A spatio-temporal knowledge-discovery platform for Earth-Science data

TCW Landgrebe
School of Geosciences
Faculty of Science
The University of Sydney
Sydney, Australia
Email: thomas.landgrebe@sydney.edu.au

RD Müller
School of Geosciences
Faculty of Science
The University of Sydney
Sydney, Australia
Email: dietmar.muller@sydney.edu.au

***Abstract*—The GPlates Geographic Information System (GIS) is a well-established tool for visualising and interacting with multimodal geo-data reconstructed through space and geological time. It overcomes the complexity of changing spatial reference frames due to plate-tectonic processes. Combining vast datasets in this manner is increasing the analysis complexity, with traditional visualisation-based approaches becoming ineffective in extracting key information and discovering hidden associations. This paper discusses the nature of these complexities, followed by the presentation of an extension to GPlates, involving the addition of interactive quantitative data-mining tools and capabilities, better suited to coping with the inherent analysis complexities. A case-study is used to demonstrate the system's unique capabilities. The integrated software infrastructure has manifested itself as a powerful knowledge discovery platform, which has the potential to lead to new high-impact discoveries in the Earth Sciences.**

I. Introduction

The GPlates plate tectonic reconstruction tool [1] is an open-source, cross platform Geographic Information System (GIS) with the unique ability to reconstruct geo-data through both space and time. The software was developed by the EarthByte group as a collaboration between The University of Sydney, Caltech and the Geological Survey of Norway, with a focus on data pertaining to "deep time" Earth Science. GPlates has played an intrinsic role as a primary tool in several high impact studies involving interactions between major Earth processes on a global scale e.g. [2], [3], [4].

The large-scale accumulation of digital geo-data, interoperability standards (e.g. GeoSciML [5]) and improved interconnectivity is now creating new opportunities within Earth Science to amalgamate traditionally disparate datasets (and communities) across several modes of data including geochemistry, geophysics, structural geology, plate tectonics etc. This in turn is facilitating answering complex scientific questions such as how continents evolved, which can only be answered by incorporating all these modalities. The impact of such an approach has significant potential in applications such as creating predictive models for mineral and energy exploration, and in designing more effective geo-hazard models.

GPlates bridges a technological gap in allowing a large variety of spatial data-types to be attached to a unified plate-tectonic reference frame, facilitating the reconstruction of Earth processes acting through geological time. In this paper we discuss the fact that combining the various datasets together results in a significant increase in the analysis complexity, limiting the effectiveness of traditional visualisation-based approaches. A quantitative spatio-temporal extension is proposed, consisting of two primary components, namely a *coregistration tool* for defining relationships between datasets in a flexible, recursive fashion, and a *data mining environment* for customising a particular analysis workflow, and making use of advanced unsupervised and supervised statistical approaches for studying the intrinsically high-dimensional data. A visual-programming environment with a plugin infrastructure is used to design sophisticated workflows without requiring software programming skills, with a library of high-level processing units allowing for workflow configuration at the appropriate conceptual level. GPlates together with these additional components is now becoming a powerful knowledge discovery tool, with flexibility and interoperability built in to maximise both the depth and breadth of possibilities.

The paper is structured as follows: in Section II the design considerations are discussed, followed by an outline of the design in Section II. The two primary modules developed are then discussed, namely a flexible spatio-temporal data coregistration tool in Section IV and an interactive data-mining environment in V. The efficiency of the system is demonstrated via a case study in Section VI, and conclusions are presented in Section VII.

The primary contribution of this paper is the presentation of the high level design considerations and technology components that have been translated into a software implementation. This technological step is helping to shift what has traditionally been a qualitative data exploration methodology (and community) into the more quantitative realm, where increasingly complex data analyses can be undertaken using powerful data mining and image processing methodologies.

II. Design considerations

In developing the knowledge-discovery framework, a number of considerations need to be made, ensuring a design that is user-friendly, flexible and extensible. In this section the primary considerations are discussed, leading to a high-level

design. In order to contextualise the discussion, the example in Figure 1 is presented, demonstrating the reconstruction of several spatial datasets at two different time intervals. Even in this relatively simple study, the large number of possible interactions and potential associations is evident.

