software intensively. The analysis and visualization of spatial data with the help of open source software has caused the emergence of new different features, which are cost effective and editable by other users. In this study, eight sample points have been used for the analysis of water quality in the Mamasın dam in the 2209/A group project of "Assessment and Modeling with GIS and RS Data of the Land Use Effects on Water Quality of Mamasın Dam" supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under its program to support graduate students. While visualizing spatial features of the points, QGIS Desktop 2.18.0 and Studio programs with open source code have been used. The RStudio program is an open source software that allows the use of the functions of the R programming language. This study is an ideal application for spatial analysis studies with the R programming language. The sample points used in the study were analyzed in the laboratories of Department of Environmental Engineering, Aksaray University. Spatial properties of the analyzed data were examined by coding in the Studio program that is free open source software. In the analysis process, first, the libraries, Leaflet(), Leaflet.extras(), rgdal(), sp(), raster(), and magrittr(), which are used in the study, have been uploaded. With the help of these libraries, the locations of the sample points are transferred to the OpenStreetMap using latitudes and longitudes of the geographic coordinate system as base map. The pH, conductivity, PO_4-P, PO_4, dissolved oxygen, and temperature information of each sample points are assigned to the variables. These variables are added as a feature for each point. The spatial characteristics of the sample points are visualized using the data variable packages and online maps as the base. After the visualization process is completed, the generated map is presented on the website created via Github.

ARTICLE HISTORY
Received 14 April 2018
Revised manuscript accepted 26 April 2018

KEYWORDS
spatial data visualization; QGIS; RStudio; R programming; open source software

CONTACT Mehmet Cetin mcetin@kastamonu.edu.tr Department of Landscape Architecture, Faculty of Engineering and Architecture, Kastamonu University, Kastamonu 37150, Turkey.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/bher.

Introduction

Open source software is a software released openly or openly distributed software and is available to anyone who wants to write source code. The distinguishing feature of such software is that the user has freedom to change it. Open source software is adaptable, robust, fast, and reliable. The world of open source code presents a new type of software production and new business models. Open source code software developed in industry standards with the collective work of global information experts is available to everyone.

Open source software, which can be easily used by all institutions and individuals in applications such as embedded systems in mobile phones, clocks supercomputers and by home users, SMEs, public institutions, and all types of schools, is a software that can be simplified, improved, and strengthened according to user-specific requirements. Open source software offers more options to users than other conventional software.

Because of these developments in the software world, open (free) source software that provides a new perspective to information technology has brought a new perspective to the studies done by researchers either from public institutions or from scientific backgrounds using GIS (Onyıl and Türk 2016). Particularly in public institutions, open source software has been used to save license fees of software (Güneş 2007). Owing to the cost requirement of licensed software, open and free source software is used intensively in scientific studies.

A literature review reveals that there are many studies that are conducted on data visualization using licensed and unlicensed software. Studies of modeling and three-dimensional (3-D) visualization of the wind turbines fields in Spain have been performed in ArcView software (Chias and Abad 2013). Oceanographic data visualized integrated GIS and scientific visualization for environment (Yafang and Yongwei 1999). 3-D visualization was performed using ArcView and Vis5D software, a scientific visualization program, in a study area in the Monterey Bay by the central California coast (Su and Sheng 1999). Using the data sets in Sweden's Gövle city, 3-D visualization was performed according to the characteristics of special places in the GIS environment (Yao and Jiang 2005). The surface of the terrain has been visualized using the wavelet pyramid method that involves compressing terrain data for online 3-D surface model visualization (Olanda *et al.* 2014). A spatial analysis of the wind potential for Malaysia was conducted (Ibrahim *et al.* 2015). In epidemiology, analyses have been performed using GIS and spatial data (Rytkönen 2004). Spatial data are combined using overlay, which is one of the most important analysis of GIS (Bimonte *et al.* 2010). Visualization of important large-scale data that was lost has been performed (Popov 2006). In large geographic data, the volume of dynamic intensity is visualized on the basis of interactive visual cluster detection (Du *et al.* 2016). The growth of trees in the city is visualized interactively (Brasch *et al.* 2009).

