# GeoVisual Analytics of health data using server side approach

Simon Moncrieff, Ulanbek Turdukulov, and Ori Gudes

**Abstract**—GeoVisual Analytics can enable a user to explore multivariate multi-temporal health datasets, to understand spatial distribution of diseases especially in relation to external factors that may influence the outbreaks. External data are presently distributed using geo web services. Web services are used in health mainly to present results leading to a supplier driven service model limiting the exploration of health data. In this paper we illustrate server side approach of designing GeoVisual Analytics environment that allow user driven exploratory analysis. The server side combines a data query, processing technique and styling methodology to rapidly visually summarise properties of a dataset. We illustrate this functionality on a typical analytical workflow used by a health researcher and demonstrate analytical functionality in cases where consistent classification and styling scheme is needed across dynamically aggregated multivariate multi-temporal datasets. And since framework builds on the existing OGC web mapping standards, it integrates the existing geo web services as well as standalone non-spatial database servers such as health data repositories.

**Index Terms**—GeoVisual analytics, server side, health research, geo web servies, user driven analysis

## 1 Introduction

Multivariate spatio-temporal health datasets are both large and complex. For example, annually about one million hospitalisations occur just within the state of Western Australia. All hospitalisation records are kept along with the information regarding patients location, gender, age and number of other additional attributes related to the incident.

Geovisual analytics can enable a user – a health researcher to explore underlying datasets, to analyse trends and understand spatial distribution of diseases. Visualisation is also important in understanding external factors that may influence the occurrence of a disease, particularly in the case of preventable diseases. Occasionally such analysis has to be performed in real time to limit spread of often poorly understood diseases (as in the recent emerging outbreaks of Ebola and MERS). Therefore, combining health related outcomes with socio-economic, demographic and environmental data, and incorporating temporal information for trend analysis, is becoming an increasingly important task in health research.

In parallel, geographic information technologies are evolving from stand-alone systems to a distributed model of independent web services. Large geographic data providers such as Australian Bureau of Statistics (ABS) and Western Australian Land Information Authority (Landgate) distribute socio-economic, demographic and environmental data as geo web services. Integrating, summarizing and visualising these disparate data sources can be a daunting task considering the size and the dynamics of both health and social-economic and demographic data.

The importance of geovisual analytics to explore large sets of geospatial data cannot be overemphasised [2, 1] and is needed in health field. However, at present, web services are not widely used in geovisual analytics (few examples can be found in [8, 11]). The data integration issue in fact points to a larger set of infrastructural problems in geovisual analytics: the existing systems rely heavily on monolithic geospatial data processing systems and lack effective support for distributed and heterogeneous computing environments [7].

In this paper we illustrate server side approach to the design of GeoVisual Analytics environment with emphasis on analytical functionality. The aim of the framework is to provide a method for the dynamic, scalable and secure spatio-temporal visualisation of health data that can utilise various geographic data sources and can be used by different users: ranging from health researchers, decision makers to a public in the future. The design combines a data query, processing technique and styling methodology to rapidly visually summarise properties of a dataset. And since framework builds on the existing OGC web mapping standards, it integrates the existing web services from ABS and Landgate and is used in conjunction with standalone non-spatial database servers such as health data repositories.

## 2 Background

Geographic visualization is a useful, although underutilized, method for examining epidemiological data [6]. Exeter et al. [4] proposed that the linkage of multiple data sets would allow for the more efficient use of health and spatial data, including environmental and socio-economic factors. Use of geo web services for health range from the mapping of layers depicting specific disease outcomes or health services [5], to providing access to specific spatial datasets in the form of interactive map layers [3]. Thematic maps are predominantly used in health using web services to present results, or visualise relatively static snapshots of a dataset [10] resulting in a supplier push service model, in which service provider decides which datasets a user should be able to access.

Although such approaches are helpful for increasing the awareness of spatial information, they limit exploratory and user driven nature of the health analysis. For instance, generating snapshots for exploring a relatively small dataset containing: 9 disease codes, 0-85 age ranges, 3 genders, 1 to N number of time intervals and in 3 spatial resolutions ranging from postcodes, suburbs to municipalities easily results in billions of thematic map layers. Such amount of layers is technically impossible to make available on demand and none of the web mapping servers can supply. Thus existing approaches to enabling web services in health data visualization fail to support exploratory nature of research requirements.

