Web based statistical data processing

G. Heinz, A.W. Mees

Römisch-Germanisches Zentalmuseum, Mainz, Germany
heinz@rgzm.de; mees@rgzm.de

1. Introduction

The Römisch-Germanisches Zentralmuseum Mainz is involved in many IT projects in archaeology. It uses various standard software packages for its GIS, 3D-scanning, database and statistical applications.

However, scientific archaeological work requires adaptations of existing packages and also especially written software routines. Because of the irregularity of many archaeological data, better suitable algorithms than available in commercial or open source software have been developed over the years.

This presentation will show newly developed analysis tools for statistical mapping (Chi-square, Mócsy value mapping, Distribution average midpoints) with datasets of more than 100000 records and will discuss the software techniques used to treat these data.

The mentioned mapping applications have in common, that the statistical calculations are performed in a Web-based environment after the user has run database queries. Dependant on available attributes of the data, queries might select e.g. certain time periods, producers, production centres. The statistical calculations are performed using the results of these queries. So the mapping is highly interactive and can easily be adapted to other data by changing the query statements. Technically the applications are generated by server based scripts, building the database query from values coming from the web query masks and then executing the query. The returned query results are used in the statistical analyses and finally displayed or mapped.

2. Statistical mapping

As a method of displaying statistically weighted dots on a map, it turned out to be particular helpful to be able to plot the expected Chi-square values of an average distribution on a map in a color range (Figure 1).

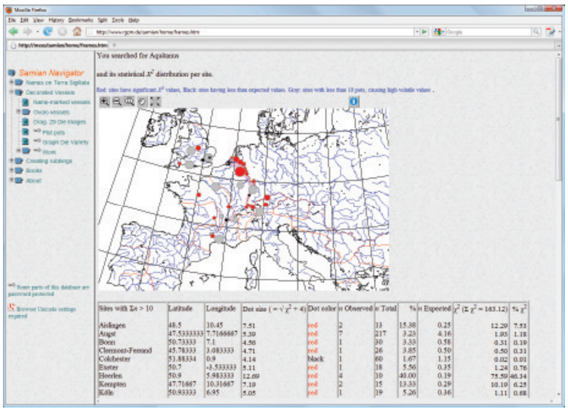

Figure 1 *Distribution map showing Chi-square values for findspots (Datasource: DANNELL et al. 2003, Aquitanus).*

Since archaeological data is often not found equally distributed and the minimum of the sum of expected values per type in a unit does quite often not yield 5, the minimum requirements for Chi-Square analysis is often not met and its usage in archaeology is often suspect.

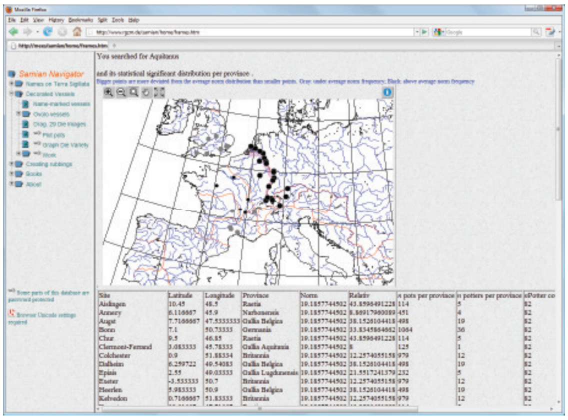

Figure 2 *Distribution map showing Mócsy values for findspots. (Datasource: DANNELL e.a. 2003, Aquitanus).*

For this purpose, an additional test according to the Mócsy coefficient was developed. This proved to be extremely helpful, since it particularly takes into account more weakly presented but still important types and units (Figure 2).

Many archaeological materials show different distribution patterns during a specified time period. A good method to show this development is to map the geographical midpoints of each find category during its chronological development. The average deviation within the distribution of each find category displayed as ellipses around the midpoint gives the archaeologist a criterion to judge whether unattributed find categories may belong to this particular group (Figure 3).

