Data mining Tutorial I: Identifying depositional environments for Australian age-coded mineral deposits

Objective

Demonstrate the combining of two relatively large time-dependent datasets. The analysis answers the question: What is the association between mineralisations and the depositional environment? This example shows that complex sub-analyses, such as investigating this as a function of the commodity type is trivial when undertaken in a quantitative fashion.

Dataset descriptions and pre-processing

- Palaeogeographic Atlas of Australia: time-dependent summarisation of sedimentological data based on several datasets, between ~ 550 Ma to present dat. AGSO/Geoscience Australia, Langford, R.P. Wilford, G.E. Truswell, E.M. Totterdell, J.M. Yeung, M. Isem, A.R. Yeates, A.N. Bradshaw, M. Brakel, A.T. Olissoff, S. Cook, P.J. Strusz, D.L., http://www.ga.gov.au/meta/ANZCW0703003727.html. In the demonstrations, a single ESRI shapefile has been generated to incorporate all data, with age-codings approximated according to the time-instant descriptions. This results in a single time-dependent data structure that is handled fluently in GPlates. For visualisation purposes, a time-dependent raster sequence has been created, annotated with familiar colours.
- OZMIN Mineral Deposits Database: Ewers, G.R., Evans, N., and Hazell, M., (Kilgour, B., compiler). 2002. OZMIN Mineral Deposits Database. [Digital Datasets]. In this demonstration a PLATE format data file has been created, storing the commodity name, and using the extreme categorical age as the age of mineralisation. Note that exploring spatio-temporal associations should consider the age-range uncertainty.

Methodology

1. Loading and visualising data in GPlates
2. Defining the desired associations using the GPlates coregistration tool
3. Compute the palaeo-distances between each rock sample and the closest plume

Step 1: Loading and visualising data in GPlates

Load GPlates from the command line as follows:

```
gplates --data-mining
```

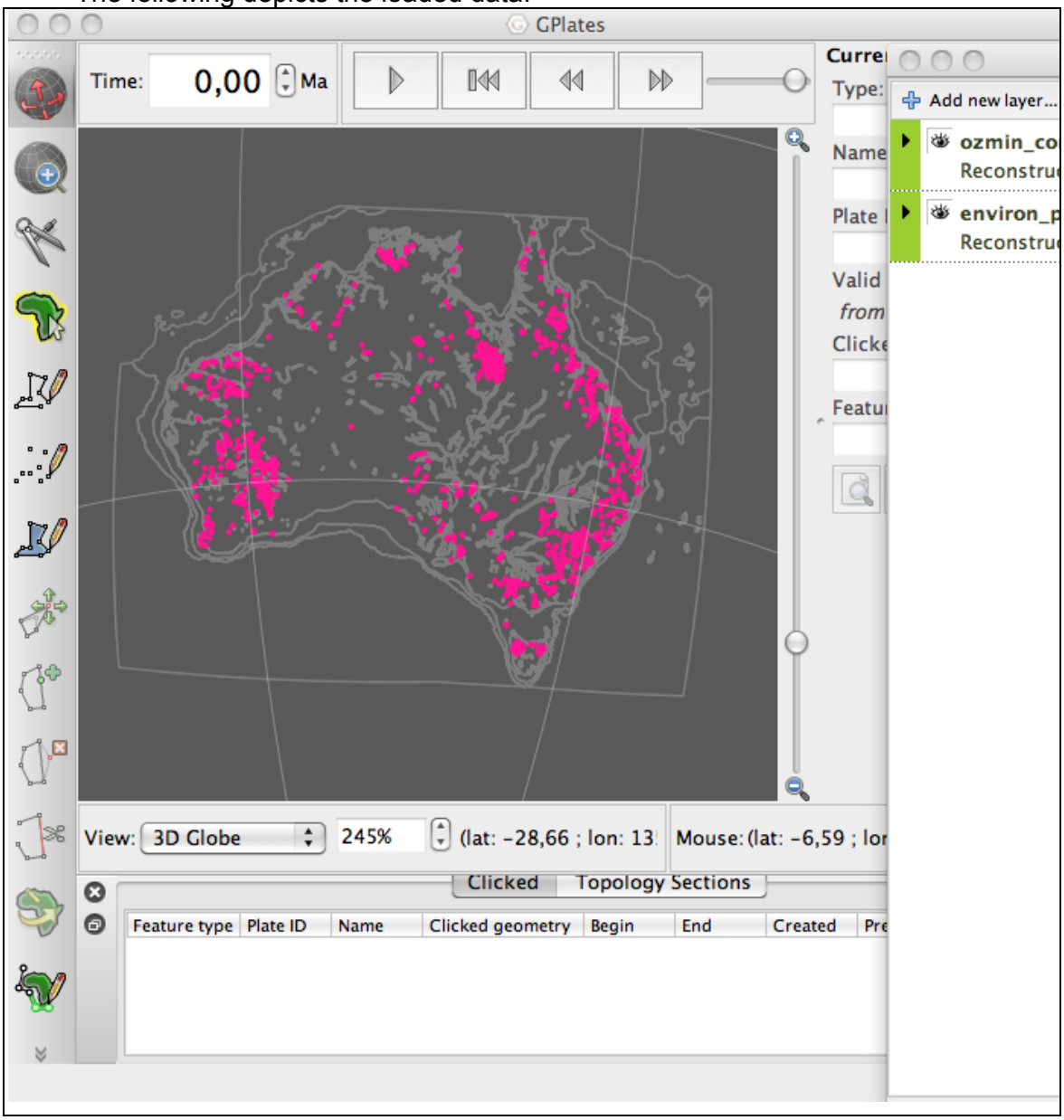
Load features and set properties as follows:

- In GPlates, open the following feature collections: Palaeogeography.shp (called the

Palaeogeography dataset)

- Then open ozmin_commodity.shp (called the *OZMIN* dataset) and configure visualisation settings: set the colouring of the OZMIN dataset to a single colour (it is currently coloured by plate ID)

The following depicts the loaded data:

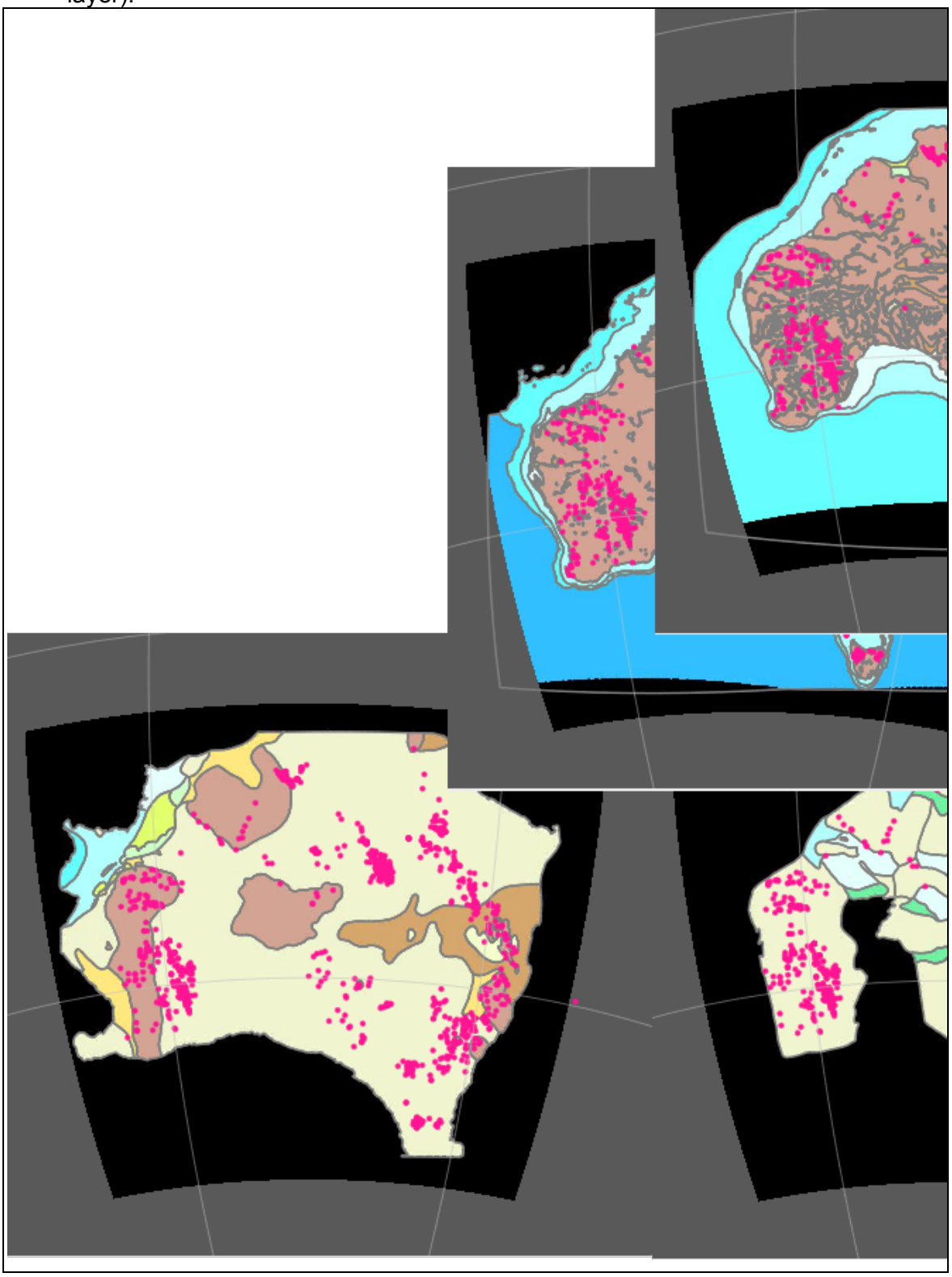


- Open the following time-dependent raster sequence (use "File", "Import Time-Dependent Raster...", and then click on "Add directory..."): timedep_raster_palaeogeog. Specify georeferencing as follows:

```
Top (lat):    -7.0
Bottom (lat): -47.0
Left (long):  109.0
Right (long): 159.0
```

In the layering tool, ensure that the raster layer is dragged to the bottom of the list

so that other data is overlaid (click and drag the coloured region on the left hand side of the layer).