The server side approach described in this paper proposes user need driven service model that allow access to such information, in conjunction with dynamic web mapping and multivariate visualisation to enable the dynamic linkage, and subsequent exploration, of health data.

## 3 A distributed geovisual analytics

A distributed visual analytics architecture requires the coupling of various components. Considering the number of issues, mainly related to the sensitivity of health records, a two tier architecture was adopted. The client is a light weight web interface consisting of *Ext*, *GeoExt*, *Openlayers*, *D3* JavaScript libraries.

- *Simon Moncrieff is with Cooperative Research Centre for Spatial Information (CRC-SI), WA, Australia. E-mail: s.moncrieff@curtin.edu.au.*
- *Ulanbek Turdukulov is with Curtin University, WA, Australia. E-mail: ulanbek.turdukulov@curtin.edu.au.*
- *Ori Gudes is with Cooperative Research Centre for Spatial Information (CRC-SI), WA, Australia.. E-mail: ori.gudes@curtin.edu.au*

*Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014; date of publication xx xxx 2014; date of current version xx xxx 2014.*
*For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.*

A server-side comprises a map server that combines a data query, processing technique and styling methodology on the fly to generate the requested thematic map image accessed through a WMS like REST API [9]. The main advantages of using a server-side technique are: first, the smaller data transfer to each client visualization; second, multiple users can be supported and the user interface is centrally maintained and updated; third, servers can be configured for highly optimised processing functionality, especially when combining datasets; forth, it allows access to results derived from potentially sensitive data-sets for analysis without such data being transferred to the client; compulsory rules and filters can be placed on the server, further restricting output where necessary. This is especially important in privacy sensitive issues such as health data.

The server consisted of two main modules (Figure 1):

- Data stores: Typically comprising one or more databases, or web services containing external data
- Server: contains the logic required to process requests from the client, acting as the interface between the client and the data store.

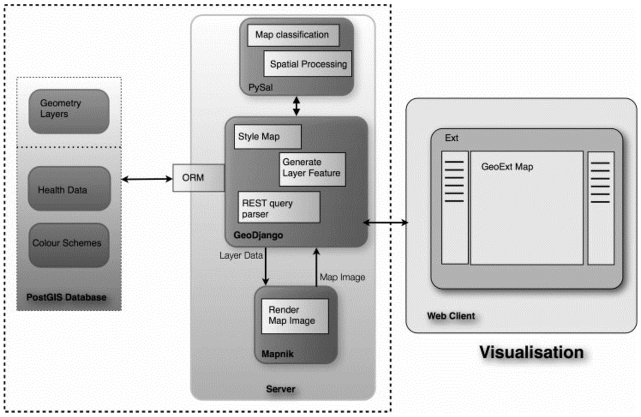


Fig. 1. System architecture of the proposed environment.

In designing the server module, the Model-View-Template (MVT) software architecture approach was adopted, as the *MVT* approach is well suited to implementing REST interfaces and supporting user interfaces with different analytical functionalities. The *GeoDjango* web development framework incorporates support for *PostGIS* spatial database; this enables the server-side logic (View) to access and to query spatial data within a *PostGIS* database (Model) when responding (Template) to a REST query string.

Using the *MVT* architecture, communication of information between the client and the server can be performed using either a markup language (XML) or a data interchange format (e.g. JavaScript Object Notation [JSON]).

## 4 Analytical functionality

Analytical and processing functionality is a core part of GeoVisual analytics. At the lowest level, the processing technique can simply return the dataset feature value. If a query returns multiple values per spatial unit, more complex processing techniques can be applied to produce an aggregate feature for each unit in the generated map layer. Examples of such processing techniques include standard aggregation and annotation techniques, such as *Count* and *Sum*. Using the aggregated data as an input to a function can generate more complex features. This method can include combining multiple datasets or determining a property of the subset of data with respect to the dataset as a whole. For example, a simple epidemiological rate is calculated as the ratio of the events of a given disease to the underlying 'at risk' population.

Further we illustrate the user driven exploratory analysis and analytical functionality in a use case involving dynamically generating health summary statistics and applying a consistent clustering and classification scheme based on a value distribution of the aggregated datasets. Consistent clustering and classification is important when dealing multivariate or multi-temporal datasets since it enables comparison of various maps.