In this study, spatial data is visualized using RStudio, an open source software, programmed in the R programming language. R is a programming language, which is used in statistical analysis and graphic production stage, and can be obtained free of charge on Internet. R software is based on the "S" language developed by Becker and Chembers. R software is an advanced version of S-PLUS, a paid software that has been introduced earlier. (Er and Sönmez 2005; İlk 2011; Doğan and Uluman 2016). Even if the R programming language was used for statistical calculations when it first appeared, packages, which help to solve many problems, have been developed by programmers. These packages are offered for free to the users. Several libraries have been written for spatial data analysis.

Field of study

Aksaray is on the Edirne, Istanbul, Ankara, Adana, Iskenderun highway and on the Samsun, Kayseri, Konya, Antalya highway. It lies between 33°E and 35°E and 38°N and 39°N. It is surrounded by Nevşehir in the east, Niğde in the south east, Konya in the west, Ankara in the north, and Kırşehir in the north east. The surface area is 7626 km^2. The area under study is the Mamasın Dam in Aksaray province (Figure 1). Mamasın Dam, which is a rock fill type on the Uluırmak Stream located 12 km east of Aksaray, is used for the irrigation. The dam was completed in 5 years and was built at the cost of 14,477,000 Turkish Liras. The dam body volume is 400,000 m^3; it is 44.90 m high; the lake volume at normal water elevation is 165.80 hm^3, and the lake size at normal water level is 16.20 km^2. Approximately 24,854 ha of area is irrigated with the dam of Mamasın. Besides, Mamasın Dam is the first irrigation area of application of Turkey and was launched as an irrigation project in 1962. An average of 45% of the potable water of the city is provided by this dam (Url-1).

Application

To use the RStudio program in the application phase, the R programming language's own interface is downloaded according to the type of operating system (Url-2). Then, RStudio is downloaded from its Internet address (Url-3). Once the programs are installed, RStudio program becomes suitable for analysis by writing scripts in the library window in the interface (Figure 2).

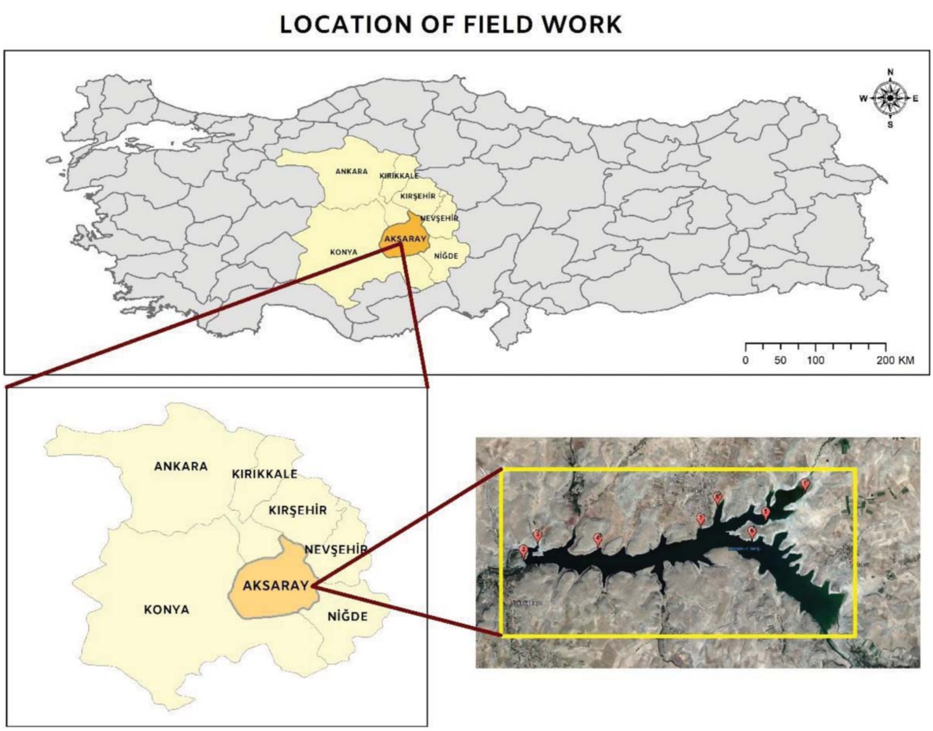

Figure 1. Field of study.