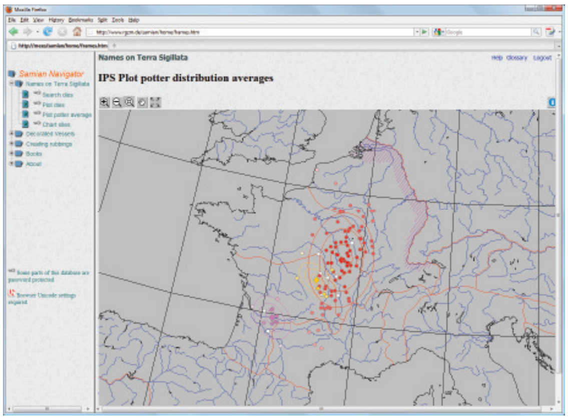

Figure 3 *Distribution map showing the midpoints of potter distributions together with the midpoints of the production centres. The ellipses are showing the standard deviations and their orientation of the potters per productioncentre between 40 and 70 AD (Datasource: HARTLEY / DICKINSON e.a. 2008).*

When we started plotting dynamic maps, internet services were not yet available with the functionality required. Therefore a self developed Java applet to plot the maps is used, in order to be able to set the display parameters via dynamic querying and have additional functionality which is not available in current web services. Environments discussed are ColdFusion, Java and PHP, in which the algorithms are coded. The backend database can be any RDMS. The completely web based workflow allows to download querying results in GMT-ready data formats in order to generate high resolution maps offline as well.

3. Managing coordinates

Since much of this mapping work requires coordinates, an internet module was added to search findspot coordinates. The basis for searching coordinates is a geographic names database from the National Geospatial-Intelligence Agency (NGA 2010). This database includes name variations and also historical names which is particularly useful when working with old archaeological literature (Corpus Inscriptionum Latinarum) and Eastern European town names which changed many times in recent history. Since many archaeological sites are not precisely located in modern datasets (e.g. La Graufesenque), it is necessary to be able to add additional records. This is realized by picking the position using current webservices offering cartographic, terrain and image information (e.g. Bing Maps, GoogleMaps) (Figure 4).

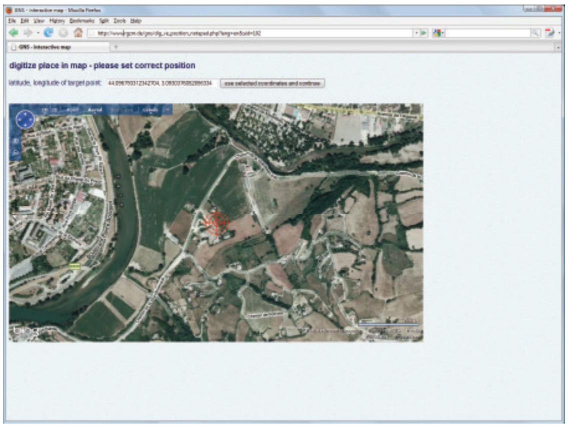

Figure 4 *Measuring the coordinates of archaeological sites using Microsoft Bing Maps.*

Each user can store online his collected coordinate sets, edit these coordinate sets, share them with other users and download them in any stage of the work in progress (Figure 5). These private coordinate sets can also be visualized using the Java based plotting applet or programs or services like GMT, GoogleEarth or Bing Maps.

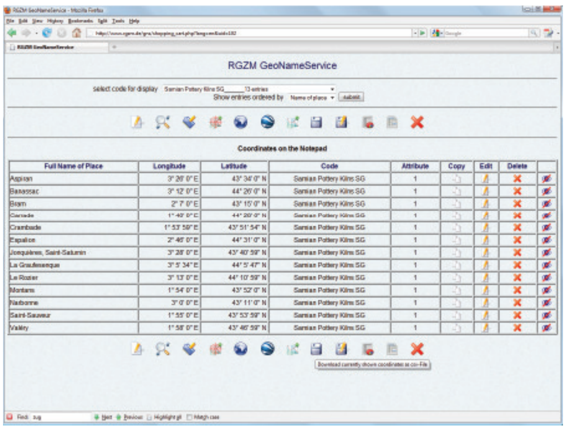

Figure 5 *User defined private coordinate list with links to map or download the collected coordinates.*

These routines are not available in commercial packages in this form and combination and therefore had to be developed in the RGZM research institute related to special scientific tasks.