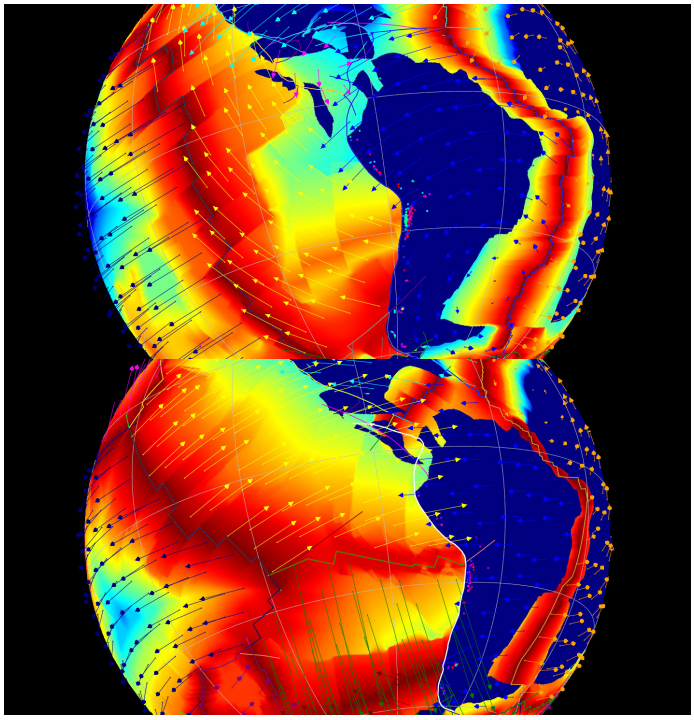

Fig. 1. An example study involving investigation of the time-dependent associations between ore deposits along the Andes, and the plate-kinematic dynamics along the Andean subduction zone (blue line) for two time slices, namely 60 million years in the past (top), and 100 million years (bottom). The arrows depict time-varying convergence velocities, coloured points along the Andes represent the relative mineralisation ages of ore deposits, and the coloured background raster represents the reconstructed age of the sea-floor.

A. Data-analysis complexities

Combining several Earth Science datasets together presents a number of challenges:

- High dimensionality: as more associations between datasets are combined, the dimensionality of the resultant derived datasets increase significantly. This makes it more difficult to identify important relationships using visualisation methodologies, and the significance of identified co-associations between variables require large datasets to justify statistically.
- Large datasets: developing increasingly complex models of Earth processes typically involves large datasets to achieve statistical confidence. Larger datasets can make identifying important patterns and relationships more difficult, and increases the computational and analysis burden.
- Spatial and temporal variations: relationships and associations varying both spatially and temporally over a moving spatial reference frame increases the complexity of assessment, and consequently identifying significant structure/associations within the data becomes more difficult. Several levels of spatial scale must also be considered.
- Uncertainty, variability, noise and missing values: these are inherent factors that are important to incorporate in order to develop models that generalise well, and detect associations that are significant.
- Mixed data types: Spatial geo-data is a mixture of a variety of different data types, from vector to raster-geometry, with a combination of categorical, ordinal and real-valued properties (meta-data). Other software tools are also often involved in an analysis workflow, the outputs of which are integrated with other spatial data. The scaling and distribution of such model outputs requires careful attention when co-analysing with other spatial data.

Data mining and statistical analysis approaches are important for dealing with these challenges. These approaches explicitly deal with measurement variability and noise, and can simultaneously assess large datasets without bias. High dimensionality is dealt with using approaches such as decorrelation, feature selection and projections to lower-dimensional spaces. The AILab Orange data-mining software [6] has been chosen as a suitable tool for this purpose, providing a large library of supervised and unsupervised components, as well as components to filter, map and select data as required. Importantly this library allows external plugins to be developed. Such a capability is required in order to develop preprocessing, translation and mapping tools specific to time-varying spatial data, resulting in data structures converted to a form suitable for data-mining.