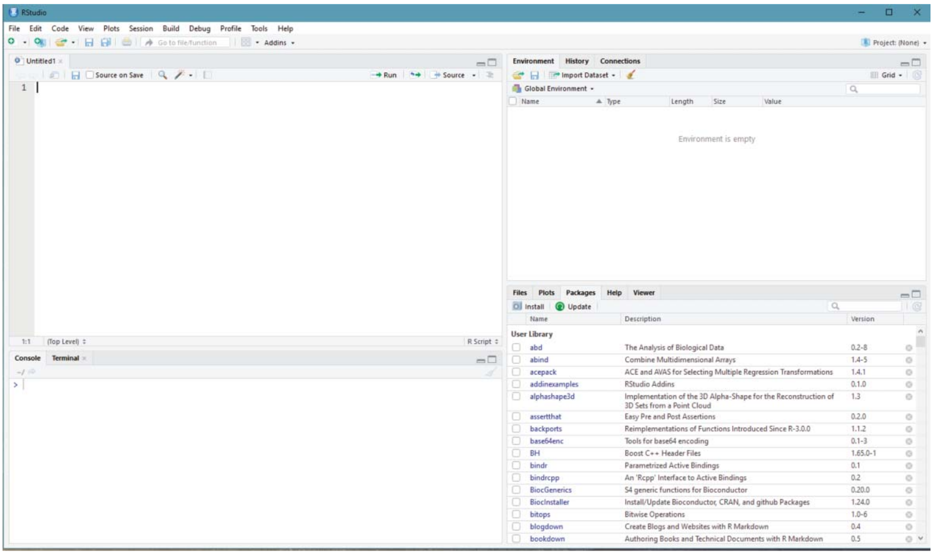

Figure 2. RStudio interface and other windows.

Figure 2 shows four basic windows. The section at the top left is called the script editor. Functional programs can be written with functions and codes in the script editor section. The lower left part is the console and the terminal window. The output of written codes can be seen in the console section. The top right part contains the environment section, where variables are displayed, the history section shows past transactions, and the connections section shows the server or web links. The windows in this section can be changed from the settings. The bottom right section contains the default folder and files section, plots showing plotted graphs, packages showing library downloads and downloaded libraries, help section, and the viewer window that shows maps and complex drawings. In this section, adjustments can be made from the settings menu.

Once the program installation is completed, the work flow diagram of the application is given in Figure 3.

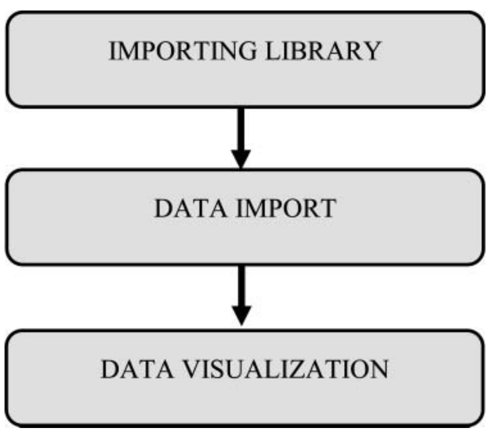

Figure 3. Workflow diagram.

Uploading libraries

To provide the basic information about R before loading the library; the ">" symbol in the R software refers to the beginning of a line and is located at the beginning of each coded new line. Commands in R software usually comprise two basic dimensions: objects and functions. These two dimensions are separated by "<-" symbol in a row. The left part of this symbol is the object, while the right part is the function (object <- function). The "<-" symbol gives the command to the R software that the specified object will be created by the specified function. Anything created in software R is treated as an "object" (Knell 2014). The object can be a variable or a statistical model. The functions are the codes the user enters into the R software to create the object.