Core of the implemented analytical functionality on the server side consists of *GeoDjango* in conjunction with *Mapnik* that was used to render the final thematic map image resulting from the data processing and map styling process, and the Exploratory Spatial Data Analysis toolbox within *PySAL* was used for both map classification (*esda.mapclassify*) and spatial processing techniques (*esda.smoothing*). Selected map classification techniques from *PySAL esda.mapclassify* library (`pysal.org/library/esda/index.html`) were incorporated into the map server. The techniques included: Natural Breaks, the Fisher Jenks (presented as Natural Breaks - Optimal) and Jenks-Caspall algorithms, quantiles, maximum breaks (presented as Boundaries), standard deviation and mean (presented as Normal Distribution) and Max-P classification (presented as Regionalisation).

## 5 Use case

Two datasets were extracted from the Australian Bureau of Statistics (ABS) data covering Western Australia, aggregated at the statistical local area (SLA) level. A test, synthetic health data-set was incorporated for testing; comprising approximately 700,000 hospitalisation unit records for the state of Western Australia categorized by International Classification of Disease (ICD-10) codes. These data were stored as unit record data with one record per hospitalisation incidence. The data were spatially aggregated to the SLA level, which corresponds to the spatial resolution currently available to the Epidemiology Department at the Department of Health, WA (DoHWA), and also the ABS statistical area 2 (SA2, roughly corresponding to the suburb level). The system is currently installed and is in use by the Department of Health accessing 11 million hospitalisation records.

We illustrate the user driven distributed Geovisual analytics environment on a typical workflow to explore the health data (Figure 2). From the data store located in DoHWA, using morbidity in combination with demographic data, the direct age standardised rate (ASR) is used to determine the health summary statistics. This statistics is linked with socio-economic data is used for visualisation and ultimately, for decision making.

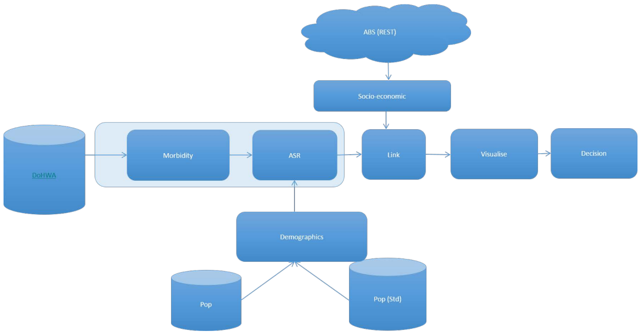
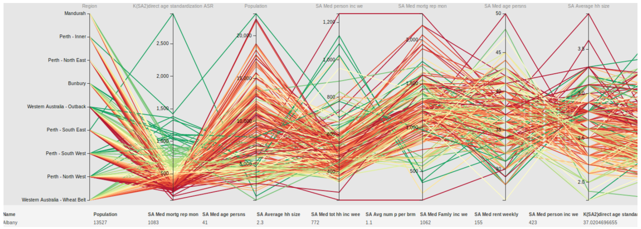


Fig. 2. A typical workflow for analysing health data.

The resulting visualisation represent multiple views of the same information, with the ability to interactively filter the data displayed in within the visualisation.

The first example of a multivariate visualisation is shown in Figure 3 a parallel coordinate plot displaying a spatial health summary statistic, along with a number of socio-economic and demographic attributes, generated using the parallel coordinate API within D3. The ASR was used to determine the health summary statistic, and was calculated for Chapter K of the International Classification of Disease (ICD10) categorisation scheme, corresponding to diseases of the digestive system, with the at risk population being determined using the ABS demographic information. The spatial resolution used corresponds to the ABS SA2. In this case, the visualisation was initialised using output from a processing web service determining the summary statistic for the hospitalisation event associated with Chapter K within the database.



| Name | Population | SA Med mortg rep mon | SA Med age persns | SA Average hh size | SA Med tot hh inc wee | SA Avg num p per brm | SA Med Famly inc we | SA Med rent weekly | SA Med person inc we | K(SA2)direct age standa |
|---|---|---|---|---|---|---|---|---|---|---|
| Albany | 13527 | 1083 | 41 | 2.3 | 772 | 1.1 | 1062 | 155 | 423 | 37.0204696655 |

Fig. 3. Parallel coordinate plot indicating spatial health summary statistic.