B. Multiple use-cases

Multitudes of workflows are conceivable for analysing problems pertaining to the Earth Sciences, and thus a software solution cannot be tailored to a few specifications. Rather the design should allow new workflows to be implemented in a flexible manner, providing interfaces facilitating custom-development of new and unique workflows. A solution that allows for abstraction, extensibility and reuse would also be important, lending itself to continuous improvement and collaboration. The approach chosen is to develop a *library* of high-level modules that are made available to a user. The user would then set up relationships between the various modules according to the required data flow.

C. Abstracting programming complexity

The users of the software are typically scientists with the need to implement complex quantitative analysis workflows. The design should allow users that are not expert programmers to customise the required work-flow, where effort is spent on solving the underlying scientific problem, whilst minimising other overheads. The approach adopted uses a *visual programming* approach, in which modules are developed to achieve a specific high-level functionality (analagous to a building

block), providing an abstraction of the underlying programming complexity. Users then connect these modules together visually, achieving a workflow that mimics the conceptual data-flow. The AILab Orange Visual Canvas [6] provides such an interface, as well as flexibility in a number of other areas.

D. Other

In order to maximise the accessibility of the software, and for compatibility with GPlates licensing, an open-source, cross-platform solution is necessary. Another requirement is a mechanism for interactivity between the data mining and GIS environments. This allows for combined visualisation and quantitative approaches. Consider for example geochemical data, where cluster analysis is applied to the multivariate elemental properties in order to identify samples with similar composition. In this case the data mining task would be to perform the cluster analysis and assign class labels, which could then be used to colour the respective spatial locations in the GIS environment.

III. High-level design

Section II discussed the primary factors taken into consideration for developing a successful design, as well as some design decisions taken. In Figure 2 the primary components of the GPlates knowledge-discovery platform are depicted. A

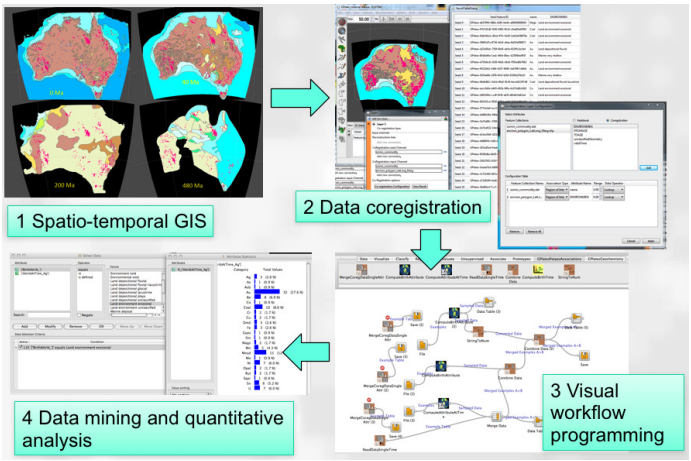

Fig. 2. High-level depiction of the GPlates knowledge-discovery platform, consisting of the interactive spatio-temporal GIS, coregistration tool, visual-workflow builder, and data-mining environment.

data coregistration tool is the interface between the spatio-temporal GIS domain and the data-mining domain. This tool allows relationships to be defined between datasets across time, and operates transparently across different data and geometry types. The coregistration tool allows multiple relationships to be specified recursively between a *seed* and any *reference* dataset. For example the distance between the seed and reference datasets can be computed, or a particular property pertaining to a reference dataset at the location of the seed (for the example in Figure 1, typical associations are computing the age of the downgoing sea-floor at each ore-deposit location, and the distance to the subduction zone). The output of this component is an array, in which each row corresponds to a particular seed geometry, and columns correspond to the chosen association. The array has a third dimension, capturing associations that vary with time, i.e. each array cell results in a time series. The coregistration tool thus translates data in the complex spatio-temporal data representation into a well-defined format that can be processed directly by signal processing and data-mining tools.