The commands to be written in the source section to download the libraries to be used in the study are as follows:

install.packages("leaflet")

install.packages("leaflet.extras")

install.packages("rgdal")

install.packages("raster")

install.packages("Magritte's").

Libraries must be installed into the program so that they can be used in the RStudio environment after the packages are loaded. The commands to be written in the load libraries are as follows:

library(leaflet)

library(leaflet.extras)

library(rgdal)

library(raster)

library(Magritte's)

Once the libraries are loaded smoothly, the data transfer phase is started.

Transfer of data

Files with various formats can be loaded to R platform. It is commonly operable with SPSS, Excel, raster, and vector-based data, and json and geojson formats. Moreover, data from databases can be processed using special packages.

One of the easiest formats to transfer data to R software is the Excel files with a comma-separated values extension. Before the data is transferred and as the RStudio program is installed, the getwd() command is entered in the directory where the data is saved. To work in the desired directory, the necessary directory configuration is done with the setwd() command. One of the most important features of the R programming language is that the

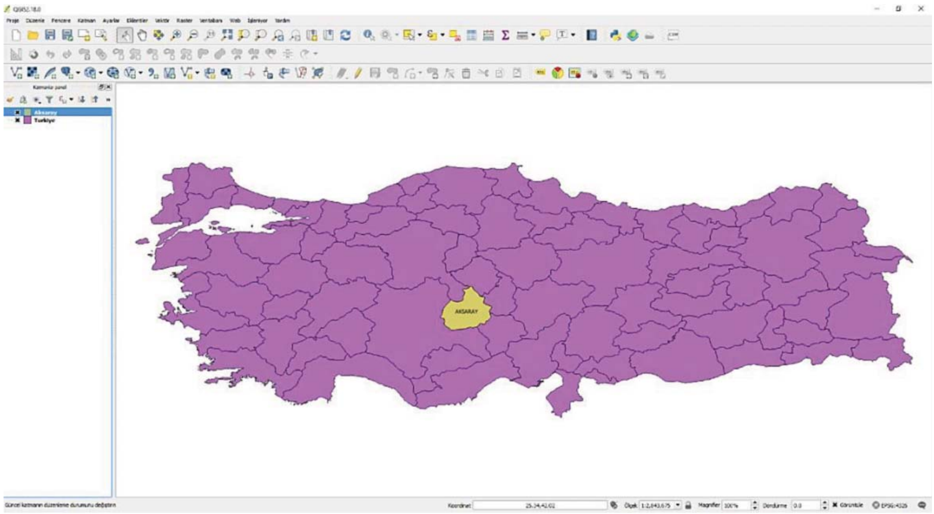

Figure 4. Production of shapefile layers.

auxiliary documents are for all libraries. Owing to this feature, all the necessary documents can be examined from the desired library and the needed information can be obtained. Before uploading the data, the desired command can be written in the help window and detailed information and examples about the package can be accessed.

After opening the vector data of Turkey in shapefile format in QGIS Desktop 2.18.0 program, the data of Aksaray province is extracted and saved with .shp extension. The file of Turkey with the shapefile format is saved separately. The shapefile files belonging to Aksaray and Turkey have been assigned to two variables named as Aksaray and Turkey with the shapefiles() function in R environment (Figure 4).