Figure 4 shows the comparison map between ICD 10 diagnostic code Z with the ABS individual income data; using consistent classification and clustering methods. Additionally, the visualisation interface also shows an SVG error bar plot that illustrates confidence intervals of the processed aggregated attribute (epidemiological summary statistic). Providing ansilliary information such as the confidence intervals is one of the advantages of the proposed methods since it provides context to the spatial information presented in the map view. Ineractivity is enabled through embedding on hover information within the SVG graph, and by linking the maps such that a change in one map triggers the same change in the remaining map. The graph emdebbed information can be extended to provide high level interpretation regarding the confidence interval.

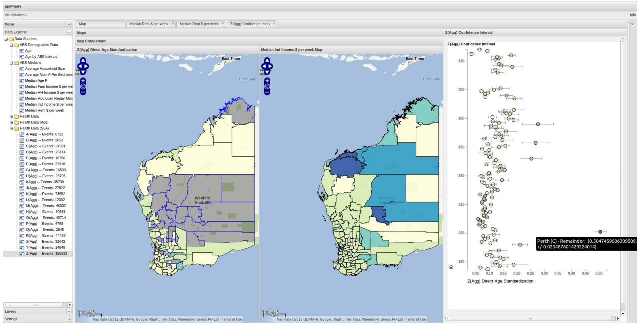


Fig. 4. Comparison map between ICD 10 diagnostic code Z with the ABS individual income data along with graph that illustrates the confidence intervals.

Figure 5 shows a spatial scatter plot matrix generated using the using combination of choropleths, bivariate choropleths, and scatter plot quantum visualisation elements. Each row comprises a data variable that constitutes the x value of the visualisation, and each column represents data variable that constitutes the y value. Consequently, the diagonal is use to display a standard choropleth. The off diagonal visualisations are used to show the spatial and non-spatial bivariate comparison between the x and y variables. The spatial visualisations comprise bivariate choropleths, and the non-spatial visualisations correspond to a standard scatter plot matrix with the histograms for each variable added for context. Due to the nature of the matrix, the variable x/y comparisons are flipped on the transpose. The map visualisations are similarly linked, with a change in one triggering a change in each of the remaining map images.

Figure 6 shows an SVG scatter plot matrix showing similar data. This scatter plot matrix enables filtering, with data point selected in one plot being highlighted in the remaining plots. The combination of presenting multiple scatter plots in the form of a matrix, and enabling filtering, facilities the rapid determination of the relationship that exists between attribute pairs. Thus, the parallel axis plot, and scatter represent different methods of viewing the same information, with each revealing different aspects of the data.

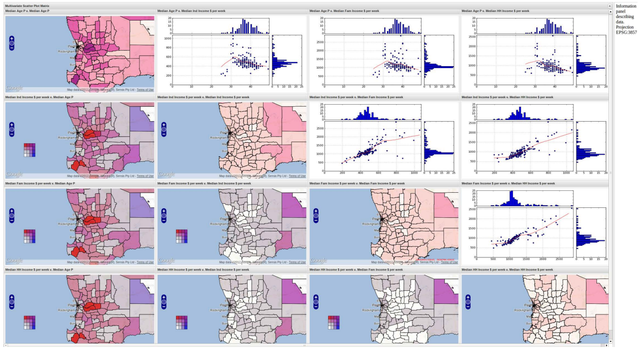


Fig. 5. spatial scatter plot matrix.

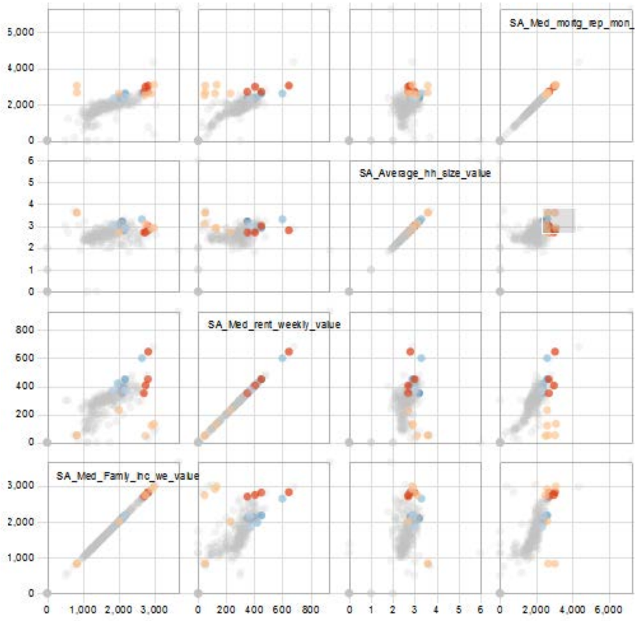


Fig. 6. Scatter plot matrix.

