A novel shiny platform for the geo-spatial analysis of large amount of patient data.

Mario Alessandro Russo[1], Francesco Guarino*[1], Monica Franzese[2], Dario Righelli[1-3], Giovanni Improta[1], Claudia Angelini[2] and Maria Triassi[1]
[1] University of Naples, Department of Public Health, Naples, Italy
[2] CNR, Institute for Applied Calculus "Mauro Picone" (IAC), Naples, Italy
[3] University of Salerno, DISA-MIS, Fisciano, Italy
*Corresponding author: Francesco Guarino email: fguarino@unisa.it

Introduction

Government and international organizations routinely collect huge amounts of data in relation to human health that might be managed through Business Intelligence (BI), a tool extremely useful for conducting epidemiological studies, monitoring and better organizing health-care systems. It has been estimated that up to 80% of all data stored in health care databases may have spatial components[1]. To fully exploit such components, there is a need of improving existing tools or developing novel spatio-temporal functionalities. Geographic information systems (GIS) as QuantumGis, SOLAP[2], etc. are potential candidates to support decisional needs, but despite their capabilities, they are still scarcely employed in association within BI applications. For these reasons, we are developing a GIS user-friendly interface in R environment in order to dynamically and interactively visualize and analyze (within BI platforms) diverse informative data layers (*e.g.,* pathology incidence data, environmental pollution, etc.). Although preliminary, we believe that this kind of tools could be suitable used for epidemiologic, environmental and economical studies by providing geographical maps and statistical data analyses of interest for different stakeholders.

Methods

The graphical user interface (GUI) platform was developed using Shiny libraries[3] and the R environment[4]. In fact, R provides an incredible large amount of statistical analysis tools, it is easily interfaced with BI platforms such as Penthao, and it is possible to use interactive functionalities and GIS-type data features. Moreover, Shiny offers the flexibility of HTML5/javascript. We organized a large database as follows: 1) Data of pathologies were retrieved from data warehouse of GESAN S.R.L., a high tech company that develops software and tool for health informative systems. In particular, GESAN provided anonymous data related to *electronic health records* (EHRs) uploaded by cooperative of roughly 3,000 general practitioners, which look after about 4,600,000 of patients. The EHRs include anonymized personal information (such as age, gender, etc.) and medical prescriptions related to several diseases associated to ICD-9 code collected in Italy (Campania Region) from 2008 to 2016 years. Moreover we also consider 2) the geographic data of administrative boundaries of Campania region downloaded from The Italian National Institute of Statistics (ISTAT) as shape file (.shp) as well as 3) the shape file of environmental data downloaded from website of Campania Environmental Protection Agency (ARPAC), 4) the GPS points of fire and abusive landfill obtained as KML file from terradeifuochi.it, a blog cured by community. Shape file were imported in the R environment through RGDAL[5] package. In general, all the spatial data were preprocessed in order to make them consistent with the reference geography (WGS84) and projection systems (UTM). The pathology datasets were converted into a spatial dataframe through BROOM[6] and SP[7] packages. The static plots were realized by means of GGPLOT2 package[8], whilst, the interactive plots were obtained through PLOTLY package[9].

Results

In this work we show the general infrastructure, the user interface and a case study.
The user interface (Figure 1) allows selecting the Region of interest from a drop down menu, then the list of municipalities comprised in the Region, and the municipalities of interest. From another drop down menu it is possible to select the type of EHRs data to visualize. Finally, in the last drop down menu the user can choose the time period to investigate. Moreover, there are additional menus that enable the overlap other informative layers (*e.g.*, related to fires of waste and/or to risk areas for agriculture, etc.). The area of each municipality was colored with a color scale in relation with the data statistics one is interest in.