The file with the extension.csv, which contains the specifics of eight sample points used in this study, has been uploaded. Each column has been named “Point_Number,” “Method,” “pH,” “Conductivity,” “PO4-P,” “PO4,” “Oxygen,” “Temperature,” “Latitude,” and

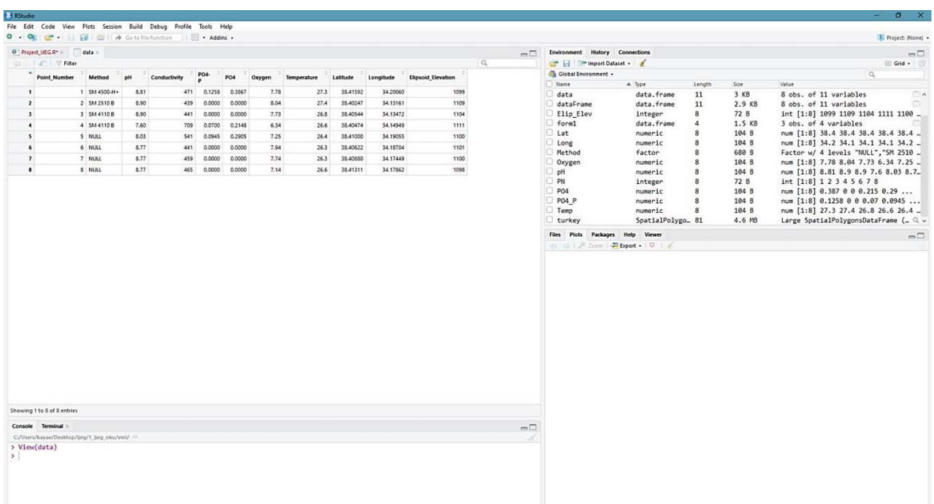

Figure 5. Example of the data set in RStudio environment.

Table 1. Spatial characteristics of the data set.

Point_Number	Method	pH	Conductivity	PO4-P	PO4	Oxygen	Temperature	Latitude	Longitude	Ellipsoid_Elevation
1	SM4500-H+	8.81	471	0.1258	0.3867	7.78	27.3.	38.41592	34.20060	1099
2	SM2510 B	8.90	439	0	0	8:04	27.4	38.40247	34.13161	1109
3	SM4110 B	8.90	441	0	0	7.73	26.8	38.40544	34.13472	1104
4	SM4110 B	7.60	709	0.0700	0.2148	6:34	26.6	38.40474	34.14949	1111
5	Null	8:03	541	0.0945	0.2905	7.25	26.4.	38.41000	34.19055	1100
6	Null	8.77	441	0	0	7.94	26.3	38.40622	34.18705	1101
7	Null	8.77	459	0	0	7.74	26.3	38.40880	34.17449	1100
8	Null	8.77	465	0	0	7.14	26.6	38.41311	34.17862	1098

"Elipsoid_Elevation" using colnames() function. The data set is displayed in a tabular format using the View() function (Figure 5).

Spatial characteristics of the data set are shown in Table 1.

The ncol() and nrow() functions are used to find the number of columns and rows of the data set. There are 11 columns and 8 rows in the data set. While visualizing in the R programming language, the most appropriate data type is the dataframe. Each column in the data set is assigned to one variable as PN, Method, pH, Conductivity, PO4_P, PO4, Oxygen, Temp, La, Long, and Elip_Elev. Data with dataframe type has been assigned to a new variable with the Data.frame() function.

Data visualization

Usage of add markers () function

Once the data set is created, Leaflet, the base map library, is loaded first. Along with the Leaflet function, the base map is accessed. Additional parameters for starting this function can be entered. The minZoom and maxZoom parameters can be entered for zooming in and out of the map. It also uses the addProviderTiles() function to add a map to the base. The addWMSTiles() function can be used if the users want to access maps from web service providers. This can be used also by combining different web-based maps. With R programming, it is easy to make initial adjustments on the map view with many additional parameters using the Leaflet library and other libraries (Url-4).

Sample points gathered from the field can be loaded to the base map in points by addMarkers() function according to latitude and longitude values. The attribute information of the points is added with the command popup = ("information") as an additional parameter. Turkey and Aksaray layers are added with addPolygons() command. On the Turkey layer, the line thickness is 0.5 pixels, the color is black, the fill color is gray, and the transparency grade is 0.2. On the Aksaray layer, the line thickness is 2.5 pixels, the color fill is yellow and the transparency grade is 0.6 (Figure 6).