Configure the animation to start at (no earlier than) 540 Ma, and end at present day, with steps of

10 Ma.

Step 2: Defining the desired associations using the GPlates coregistration tool

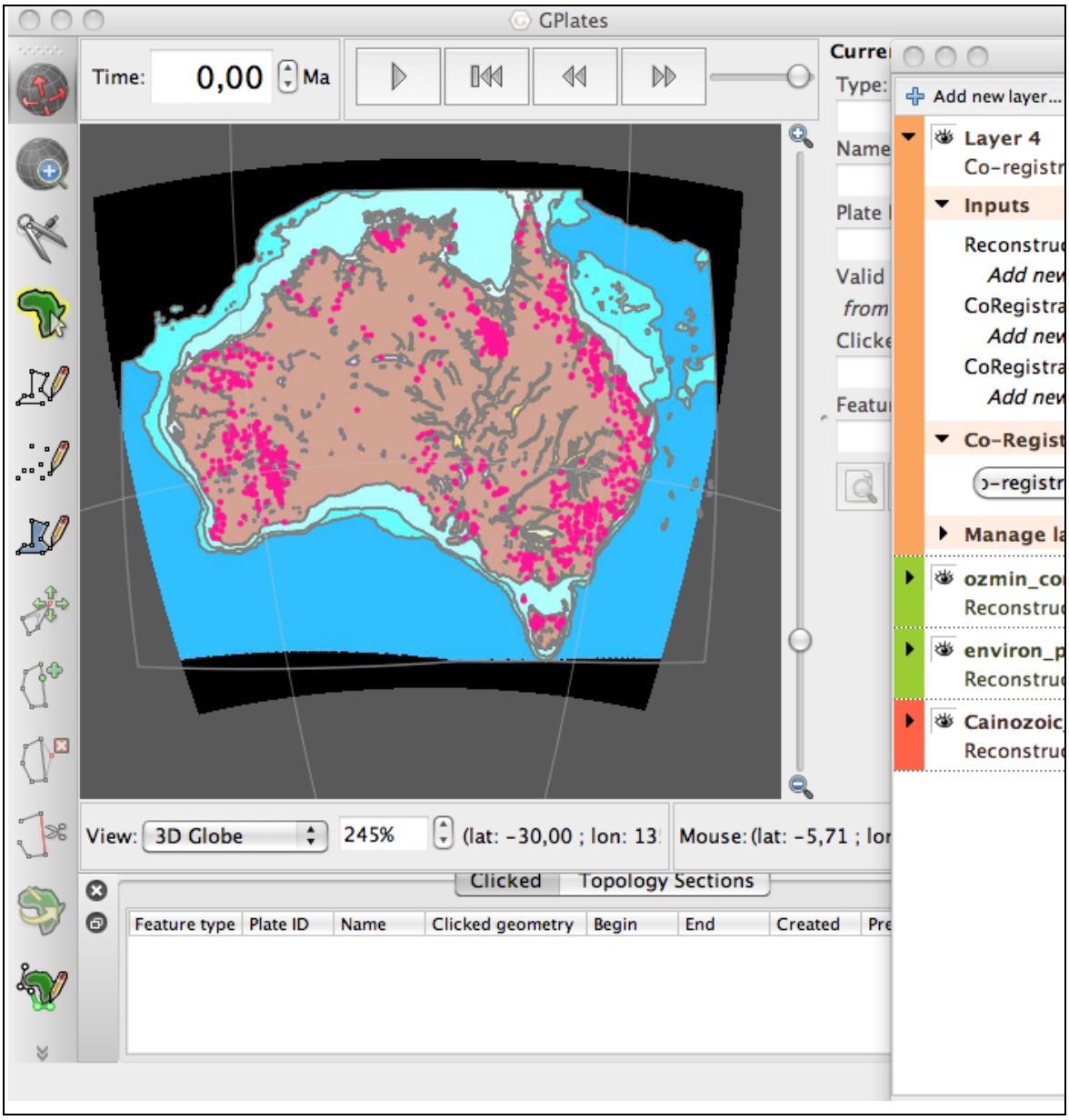
In this analysis, we wish to investigate the sedimentological environment in which an ore-deposit has formed. This is to be determined by computing the environment in which an ore deposit occurs for all times, and then subsequently determining the property at the "birth-age". Later we discuss how we can look at relationships as a function of the commodity type. Performing this analysis entails investigating two data properties, namely the "Environ" property of the Palaeogeography dataset, and the "name" parameter of the OZMIN dataset, representing the commodity name. Two steps are involved, namely defining the data association, and analysing the data via the data mining tool (next step).

1. Configure GPlates coregistration tool
2. Export coregistration results

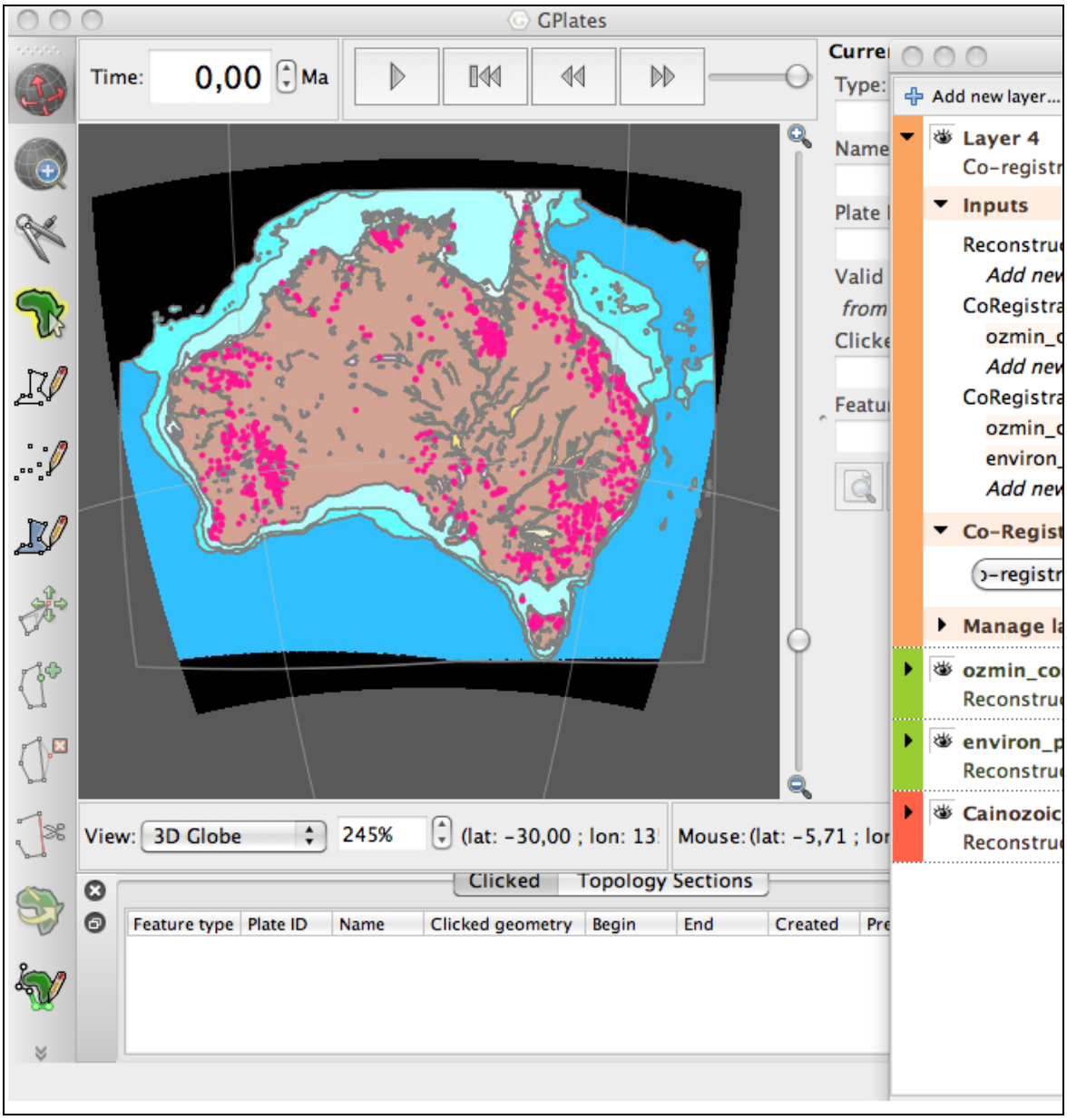
Part 1: Configure GPlates coregistration tool

Data coregistration is performed via the Layers dialog in GPlates (show via the "Window" menu item if absent). The following steps define the required data association:

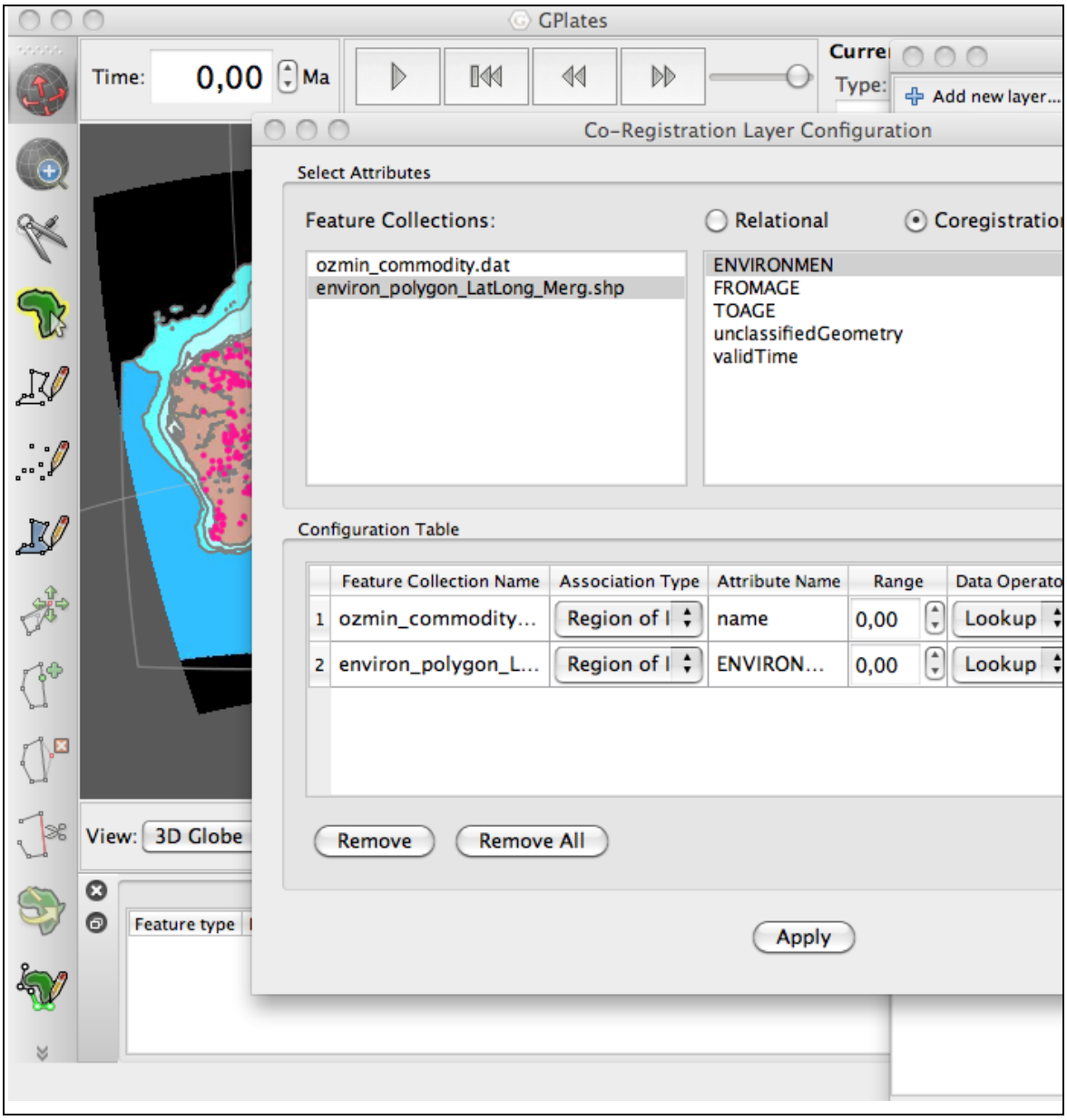
1. Define a "co-registration" layer from the "Add new layer..." button on the Layers dialog (top). A new layer will thus be shown on the layer dialog.
   Select the new layer, and expose inner parameters by selecting the triangle button on the left side of the layer. The following depicts exposed parameters:



2. First a coregistration seed channel is defined, which is essentially the independent variable in the analysis. Coregistration inputs are then dependent variables, depicting relations and associations with respect to the seeds. In the layer tool, select the OZMIN dataset as the seed. Both the OZMIN and palaeogeography datasets are then to be selected as coregistration input channels; the OZMIN dataset is included here so that the commodity type can be included in the analysis. The layer parameters should look as follows:

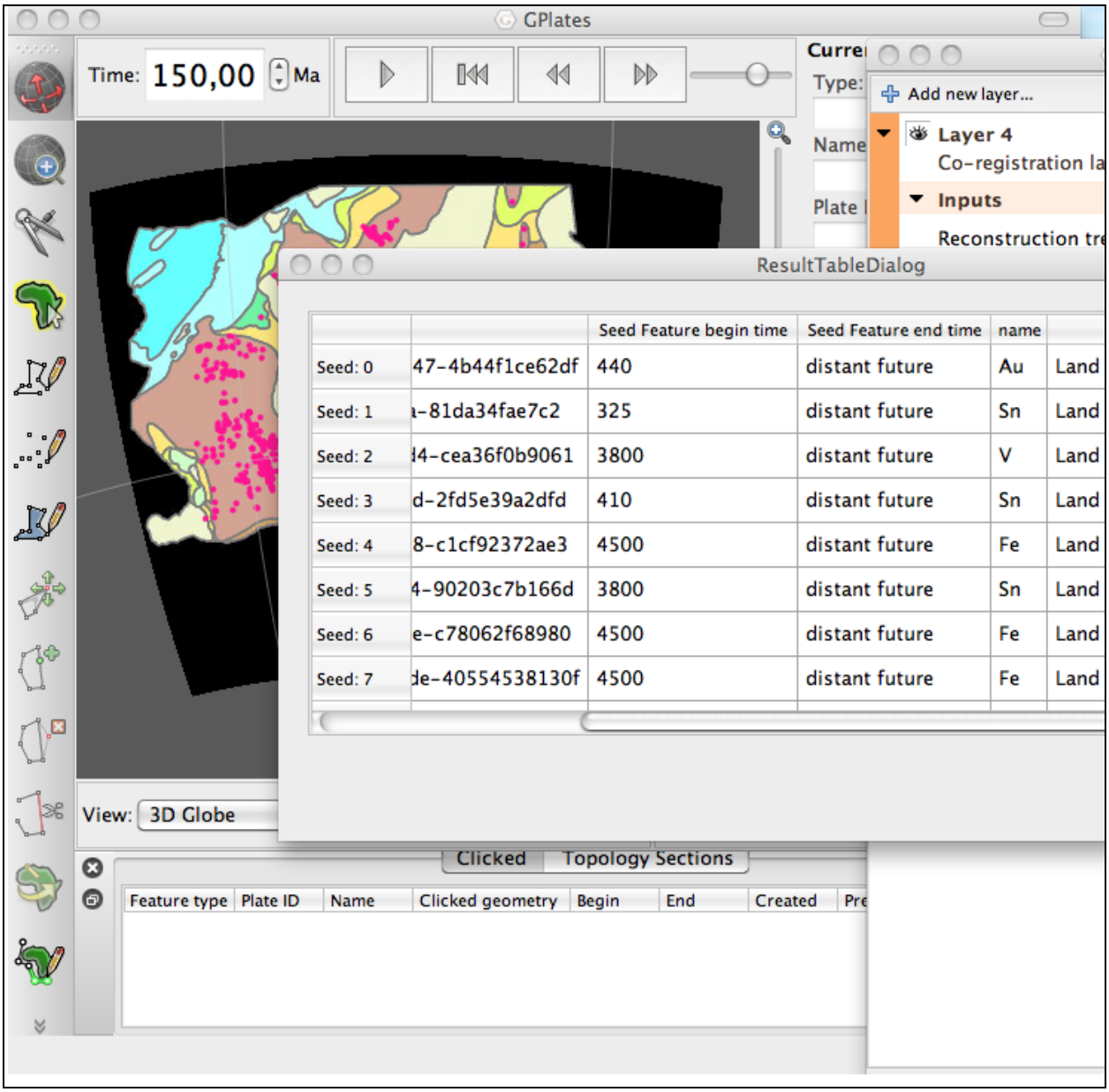


3. In the next step, the data association is configured by selecting the "Co-registration Configuration" on the layer dialog. First the commodity name is added by selecting the commodity dataset, selecting the "Coregistration" option (as opposed to "Relational"), and then selecting the property *name*. Selecting the "Add" button will then add this to the configuration table. The palaeogeography dataset should then be selected similarly, and the *ENVIRONMEN* attribute selected, following by adding to the configuration table. The coregistration configuration should look as follows:

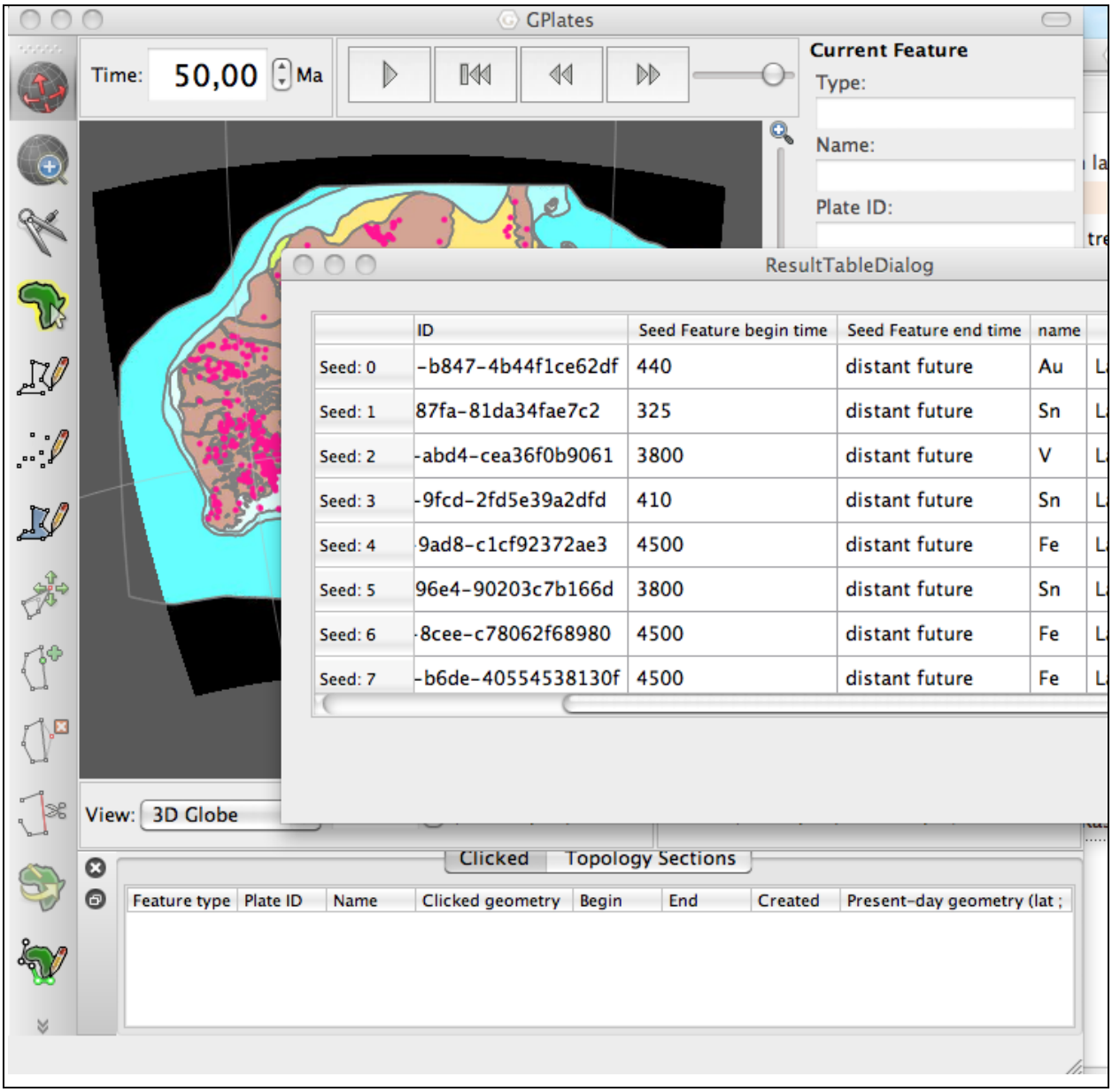


4. Clicking "Apply" will then add these computations to GPlates, with co-registration computations performed at each time instant.

Now that the coregistration has been configured, we can visually inspect how the defined associations change as we vary the GPlates time slider. The coregistration values can be viewed by choosing a desired time, and selecting "View Result" on the layer dialog corresponding to the coregistration layer. (The first three attributes are always automatically set to be the GPlates-ID and begin and end time of the seed, respectively.) At a time of 150 Ma, the following results are obtained:



Similarly, the following depicts results at 50Ma:



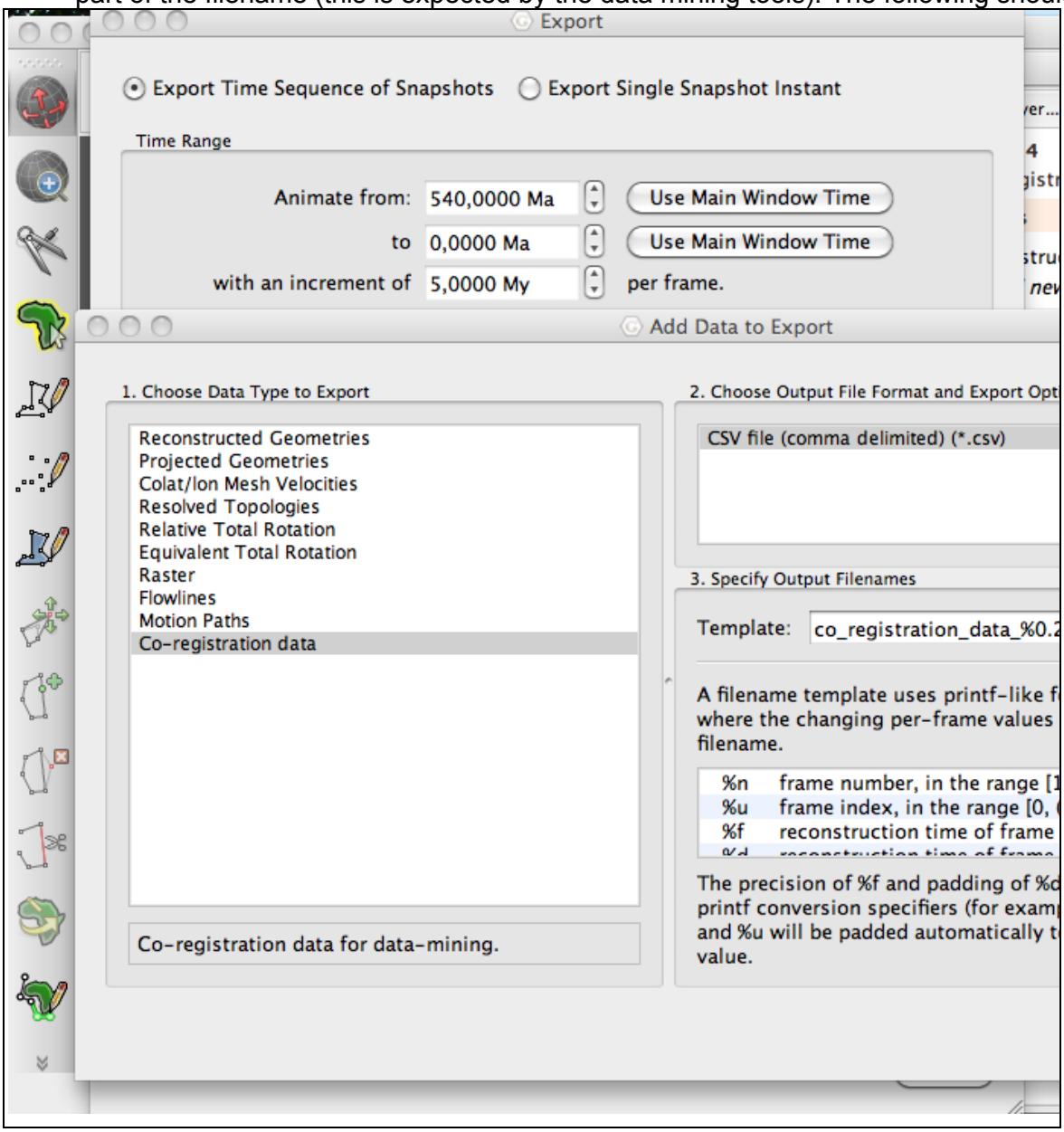
Note that seed points not yet defined at a particular time instant do not appear on the output (thus different rows for different times). Additionally, due to GPlates' method of determining the associations for all times, the result table can look different anyway.

Part 2: Export coregistration results

The final coregistration step is to export the coregistration results to an output directory. This is an interim step while GPlates is coherently integrated with the data mining suite. The coregistration export will output comma-separated-value (.csv) files for each time instant defined by the "Configure animation" option in the GPlates menu. For this analysis, the following steps should be followed:

1. In GPlates, the export facility is to be used for the coregistration result output. This is initiated via the menu ("Reconstruction" followed by by "Export...")
   The Add data button should be selected, followed by selection of the "Co-registration data"

export type, and highlighting of the CSV file output format option. Please leave other parameters at their default values, and ensure that the sub-string "co_registration" forms part of the filename (this is expected by the data mining tools). The following should result:



2.
3. Once the export has been selected, the target directory should be defined. Please create a new empty directory and define this to be the export target directory. The results are then generated by selecting the "Begin Animation" button.