A *visual programming* software tool provides a library of high-level components, together with a design *canvas* upon which users develop a desired workflow by interactively placing and connecting module components. These components abstract programming complexity, allowing the workflows to be developed in an intuitive fashion without unnecessary programming complexities. Finally an interactive *data-mining* infrastructure provides capabilities to assess data, discover unseen associations, and train statistical models on data originating from the coregistration tool. Feedback is provided to the GIS to allow spatial and temporal feedback from a quantitative analysis (as demonstrated in the case study).

IV. Data coregistration tool

The coregistration tool is used to recursively define desired relationships between datasets, and is transparent to both geometry and attribute types. Defining an association between a seed and reference dataset involves four primary steps, namely *seed* dataset selection, *reference* dataset selection, association-type selection, and then defining the association parameters. The tool is used recursively to systematically build a table of associations, allowing for large numbers of simultaneous analyses. The output is a table consisting of a number of rows (corresponding to each seed), and columns (the number of coregistration operations defined), repeated across time according to the number of time steps. This output is then fed to the input of a data-mining tool. An example demonstrating the definition of two time-varying relationships between a seed and reference dataset is shown in in Figure 3. An example coregistration output is then shown in Figure 4 at a particular point in time. In this example mineral deposits are compared to their geological environment through time.

A. Seed dataset definition

A seed dataset consists of a base dataset with respect to which other associations or relationships are to be computed e.g. known gold ore-deposit locations where we would like to investigate their relationships with other datasets. The output of the coregistration tool will consist of a data structure with a number of rows consisting of the size of the seed dataset.

B. Reference dataset definition

In a second step a dataset for which a relationship is to be computed with respect to the seed dataset is chosen. For example we might choose to study the gravitational anomaly data at the locations of known gold deposits.

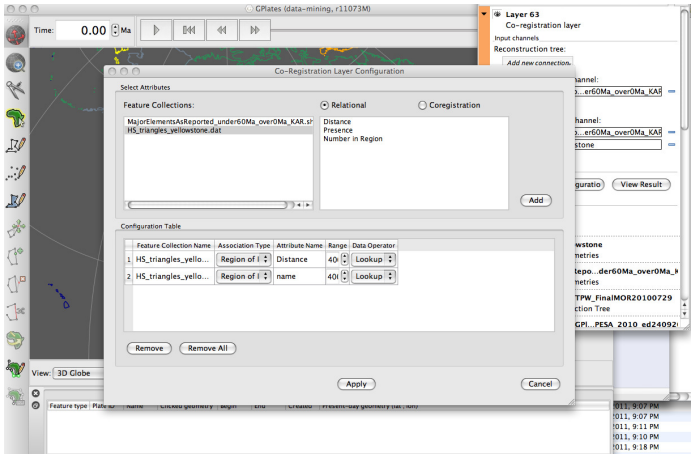

Fig. 3. Example showing specification of two coregistration specifications. The distance from the seed dataset (MajorElementsAsReported) is computed between the nearest reference dataset (Hotspots) and it's corresponding name. The time-series of these distances can then be computed through time.

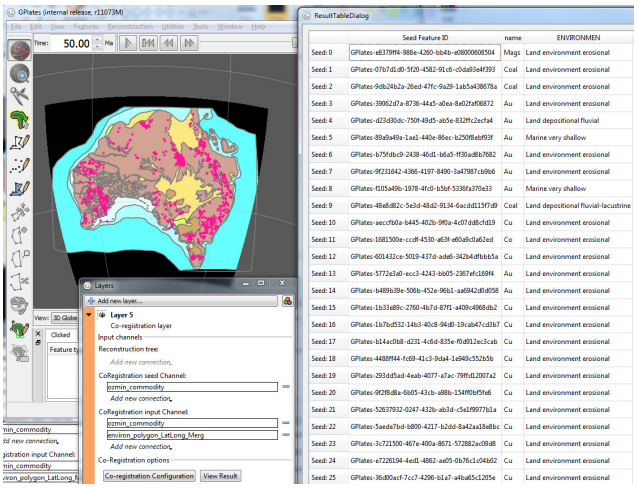

Fig. 4. Example coregistration tool output for a particular point in time. In this example mineral deposits in Australia are analysed with respect to palaeo-geography i.e. the time-varying geography.

