Data Mining Tutorial II: Spatio-temporal associations between North American volcanism and hot spots

Objective

The North American Vulcanism Geochemistry Dataset (called the *NAVDAT* dataset) comprises a large set of volcanic rock measurements in Western USA, with both age and geochemistry information. The data covers a complex geological setting, where there has been a combination of margin- and plume-based processes. In this case study, we try to partition data that was associated with the passing of the North American continent over the Yellowstone hot spot. Such a partitioning allows present-day regions to be identified that evolved due to the plume interaction. Geochemistry of the two (now segmented) populations allow for more indepth studies and validation. The case study illustrates the integration and quantitative analysis of data varying across both space and time.

Methodology

1. Load NAVDAT, plume, plate boundary data and rotation files into GPlates
2. Attaching the NAVDAT dataset to the North American plate
3. Compute the palaeo-distances between each rock sample and the closest plume
4. Extract the palaeo distance and hot spot for each rock sample at its respective age of appearance
5. Filter out rocks below a chosen distance threshold, and those associated with Yellowstone
6. Plot the resultant segmented sub-dataset spatially and confirm hypothesis of two completely different populations by considering geochemistries between the two segmented datasets
7. Outlook: Interactive analysis

Pre-requistes

Experience with GPlates (loading data, reconstruction animation and coregistration) as well as with the Orange data mining suite is necessary; the reading of the data mining tutorial 1, Identifying depositional environments for Australian age-coded mineral deposits, is recommended as it describes important steps in more detail. (The process of assigning plate IDs ("cookie-cutting") is also used here, but it can be skipped; another tutorial describes this more in depth.)

Step 1: Loading and visualising data in GPlates

Start GPlates from the command line as follows:

```
gplates --data mining
```

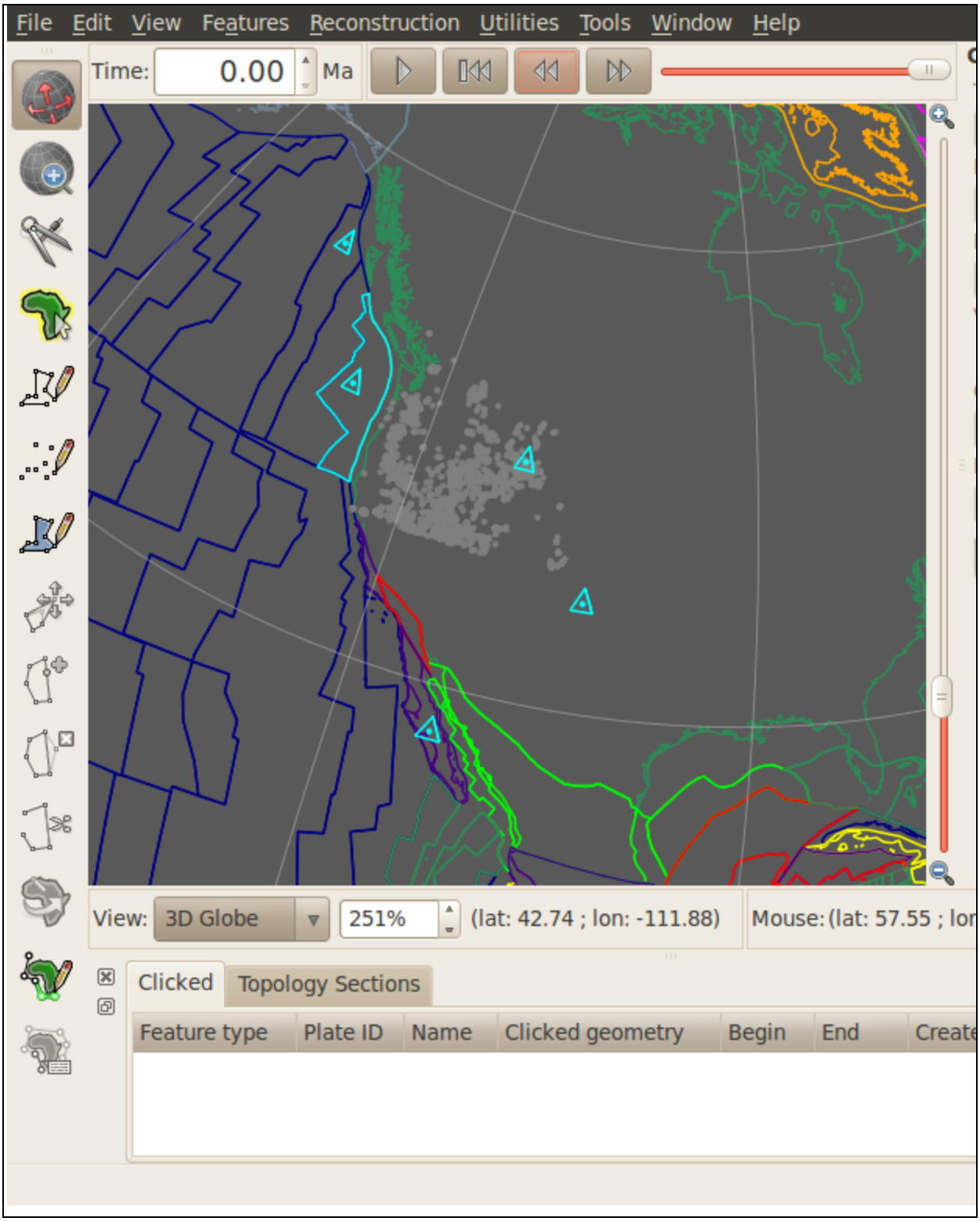
Open the following feature collections:

- NAVDAT volcanic rock measurements - this dataset is an extraction of the source data,

with samples under 60 Ma extracted, as well as samples aged using the KAR technique. To save even more computing time, the file "NAVDAT_filtered.shp" additionally had a bounding box filter applied to remove data far from the interest zone. For convenience another version of that dataset, "NAVDAT_filtered_attached.shp", is also given which is actually the output of step 2, thus allows for skipping that step.