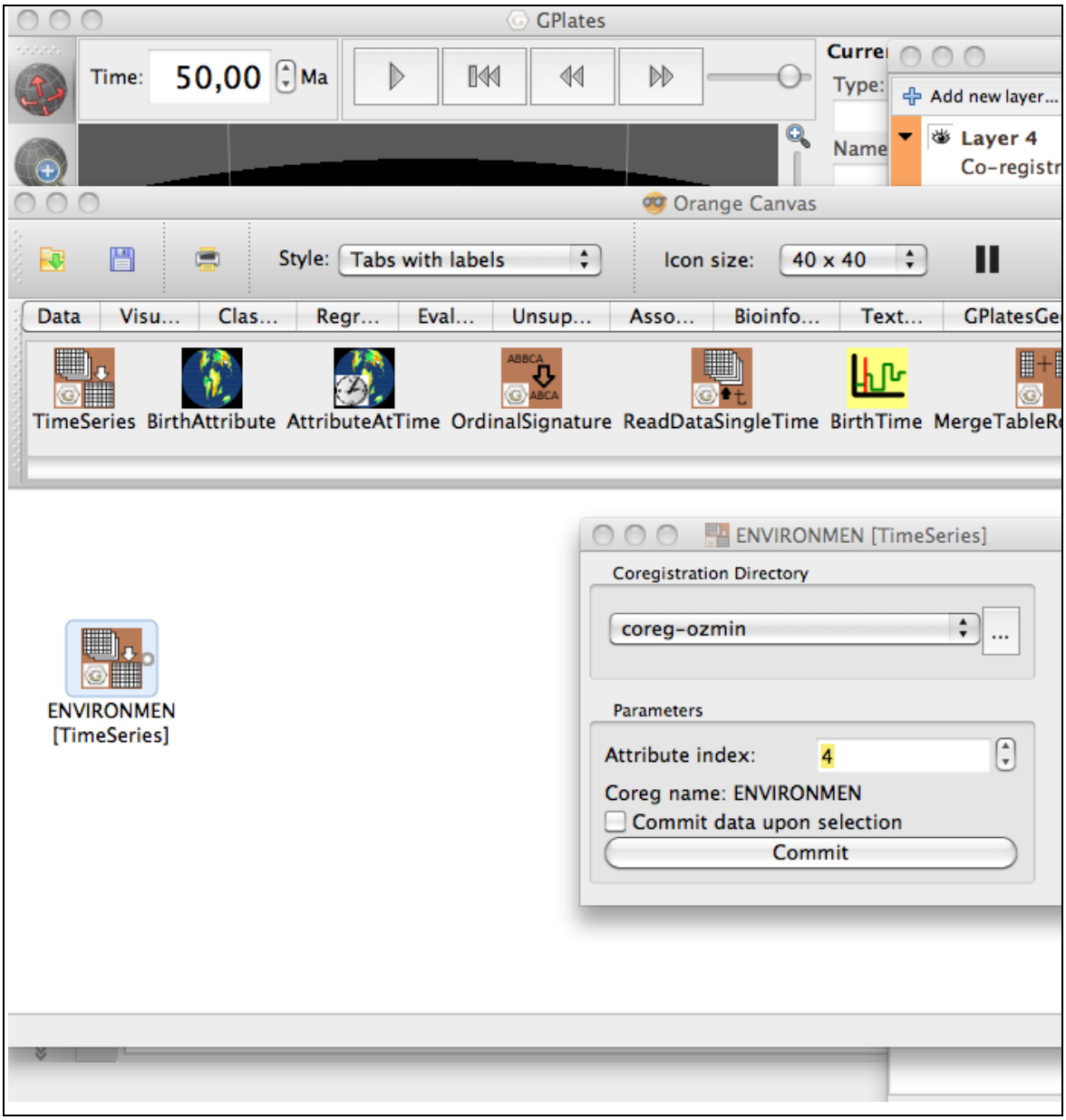
Step 3: Configuring an Orange schematic design for the analysis

Once the coregistration analysis results have been exported via GPlates, the next step is to perform data mining and analysis of the time-dependent results. The *Orange* data mining tool is being used for this, with GPlates specific-plugins underway. A visual-programming environment

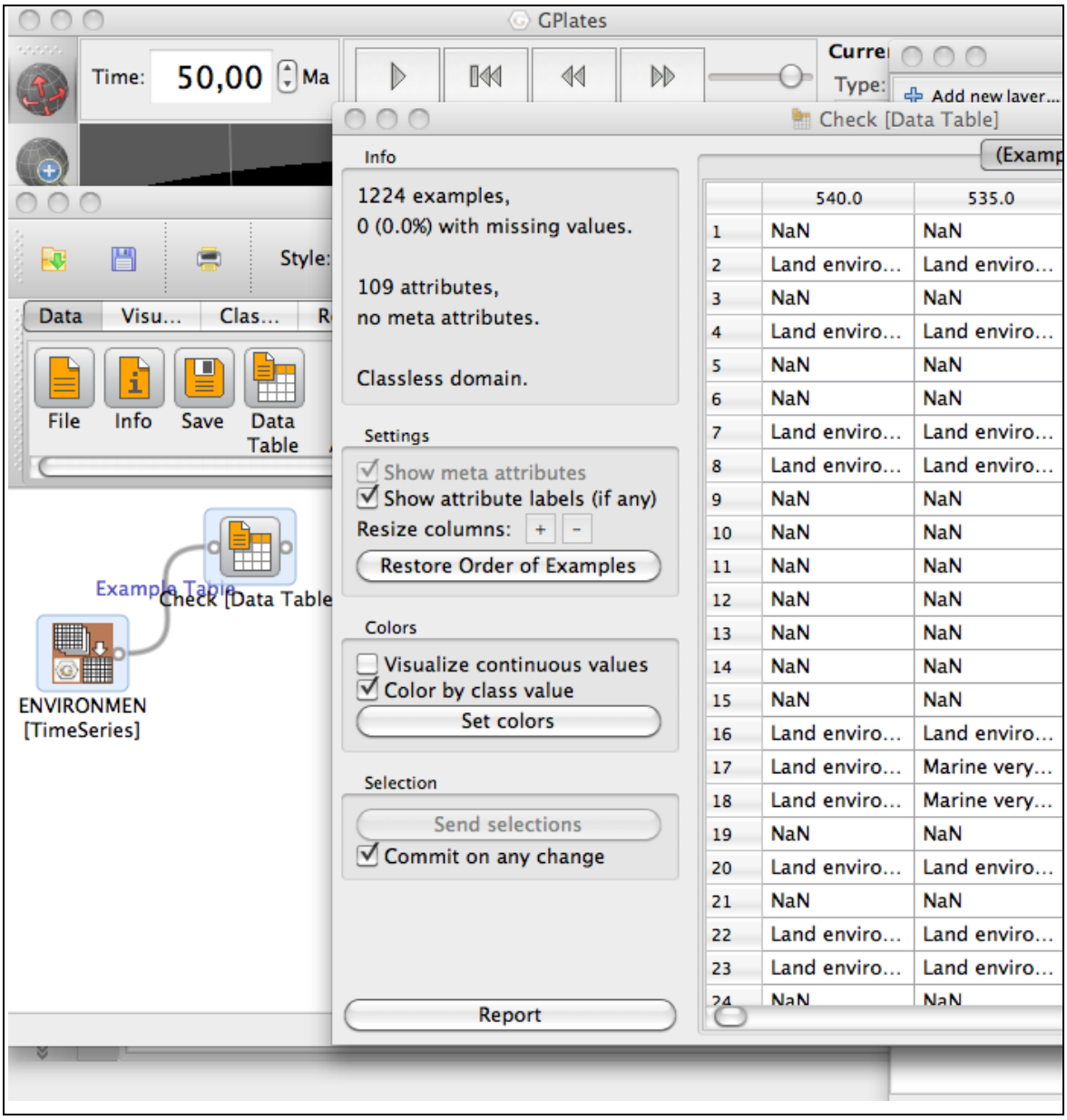
is being used to abstract analysis complexity and improve flexibility. Current developments will also see Orange directly integrated with GPlates, but at this preliminary stage Orange has to be started as a separate program.

Once Orange has been started, various data mining and analysis tools, called *Widgets* can be selected from collections. A *GPlatesPalaeoAssociations* collection has been developed for the analyses required here. The following steps should be undertaken for this analysis:

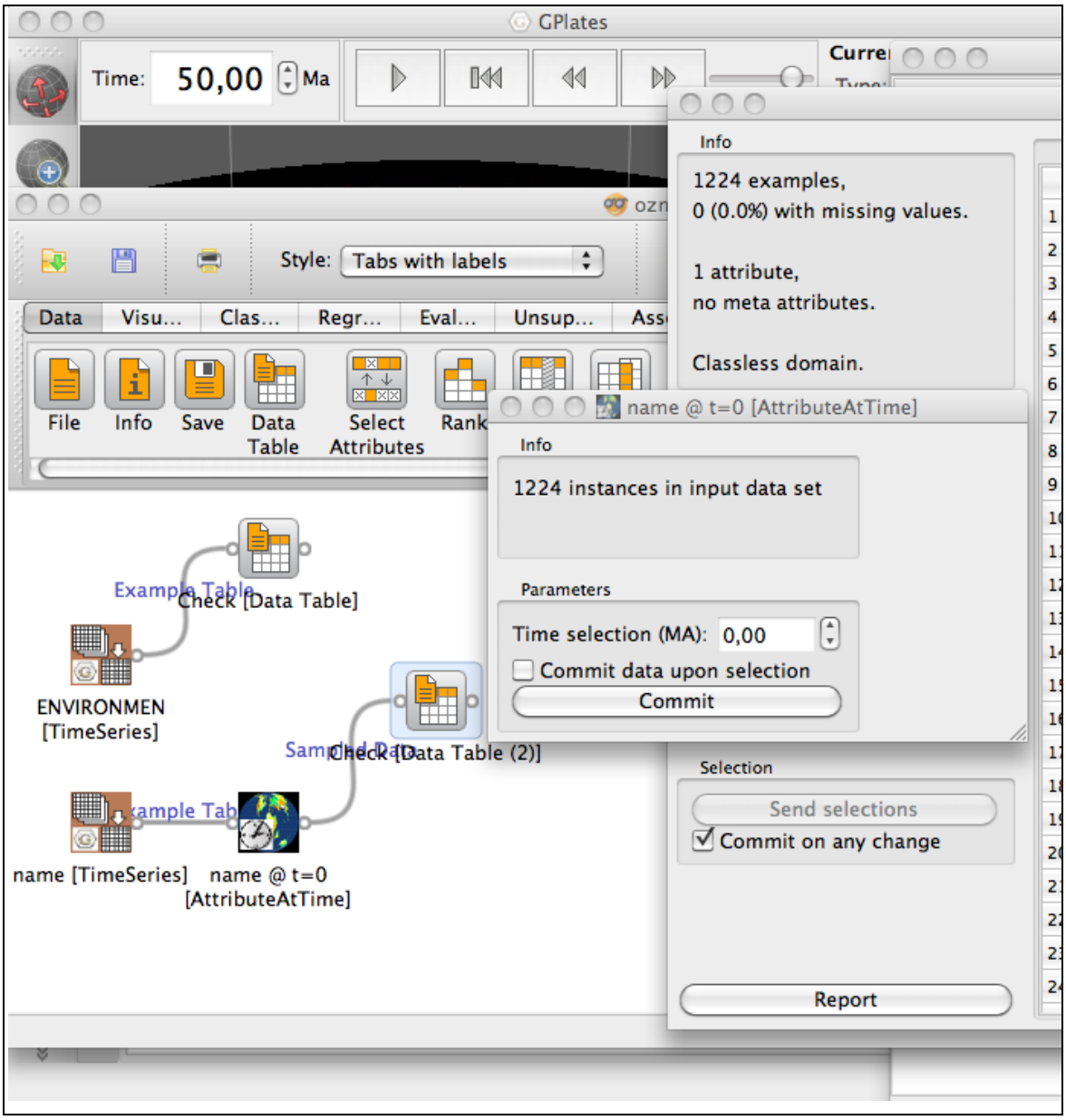
In a first step, the "TimeSeries" widget from the *GPlatesPalaeoAssociations* collection should be dragged onto an empty canvas. This widget analyses all coregistration outputs across time, and returns a homogenous table for the chosen coregistration attribute corresponding to each seed geometry for all times. Thus this widget forms a time-series data structure, ready for subsequent analysis. After double-clicking on the widget to expose parameters, the directory which the coregistration results were exported to should be chosen. The "attribute" parameter then allows for selection of the coregistration analysis of interest (scrolling through attribute indices will result in the associated attribute names of the cregistration being shown). For this exercise, the fourth attribute index should be chosen, pertaining to the *ENVIRONMEN* attribute. (As already mentioned above indices 0, 1 and 2 are always reserved for the GPlates-ID and begin and end time of the seed in question.) The following depicts the first step (it is a good habit to rename widgets appropriately for better illustration):



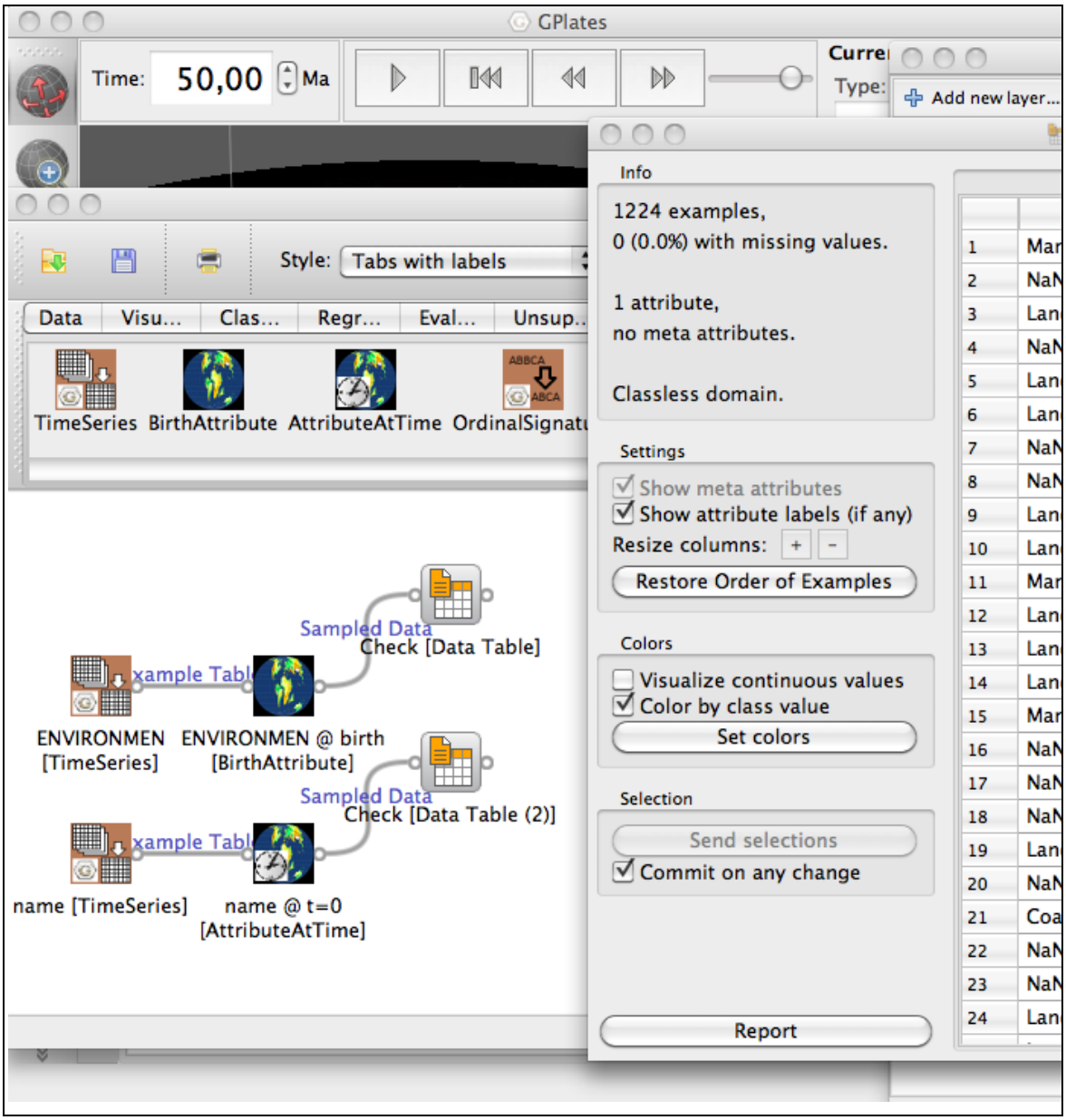
Dragging a "Data Table" widget from the *Data* collection, and connecting it to the output of the "TimeSeries" widget allows the data structure to be studied in time-series format (tables denoted with "Check" are just for checking intermediate results):



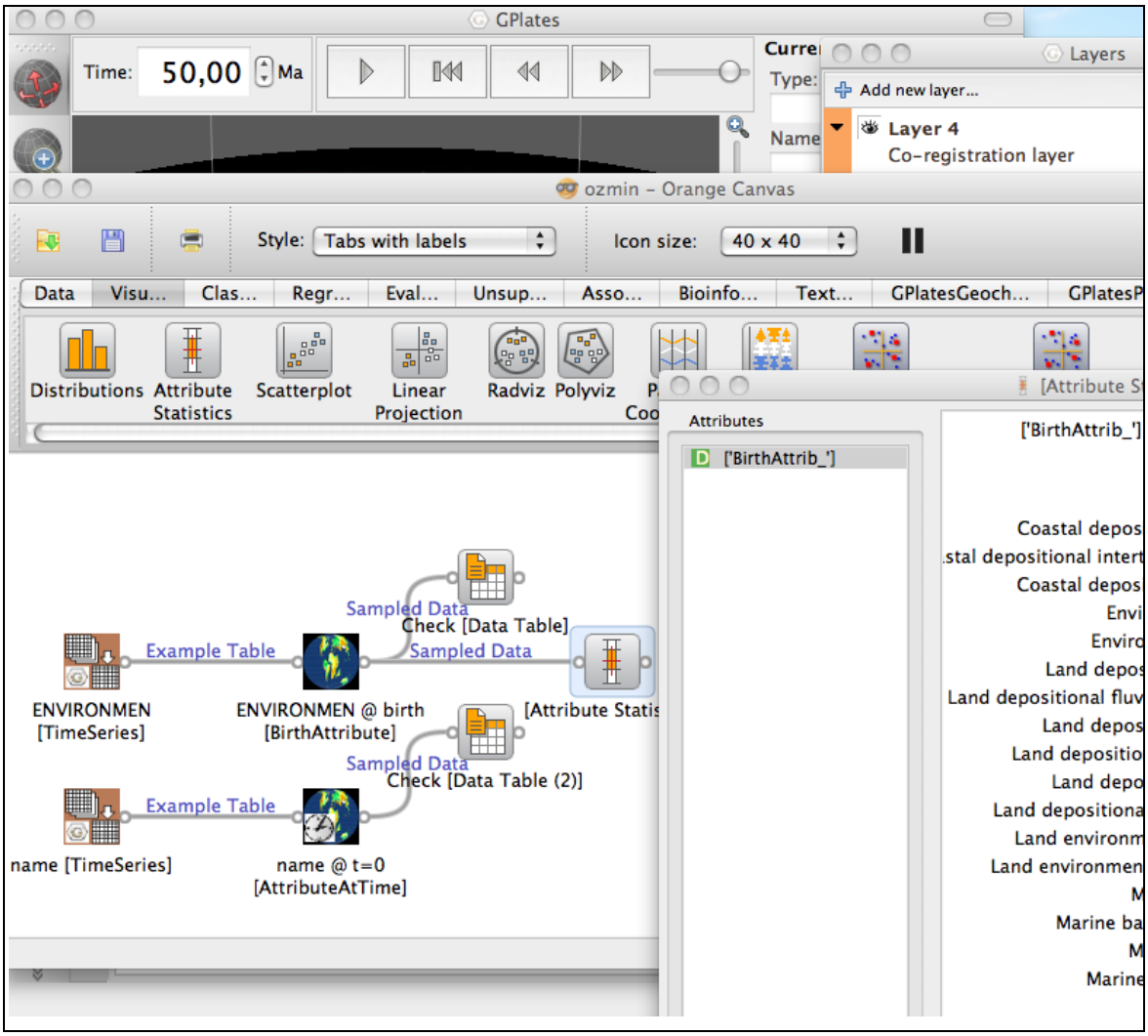
- Next a second "TimeSeries" widget should be dragged onto the canvas, and the *name* attribute selected. Since this attribute does not change over time, the present day attribute represents the commodity type over all times, extracted by attaching the "AttributeAtTime" widget to the output, set to extract results at present day. A single vector of results is obtained, as depicted below:



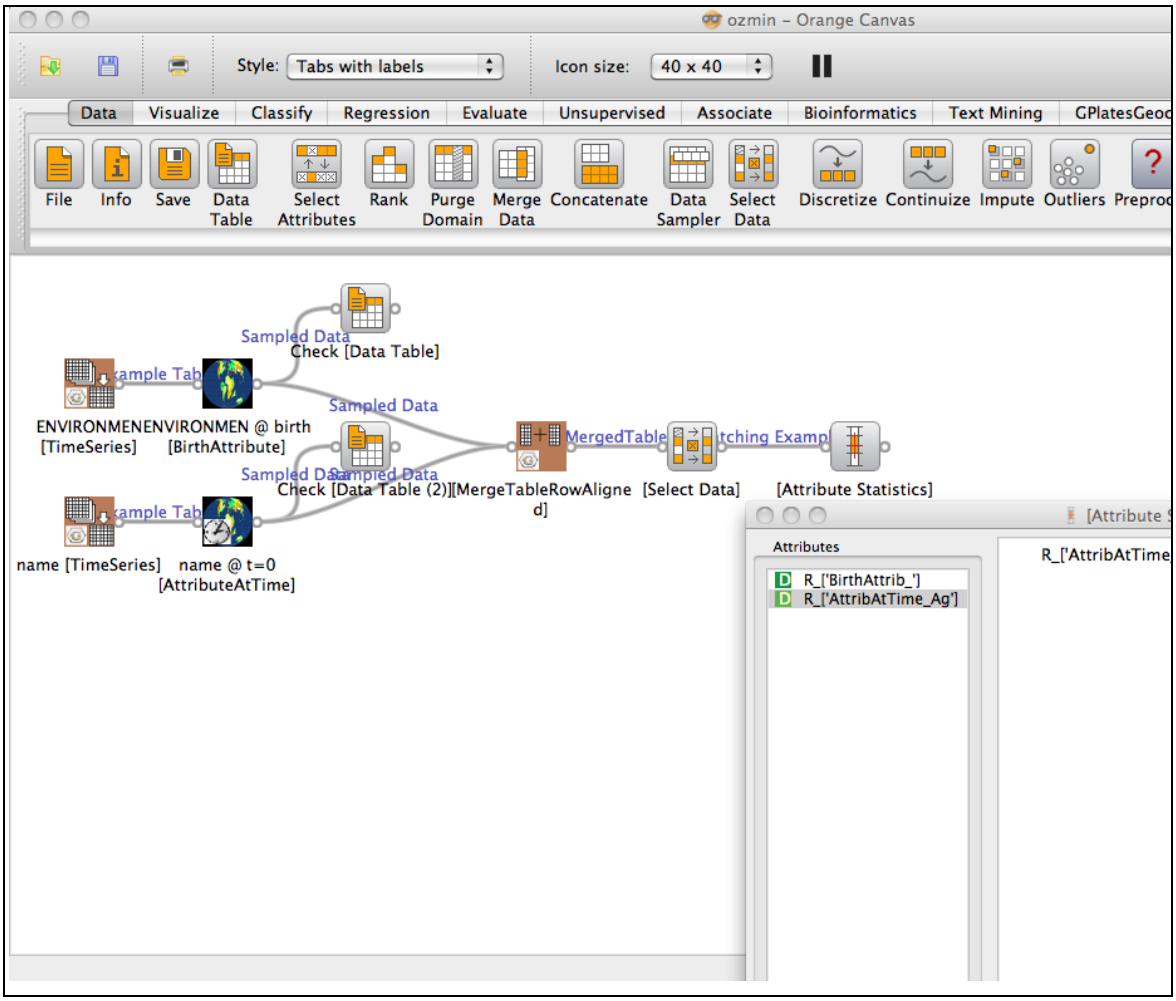
- Next the "BirthAttribute" widget from the *GPlatesPalaeoAssociations* collection should be connected to the output of the first "TimeSeries" widget. This widget is useful for establishing palaeo-relationships at the time of formation. This is achieved by detecting the point in time at which the seed becomes valid, and then recording the selected attribute at that time:



Note that many of the output results are shown as "NaN", i.e. invalid. This is because many of the seed features formed before 540 Ma, and relevant palaeo-geographic information is not available. The "Attribute Statistics" widget from the *Visualise* collection is useful to look at the overall statistics of the results:

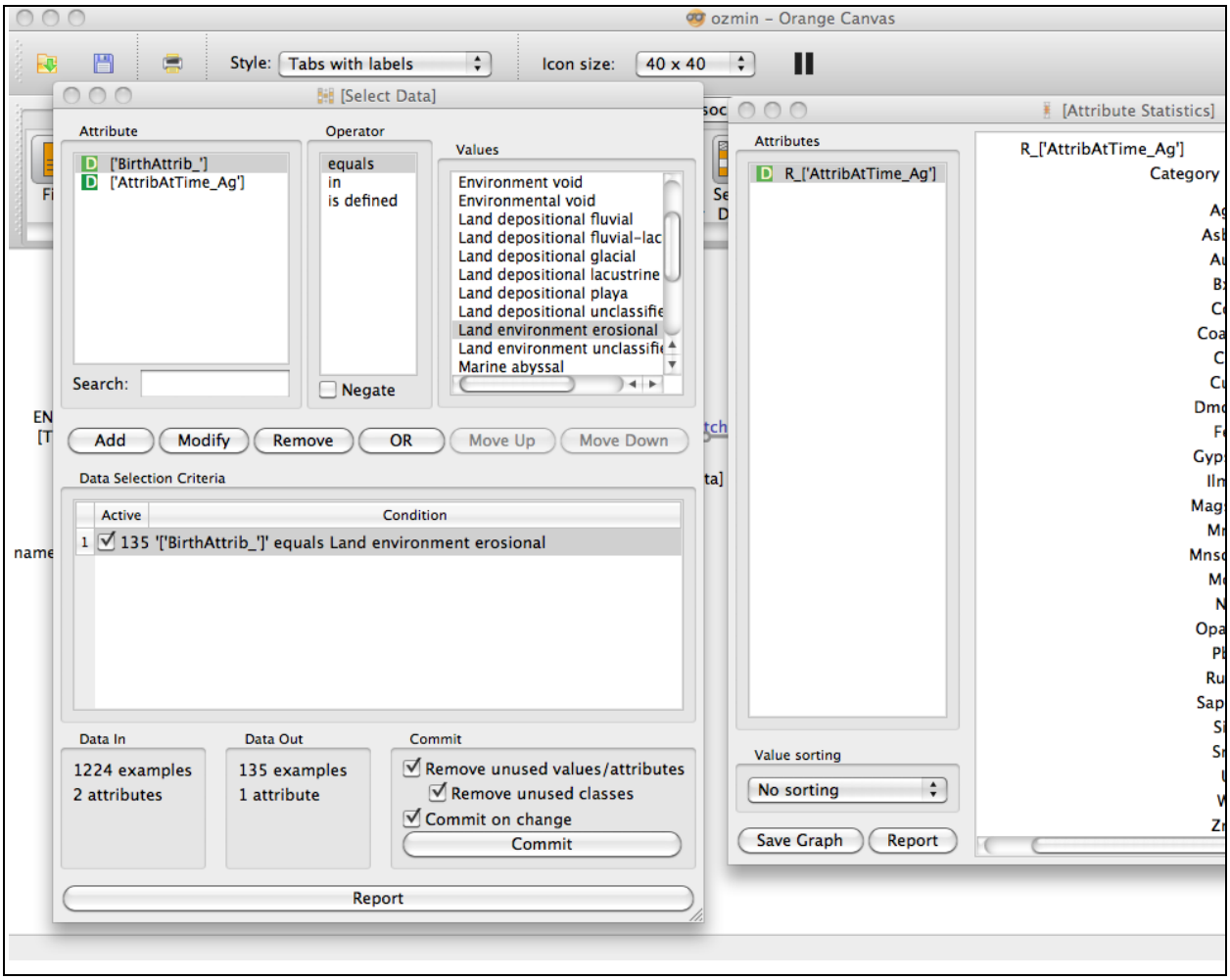


- It can be seen that the 61.6 percent of the data is invalid (mineralisations occurred before 540 Ma), with the bulk of the remaining data falling within 6 other categories e.g. *Land environment erosional*.
- In order to study the relationships between the birth environment and different commodity types, the "CombineData" widget from the *GPlatesPalaeoAssociations* plugin joins the two respective datasets together into a single data structure. Caveat: when connecting inputs to this widget, the widget should be opened, and the variables from each source selected manually.
  To complete the analysis, the "Select Data" widget should be selected from the *Data* plugin, followed by the "Attribute Statistics" attribute from the *Visualise* plugin. The "Select Data" widget can be used interactively to filter results according to either the environment or commodity types. The following schematic completes the exercise:

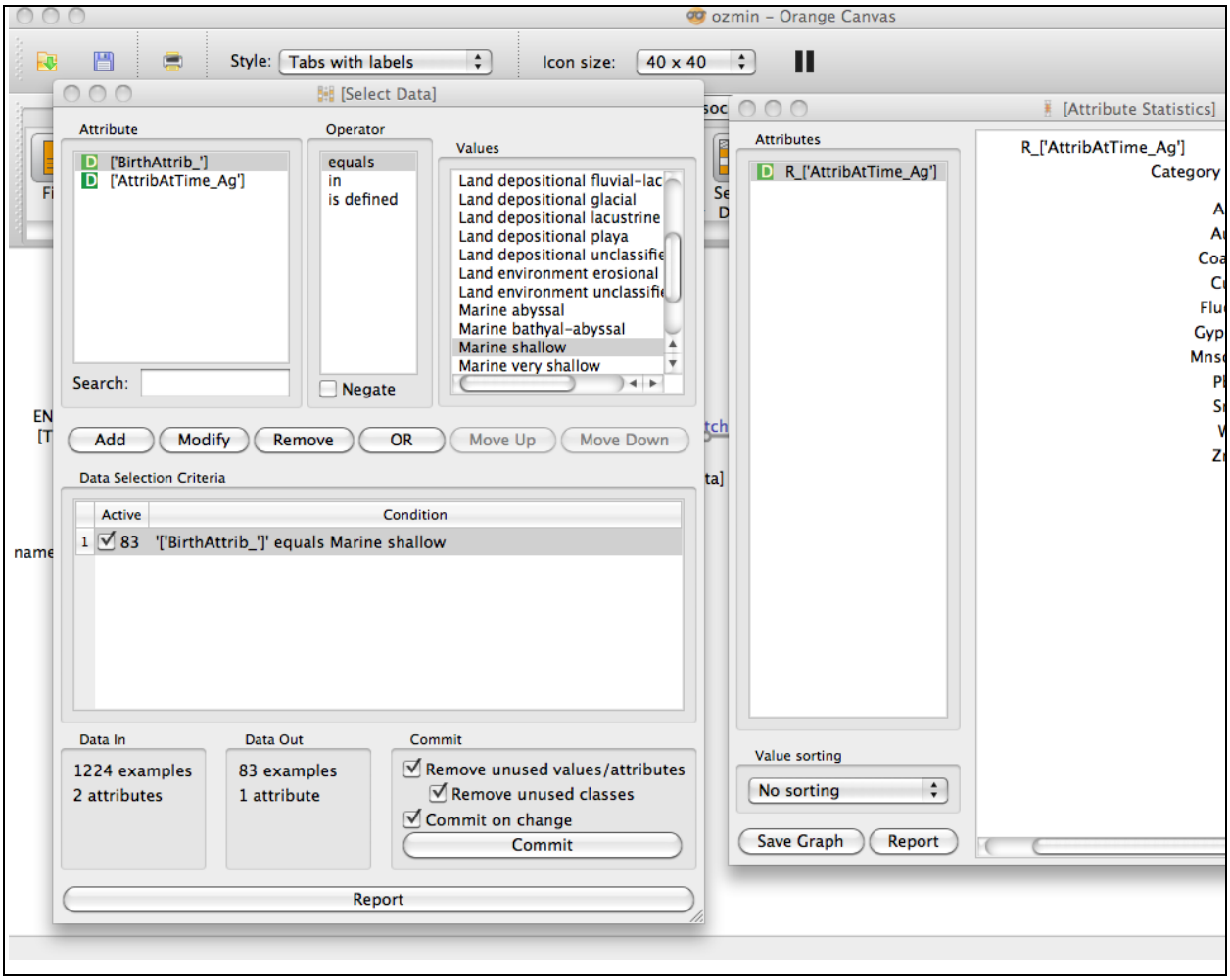


-

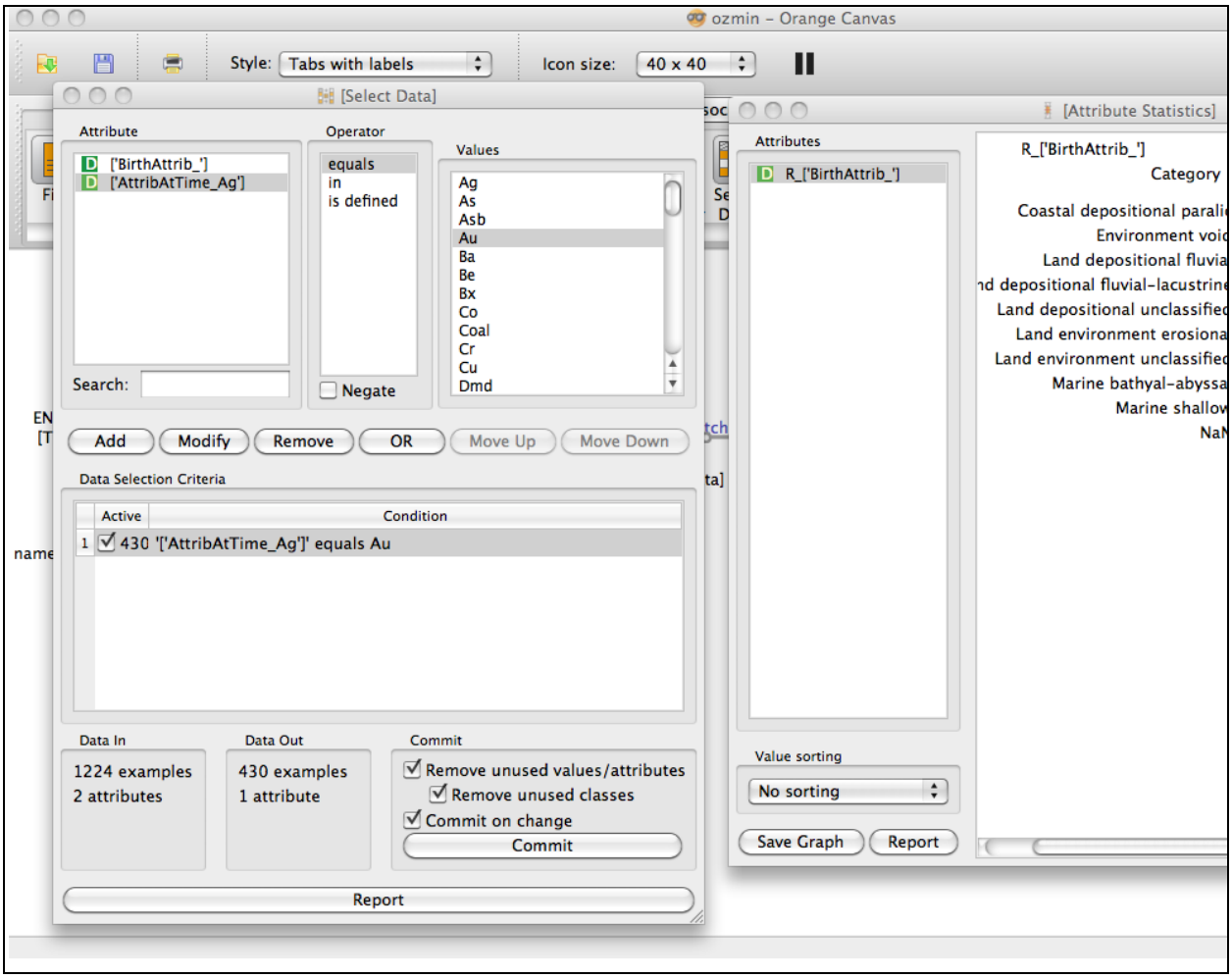
Three analyses are shown to demonstrate how the resultant Orange schematic can be used for investigating spatio-temporal associations. The "select data" widget allows for the required analysis interactivity.
Commodities that formed in *Land environment erosional* environments:



Commodities that formed in *Marine shallow* environments:



Environments in which *Gold* formed:


ozmin – Orange Canvas
Style: Tabs with labels
Icon size: 40 x 40
[Select Data]
Attribute
['BirthAttrib_']
['AttribAtTime_Ag']
Search:
Operator
equals
in
is defined
Negate
Values
Ag
As
Asb
Au
Ba
Be
Bx
Co
Coal
Cr
Cu
Dmd
Add
Modify
Remove
OR
Move Up
Move Down
Data Selection Criteria
Active
Condition
1 430 '['AttribAtTime_Ag']' equals Au
Data In
1224 examples
2 attributes
Data Out
430 examples
1 attribute
Commit
Remove unused values/attributes
Remove unused classes
Commit on change
Commit
Report
[Attribute Statistics]
Attributes
R_['BirthAttrib_']
Value sorting
No sorting
Save Graph
Report
R_['BirthAttrib_']
Category
Coastal depositional parali
Environment voi
Land depositional fluvia
d depositional fluvial-lacustrin
Land depositional unclassifie
Land environment erosiona
Land environment unclassifie
Marine bathyal-abyssa
Marine shallow
NaN