C. Defining the association type

Two types of associations are possible. The first are relational associations, in which statistics pertaining to the *proximity* between the seed and reference datasets are computed, such as their relative distances through time, the number present within a region of interest etc. An example would be the distance between gold deposits and the nearest faults. The second type is a *neighbourhood* association, in which properties or characteristics of data in the same neighbourhood as the seed data is computed. For example, we may wish to compute the gravitational response or mean elevation at the location of each gold deposit.

D. Association operator definition

Once the reference dataset (with corresponding association type) has been defined, the final step is to define the nature of the association, resulting in a scalar outcome. For example, when computing the gravitational response at seed gold deposit locations, we would define the necessary parameters at this point, for example specifying the mean value within a radius of 2km.

V. Interactive data-mining environment

The output of the coregistration tool is essentially a 3-dimensional matrix structure, with rows corresponding to *seed* geometry entities, and columns to defined relationships with other datasets. The third dimension characterises the temporal variations in the associations. The AILab Orange [6] software tool has been combined with GPlates to fulfil the data-mining role. It consists of a comprehensive library of functionalities, with the following notable features:

- Data manipulation tools: Several powerful components are present to flexibly combine, concatenate, filter by multiple criteria, sort and map data.
- Interactive visualisation tools: Statistical summaries, visualisation of projected data, and scatter-plots aid in finding associations and understanding the data.
- Machine learning tools: A comprehensive set of unsupervised and supervised classification components exist, which are essential for the desired knowledge-discovery capabilities e.g. finding structure in data, clustering similar multivariate patterns, building statistical models and testing hypotheses.

An important aspect is the ability to develop custom-components, complementing the library of existing general tools with application-specific ones. A plugin library has been developed for GPlates, consisting of a number of specific components for extracting features from time-series, investigating the geological environment in the past, or data-patterns leading to the occurrence of a particular phenomenon e.g. an ore deposit. AILab Orange also provides a visual programming environment, thus allowing workflows to be developed interactively by connecting required components together. As discussed earlier, this capability is very attractive for this design in abstracting underlying programming complexities. An example of the visual programming environment, also depicting the GPlates plugin library, is shown in Figure 5 below, demonstrating how a specific data-mining workflow has been developed to solve a particular problem.

VI. Case study: unravelling a complex geodynamic environment

The present-day configuration of the Earth's crust in North America between the Yellowstone hotspot and the adjacent region extending to the west coast evolved into its present state via a complex combination of magmatic processes acting both at the plate boundary and from below due to a mantle plume. This interaction occurred over the past 60-80 million years, as the North American continent shifted over the Yellowstone hotspot (see e.g. [7]). In this case study we combine a number of datasets through space and time using the GPlates knowledge-discovery tool, with the objective

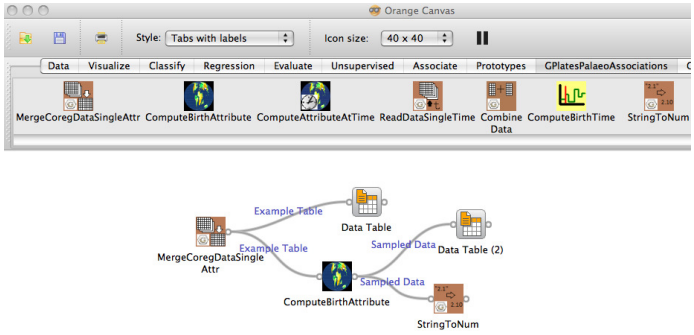

Fig. 5. The Orange visual-programming environment, with GPlates plugin components shown, and a simple example workflow.

of partitioning crustal regions that were influenced by the plume interaction. This can have significant implications for understanding the nature of geological processes, including ore-deposit prediction. The following datasets were combined:

- The North American Volcanic and Intrusive Rock Database (NAVDAT) [8], consisting of aged geochemical samples, with the data filtered to include samples under 60 Ma, within the greater vicinity of Yellowstone.
- Global present-day hotspot dataset developed by the EarthByte group.
- Earthbyte global plate model [4] to define the motion of the North American plate over time.