## 6 Conclusion

The approach to web mapping proposed in this paper constitutes the first phase in the development of a full geo analytical platform for exploring the properties of large health data-sets. Consequently, there are a number of issues remaining to be addressed as future work relating to usability, privacy and security. Usability with respect to a GeoVisual analytical functionality remains an open issue, particularly given the workflows and complexities of analysis envisaged by such a system. Nevertheless, the presented approach allows for extensible analytical functionality and more importantly, it breaks the barriers of supplier driven model servicing only the static results, towards real-time, interactive exploratory analysis of health data in conjunction with data sources from other geo web services.

## Acknowledgments

This work has been supported by the Cooperative Research Centre for Spatial Information, whose activities were funded by the Australian Commonwealths Cooperative Research Centres Programme.

## References

[1] G. Andrienko, N. Andrienko, J. Dykes, S. I. Fabrikant, and M. Wachowicz. Geovisualization of Dynamics, Movement and Change: Key Issues and Developing Approaches in Visualization Research. *Information Visualization*, 7 (3/4)(3/4):173–180, 2008.

[2] G. Andrienko, N. Andrienko, P. Jankowski, M.-J. Kraak, D. Keim, A. MacEachren, and S. Wrobel. Geovisual Analytics for Spatial Decision Support. Setting the Research Agenda. *International Journal of Geographical Information Science*, 21 (8)(8):839–857, 2007.

[3] C. Cinnamon, J., M. Rinner, S. Cusimano, T. Marshall, R. Hern, R. H. Glazier, and M. Chipman. Evaluating Web-based Static, Animated and Interactive Maps for Injury Prevention. *Geospatial Health*, 4(1):3–16, 2009.

[4] D. J. Exeter, S. Rodgers, and C. E. Sabel. ”Whose data is it anyway?”, the implications of putting small area-level health and social data online. *Health Policy*, 2013.

[5] D. H. Foley, R. C. Wilkerson, S. Birney, I.and Harrison, J. Christensen, and L. M. Rueda. Mosquitomap and the mal-area calculator: new web tools to relate mosquito species distribution with vector borne disease. *International Journal of Health Geography*, 9(11), 2010.

[6] K. Joyce. To Me Its Just Another Tool to Help Understand the Evidence: Public Health Decision-makers Perceptions of the Value of Geographical Information Systems (GIS). *Health and Place*, 15(3):831–840, 2009. doi:10.1016/j.healthplace.2009.01.004.

[7] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler. Challenges in Visual Data Analysis. In *10th International Conference on Information Visualisation*, pages 9–16, Los Alamitos, USA, 2006. IEEE Computer Society Press.

[8] M. Kramis, C. Gabathuler, S. I. Fabrikant, and M. Waldvogel. An XML-based Infrastructure to Enhance Collaborative Geographic Visual Analytics. *Cartography and Geographic Information Science*, 36(6):281–293 (13), 2009.

[9] S. Moncrieff, G. West, J. Cosford, N. Mullan, and A. Jardine. An open source, server-side framework for analytical web mapping and its application to health. *International Journal of Digital Earth*, 7(4):294–315, 2014. http://dx.doi.org/10.1080/17538947.2013.786143.

[10] T. Samarasundera, E., T. Walsh, A. Cheng, K. Koenig, A. Jattansingh, Dawe, and M. Soljak. Methods and Tools for Geographical Mapping and Analysis in Primary Health Care. *Primary Health Care Research Development*, 13(1):10–21, 2012.

[11] U. Turdukulov, C. A. Blok, B. Köbben, and J. Morales. Challenges in data integration and interoperability in GeoVisual Analytics. *Journal of Location Based Services*, 4(3-4):166–182, 2010.