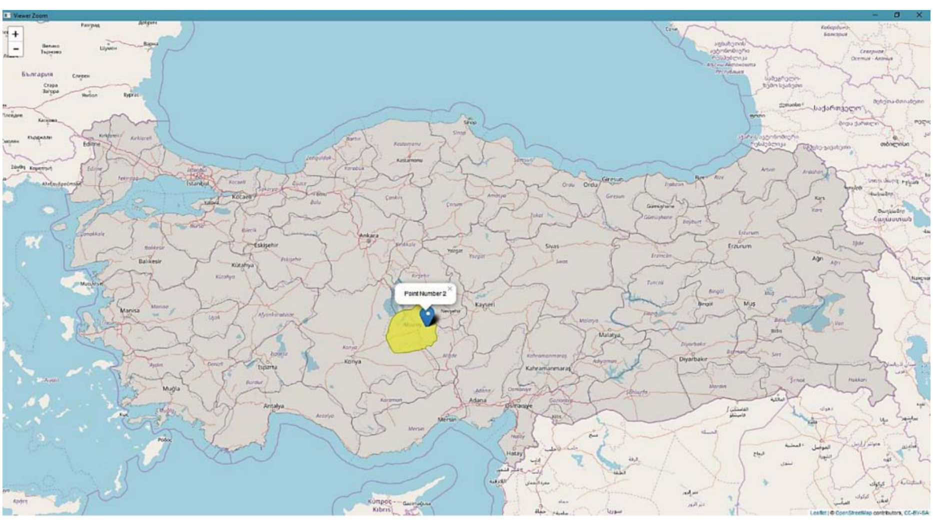

Figure 6. Demonstration of variables on the result map.

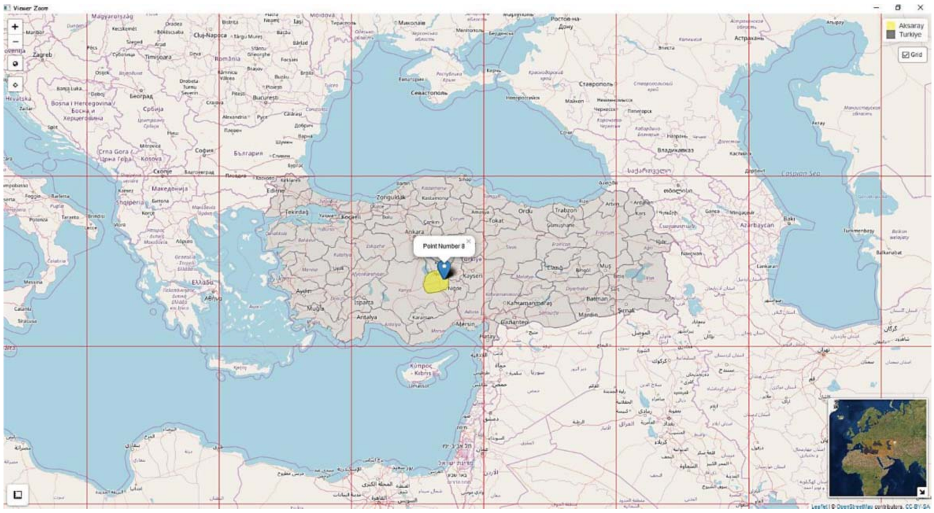

Figure 7. Adding additional features.

After these basic marker properties, if we have created our own icons, we can use these icons to shape our points with makeIcon() function.

Additional features can be added to the map we have created. With the addMiniMap() command, the positions of base maps, with different dimensions, to be decided by the user can be added onto the result map. With the help of the addEasyButton() function, the user can use a button on the map to display the map in the first level (or help return to desired level). This function can also be used to determine the location of users. The addLegend() command adds a legend. With the addMeasure() function, a ruler that allows metric measurements on the map is added. With this function, we can calculate the size of any desired area in different units. With the addGraticule() function, an equally spaced grid is added on the online map. In practice, it is set to one grid per every 6°. Optionally, the grid can be turned off if desired by placing a control button. As an additional parameter to the grid network, html code # FF0000 for color and the value of 2 for line thickness are entered (Figure 7).