These datasets are depicted in Figure 6, reconstructed back 22 Ma. The data-analysis objective of this study is firstly to

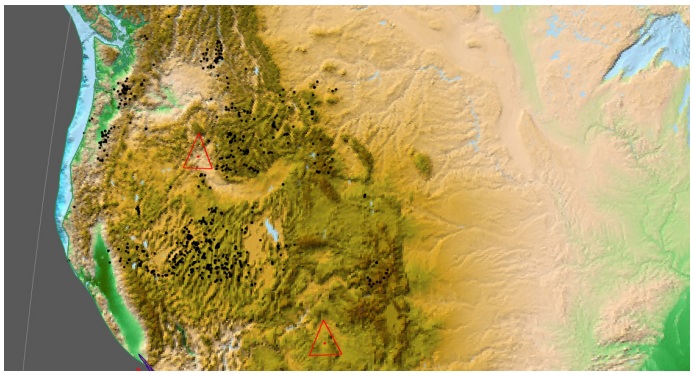

Fig. 6. Case study datasets as depicted in GPlates at 22 Ma in the past, overlaid on the ETOPO1 global relief model [9]. Red triangles depict hotspots (with the upper-right point corresponding to Yellowstone), and black points correspond to age-coded geochemistry.

determine the proximity between the NAVDAT dataset and the yellowstone hotspot at the "birth" age (i.e. the age at which the rocks formed) of each rock sample. Rock samples originating within close proximity of the hotspot are considered to be associated with it, allowing the NAVDAT dataset to be partitioned into two parts, i.e. samples associated with the hotspot, and the complement. The subsequent geochemical populations are then studied to investigate chemical compositions, allowing the nature of the plume-lithosphere to be analysed independently of the other sources of magmatism in the region.

The GPlates coregistration tool is used to specify the relationships of interest. The NAVDAT dataset is chosen as the seed dataset, and the hotspots as the reference. The distance and name pertaining to the nearest hotspot are then defined, specified as a relational association type. The geochemistry metadata of the seed dataset is defined as part of the coregistration. Simulation parameters are then configured to perform the coregistration between 60 Ma in the past to present-day, with 1 Ma time intervals. The resultant coregistration results are then imported into the data-mining tool, with the work-flow design depicted in Figure 7. In this workflow, the hotspot distances at the time of birth are computed, as well as the name of the nearest hotspot. Once these calculations have been performed, the results are appended to the geochemistry metadata. The entire dataset is then filtered by the distance to the nearest hotspot, and samples closest to Yellowstone. The partitioned geochemistries are then analysed, and the target points plotted in GPlates.

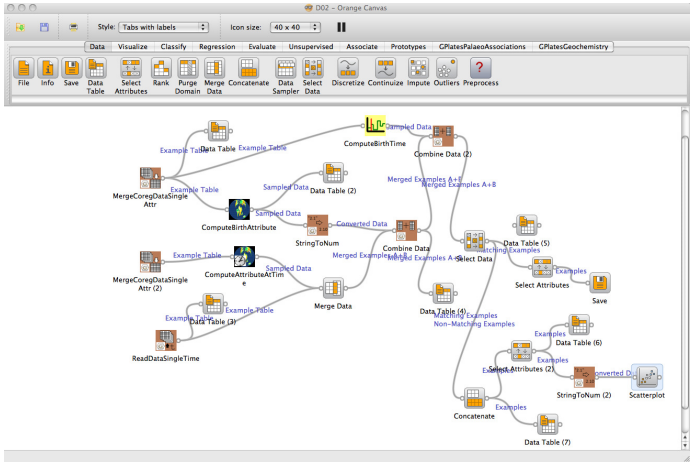

Fig. 7. Visual workflow for the NAVDAT case study, involving extraction of time-varying associations, filtering of data close to the Yellowstone plume, and investigation of the resultant partitioned geochemistries.