4. Statistical webservices

A logical consequence of this entirely internet based work was also to concentrate on the porting of the popular parts of good old WinBasp package (SCOLLAR 2005) as well, which had huge popularity among archaeologists, but had always the problem of its missing ODBC interface, so that students with no command line experience suffered enormous difficulties in importing their MSAccess or MSExcel data into this package. With the newly established RGZM portal www.rgzm.de/adp, everyone can now use the web to enter their MSAccess or MSExcel output saved as ASCII datasets to generate 3D rotating Correspondence Analysis (Figure 6) and run Seriations (Figure 7), using as a back end the open source statistical software from the R project plus additional packages for special calculations.

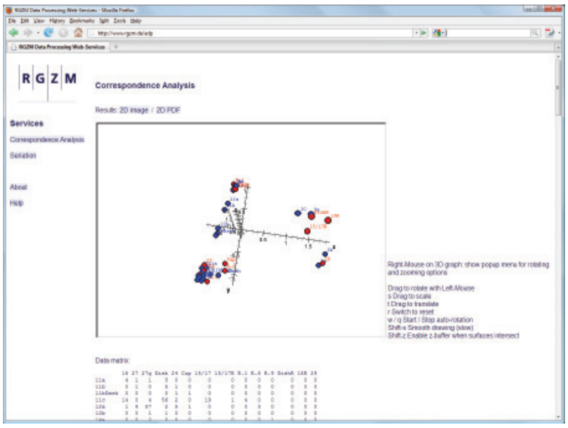

Figure 6: *Correspondence Analysis of an ASCII data file uploaded by a user. (Data source: HARTLEY / DICKINSON e.a. 2008, Aquitanus).*

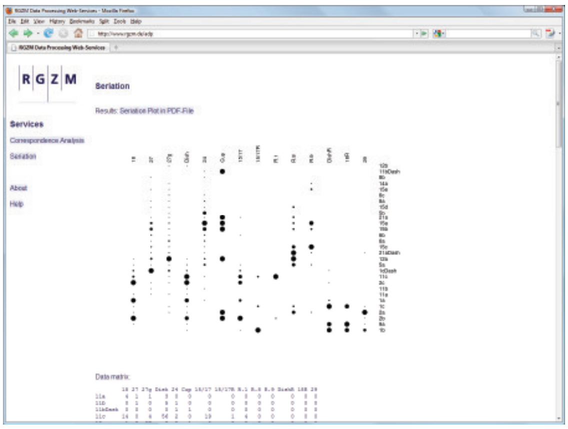

Figure 7: *Seriation of an ASCII data file uploaded by a user. (Data source: HARTLEY / DICKINSON e.a. 2008, Aquitanus).*

The above mentioned Chi-square, Mócsy coefficients and standard deviations of average distribution midpoints mapping will be made available there for everyone to use.

The presented examples show an environment with a selection of functions resulting from the archaeological practice in projects with large numbers of find related database records with spatial information. Using a web interface with suitable techniques allows sharing the data and analysis tools and a highly interactive approach. However, making these functionalities available as web services to a much wider range of archaeologists does even more mean that an intrinsic understanding of statistical basics and algorithms mentioned before should become part of the archaeological education.

Literature

DANNELL et al. 2003: G.B. Dannell / B.M. Dickinson / B.R. Hartley / A.W. Mees / M. Polak / P.V. Webster / A. Vernhet (ed.), Gestempelte südgallische Reliefsigillata (Drag. 29) aus den Werkstätten von La Graufesenque Gesammelt von der Association Pegasus - Recherches Européennes sur La Graufesenque. Römisch-Germanisches Zentralmuseum, Forschungsinstitut für Vor- und Frühgeschichte. Kataloge vor- und frühgeschichtlicher Altertümer 34 (Mainz 2003).

HARTLEY / DICKINSON et al. 2008: B.R. Hartley / B.M. Dickinson, with G. B. Dannell, M.G. Fulford, A.W. Mees, P.A. Tyers, R.H. Wilkinson, Names on Terra Sigillata. An Index of Makers' Stamps & Signatures on Gallo-Roman *Terra Sigillata* (Samian Ware). Institute of Classical Studies 102, London 2008.

SCOLLAR 2005: http://www.uni-koeln.de/~al001/

NGA 2010: http://earth-info.nga.mil/gns/html/index.html