The map created in the RStudio environment is saved as WebProject.html in HTML format. It is shared on the website created via Github.

Discussion and conclusion

Countries develop their own software using open source software and avoid license fees using such software in their applications (Güneş 2007). Open source software is advantageous because it provides free data processing, analysis, and presentation of data. As it is widely being used in many institutions and universities, it has become easy to implement. In public institutions or municipalities, construction plans, urban information systems, and other information systems can be created and it allows the opportunity to be presented on the web to the users free of charge.

Budget savings can be easily achieved using open source software in public institutions and municipalities.

Funding

This study has been supported by TÜBİTAK (The Scientific and Technological Research Council of Turkey) within the context of "Assessment and Modeling with GIS and Remote Sensing Data of the Land Use Effects on Water Quality of Mamasin Dam" with the application number of 1919B011502497.

ORCID

Mehmet Cetin http://orcid.org/0000-0002-8992-0289

References

Bimonte S, Schneider M, Mahboubi H, *et al.* 2010. Merging spatial data cubes using GIS overlay operator. J Dec Syst 19(3):261–90. doi:10.3166/jds.19261-290

Brasch S, McPherson EG, and Linsen L. 2009. Interactive data visualization for animating urban tree growth. Int J Model Simul 29(3):306–14. doi:10.1080/02286203.2009.11442538

Chias P, and Abad T. 2013. Wind farms: GIS-based visual impact assessment and visualization tools. Cartography Geograph Inf Sci 40(3):229–37. doi:10.1080/15230406.2013.809231

Doğan CD, and Uluman M. 2016. R Software and Its Use in Statistical Data Analysis, Elementary Education Online, 2016; 15: 2, 615–634 [Online]: http: //ilkogretim-online.org.tr, doi:10.17051/io.2016.24991

Du F, Zhu A, and Qi F. 2016. Interactive visual cluster detection in large geospatial datasets based on dynamic density volume visualization. Geocarto Int 31(6):597–611. doi:10.1080/10106049.2015.1073364

Er F, and Sönmez H. 2005. The usage of R windows for the fundamental statistics education. J Eng & Arch Fac Eskisehir Osmangazi Univ XVIII7(2):1–86.

Güneş İ. 2007. Economic Benefits of Using Open Source Software in Public Institutions: Pilot Implementation Proposal for Local Administrations, Selçuk University Karaman FEAS Journal, Local Economies Special Issue

Ibrahim MZ, Yong KH, Ismail M, *et al.* 2015. Spatial analysis of wind potential for malaysia. Int J Renewable Energy Res 5(1):2015.

İlk Ö. R. 2011. Introduction to Software Book, METU Development Foundation Publishing

Olanda R, Pérez M, Orduña JM, *et al.* 2014. Terrain data compression using wavelet-tiled pyramids for online 3D terrain visualization. Int J Geograph Inf Sci 28(2):407–25. doi:10.1080/13658816.2013.829920

Popov S. 2006. Large-scale data visualization with missing values. Ukio Technologinis ir Econominis Vystymas 12(1):44–49.

Rytkönen MJP. 2004. Not all maps are equal: GIS and spatial analysis in epidemiology. Int J Circumpolar Health 63(1), 9–24. doi:10.3402/ijch.v63i1.17642

Yafang S, and Yongwei S. 1999. Visualizing upwelling at monterey bay in an integrated environment of GIS and scientific visualization. Marine Geodesy 22(2), 93–103. doi:10.1080/014904199273515

Yao X, and Jiang B. 2005. Visualization of qualitative locations in geographic information systems. Cartography Geograph Inf Sc 32(4), 219–29. doi:10.1559/152304005775194683