Partitioning of the NAVDAT dataset into observations associated with the hotspot and the complement is depicted in Figure 8. Varying the proximity threshold, and plotting the subsequent partitioning spatially (i.e. in the GPlates GIS tool) aids in selecting a suitable threshold, and allows for visual inspection of the hotspot trail as it interacted with the lithosphere. The partitioning then allows for a follow-up analysis using the geochemistry metadata. Two typical geochemical elements studied when considering plume-related magmatism involve comparing the elements Yb and La. This analysis is shown in Figure 9 for the two partitioned populations (using a 260km distance threshold), with some separation of the populations evident. This can be used for further, more in-depth analysis. This case study demonstrates how a complex spatio-temporal analysis involving a moving spatial reference frame can be decomposed and solved in a relatively simple fashion using the proposed framework.

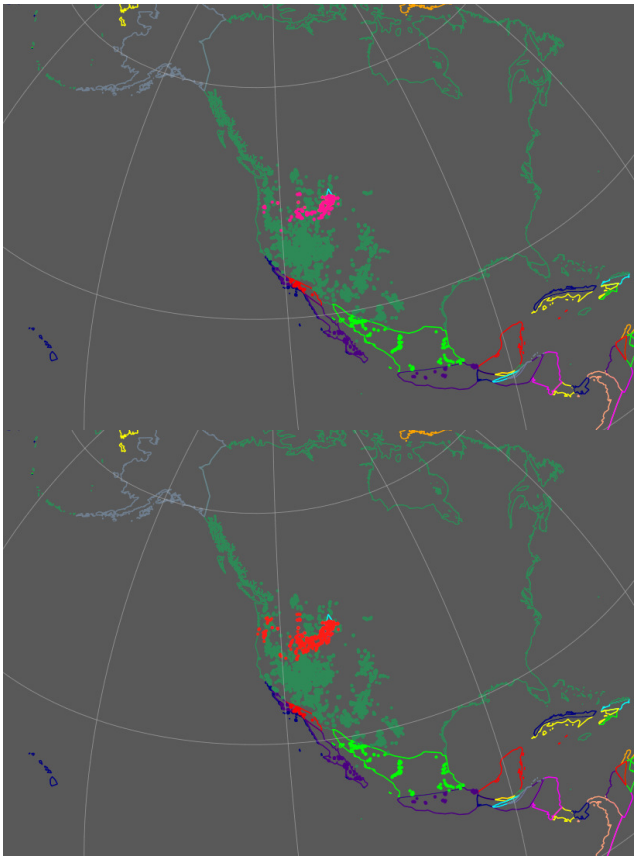

Fig. 8. Partitioning of the NAVDAT dataset subsequent to data-mining, demonstrating the importance of feedback between the data-mining and GIS components. The top figure depicts the partitioned dataset in pink, with the remainder in green, having used a 260km distance threshold. A 360 km threshold partitioning is depicted in the lower plot, with the red points depicting a now larger sphere of influence.