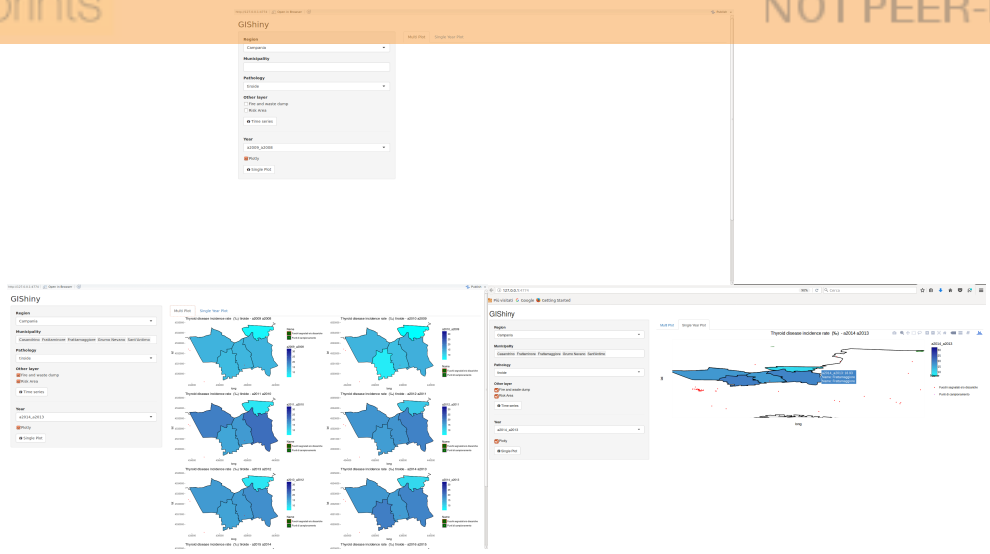

Figure 1

The user can choose to view a time series in a static map to compare the variation of the data in different years by pressing “Time series” button. If the user would visualize the data of a specific year, it is possible to press the “Single plot” button. In this case the GUI will show a static map automatically zoomed on the area of interest. On the contrary, if the user ticks the third flag menu called “Plotly” and the “Single plot” button, a navigable and interactive map will be showed. As case study, we elaborated data from five municipalities of Campania region known to be part of the so called “Terra dei fuochi” (*e.g.*, Casandrino, Frattamaggiore, Frattaminore, Grumo Nevano and Sant’Antimo) that were deeply covered (at least greater than 50% for each municipality) by the cooperatives of general practitioners afferent to GESAN data warehouse in relation to number of inhabitants. We will show results related to chronic diseases related to thyroid analysing the data of roughly 100,000 people for about 8 years. The patients that had at least two medical prescriptions associated with ICD-9 code related to thyroid chronic diseases were selected. In order to aggregate all the chronic thyroid diseases we translated the ICD-9 code into ICPC (International Classification of Primary Care) code. The incidence rate (IR) of chronic disease was calculated as the number of new cases per population at risk in a given time period using subset function included in DPLYR package[10]. These data were elaborated and geo-referenced within the designed pipeline in order to obtain a map composed by different overlapped layers useful to investigate potential relationship between chronic diseases and environmental pressures.

Conclusions

The GUI here described represents a first result of a working in progress project. At the present it is able to process data from geographic and statistical point of view, and to render a map representing a time series of pathology incidence, or a interactively map showing IR for the selected years of interest. Future release, will furthermore improve geospatial statistical utilities, moreover will be able to handle also data such as pharmaceutical expense, economic data, and it will be integrated in open source software of BI as plug-in offering new functionality to these platforms.

References

1 An introduction to geographic information systems: linking maps to databases. Franklin, C. (1992).
2 SOLAP technology: Merging business intelligence with geospatial technology for interactive spatio-temporal exploration and analysis of data. Rivest, S. *et al.* (2005).
3 shiny: Web Application Framework for R. (Chang, W. C., *et al.*, 2017).
4 RStudio: Integrated development environment for R (RStudio, Boston, MA, 2012).
5 rgdal: Bindings for the Geospatial Data Abstraction Library. (Bivand, R. K., T.; Rowlingson, B.;, 2016).
6 broom: Convert Statistical Analysis Objects into Tidy Data Frames. (Robinson, D., 2017).
7 Classes and methods for spatial data in R. (Pebesma, E. J. R. S. B., 2005).
8 ggplot2: Elegant Graphics for Data Analysis. (Wickham, H., 2009).
9 plotly: Create Interactive Web Graphics via 'plotly.js'. (Sievert, C. P., C.; Hocking,T.;Chamberlain,S.; Ram, K.; Corvellec, M.; Despouy, P.; , 2016).
10 dplyr: A Grammar of Data Manipulation. (Wickham, H. F., R. , 2016).