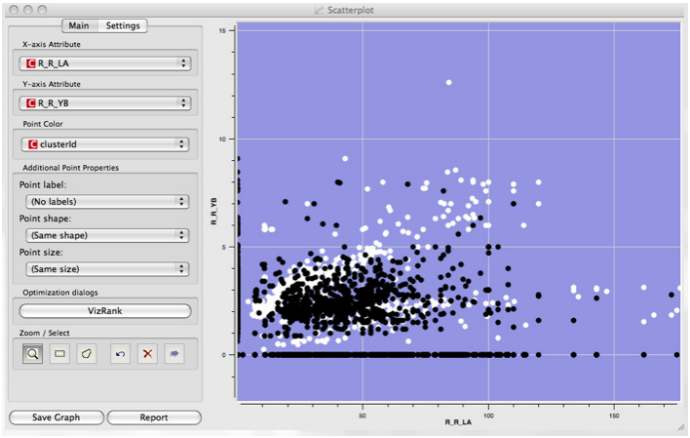

Fig. 9. Geochemical analysis for the partitioned NAVDAT dataset regarding the Yb and La elements. Good separation of the populations suggests the data coregistration has successfully partitioned the lithospheric segments corresponding to the path of the plume.

VII. Conclusion

This paper presented a spatio-temporal knowledge-discovery platform design with application to the Earth Sciences, manifested as an extension to the GPlates plate-reconstruction GIS tool. The importance of incorporating the many modes of data (which are becoming increasingly available) across the geosciences leads to the ability to tackle more complex and pertinent scientific questions. The simultaneous assessment of these large and diverse datasets through space and time poses new challenges, and calls for novel approaches in interacting with and understanding the data. In this paper, we discussed these challenges, and how they led to the development of a corresponding software design and implementation. The design consists of two primary components, namely a *coregistration tool* for recursively defining explicit quantitative relationships between datasets in a flexible fashion, and a *data-mining environment* for subsequent assessment of the derived datasets using powerful data manipulation and statistical analysis approaches. The ability to develop specialised plugins and abstract workflow definition complexity via a visual programming environment are two essential aspects incorporated into the design. The overall solution is demonstrated via a complex spatio-temporal case study. It is anticipated that this knowledge-discovery environment will lead to significant scientific advances, and be adopted by a broad community of users due to it's flexible, open design.

Acknowledgment

The authors acknowledge support for this research via the by ARC grant number FL0992245.

References

[1] EarthByte, *GPlates - a cross-platform, open-source real-time spatio-temporal GIS*, The University of Sydney, CalTech, The Norwegian Geological Survey, 2011. [Online]. Available: http://www.gplates.org

[2] J. M. Whittaker, R. D. Mueller, G. Leitchenchov, H. Stagg, M. Sdrolias, C. Gaina, and A. Goncharov, "Major australian-antarctic plate reorganization at hawaiian-emperor bend time," *Science*, vol. 318, pp. 83–86, 2007.

[3] T. Torsvik, R. Mueller, R. Van der Voo, B. Steinberger, and C. Gaina, "Global plate motion frames: Toward a unified model," *Reviews in Geophysics*, vol. 34, pp. 1–44, 2008.

[4] R. D. Mueller, M. Sdrolias, C. Gaina, and W. Roest, "Age, spreading rates, and spreading asymmetry of the world's ocean crust," *Geochemistry Geophysics Geosystems-G3*, vol. 9, 2008.

[5] *Geoscience Markup Language (GeoSciML)*, Commission for the Management and Application of Geoscience Information (CGI), 2011. [Online]. Available: http://www.geosciml.org/geosciml/2.0/doc/

[6] J. Demsar, B. Zupan, G. Leban, and T.Curk, "Orange: From experimental machine learning to interactive data mining," *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2004.

[7] B. B. Hanan, J. W. Shervais, and S. K. Vetter, "Yellowstone plumevcontinental lithosphere interaction beneath the snake river plain," *Geology*, vol. 36, no. 1, pp. 51–54, January, 2008. [Online]. Available: http://geology.gsapubs.org/content/36/1/51.abstract

[8] J. Walker, T. Bowers, R. Black, A. Glazner, G. Lang Farmer, and R. Carlson, "A geochemical database for western north american volcanic and intrusive rocks (navdat)," *Geoinformatics: Data to Knowledge, Book Section*, no. 397, pp. 61–71, 2006.

[9] C. Amante and B. Eakins, "Etopo1 1 arc-minute global relief model: Procedures, data sources and analysis," *NOAA Technical Memorandum NESDIS NGDC-24*, p. 19, March 2009.