- Global hot spots - HS_triangles.dat
- Coastlines - Global_EarthByte_GPlates_Coastlines_20091007.dat
- Plate polygons - Global_EarthByte_GPlates_PresentDay_PlatePolygons_20091015_ASEG_PESA_2010_ed24092010.gpml, these are used to attach the NAVDAT dataset to the appropriate plate (thus is not needed when skipping that step, see above)
- Rotation file - Global_EarthByte_TPW_FinalMOR20100729.rot

Configure reconstruction animation settings between 60 Ma and 0 Ma, with a 1 Ma step.
The following depicts the loaded data (HS are shown as blue triangles; the order of the layers and the colouring might need to be changed to get the same picture):

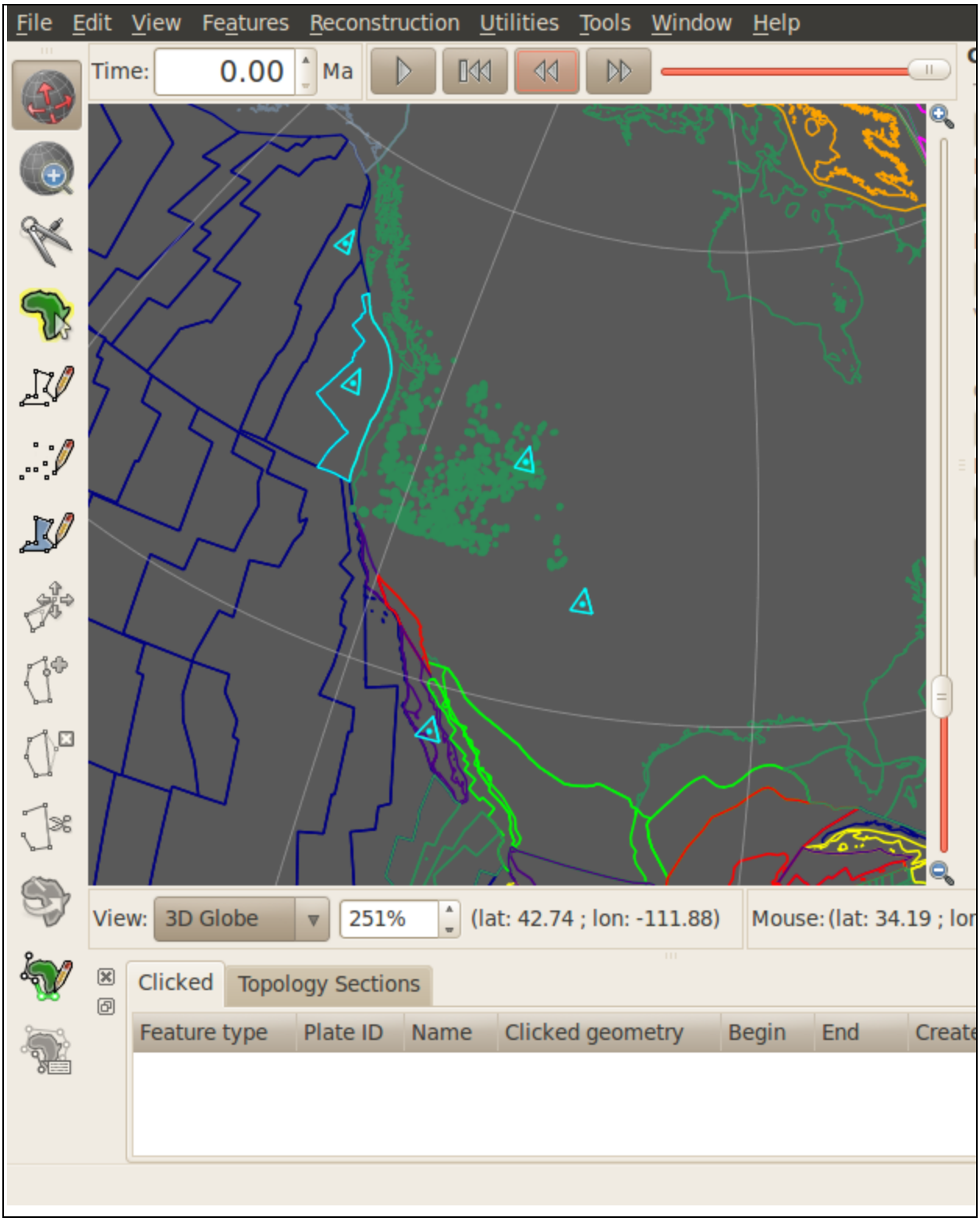


Step 2: Attaching the NAVDAT dataset to the North American plate

Since the volcanic rocks measurements do not come with reconstruction information we need to assign to it approriate plate IDs ("cookie cut" it) like already shown in another tutorial.

1. Select Assign Plate IDs ... from Feature menu.
2. Choose the plate polygons as partitioning polygons.
3. Choose the NAVDAT file ("NAVDAT_filtered.shp") as feature to partition.
4. Finally select "cookie cut" as feature partitioning method, while retaining the reconstruction time at the present and the reconstruction plate ID as only property to copy.

After a few minutes (on a dual core machine) GPlates should have calculated the plate ID assignment for the specified time period. As a result the colours of the NAVDAT points should have adopted the colour of the corresponding plate (unless another colour scheme has been chosen) and when going back in time they should move with that plate.
The loaded data after attaching it to the North American plate:



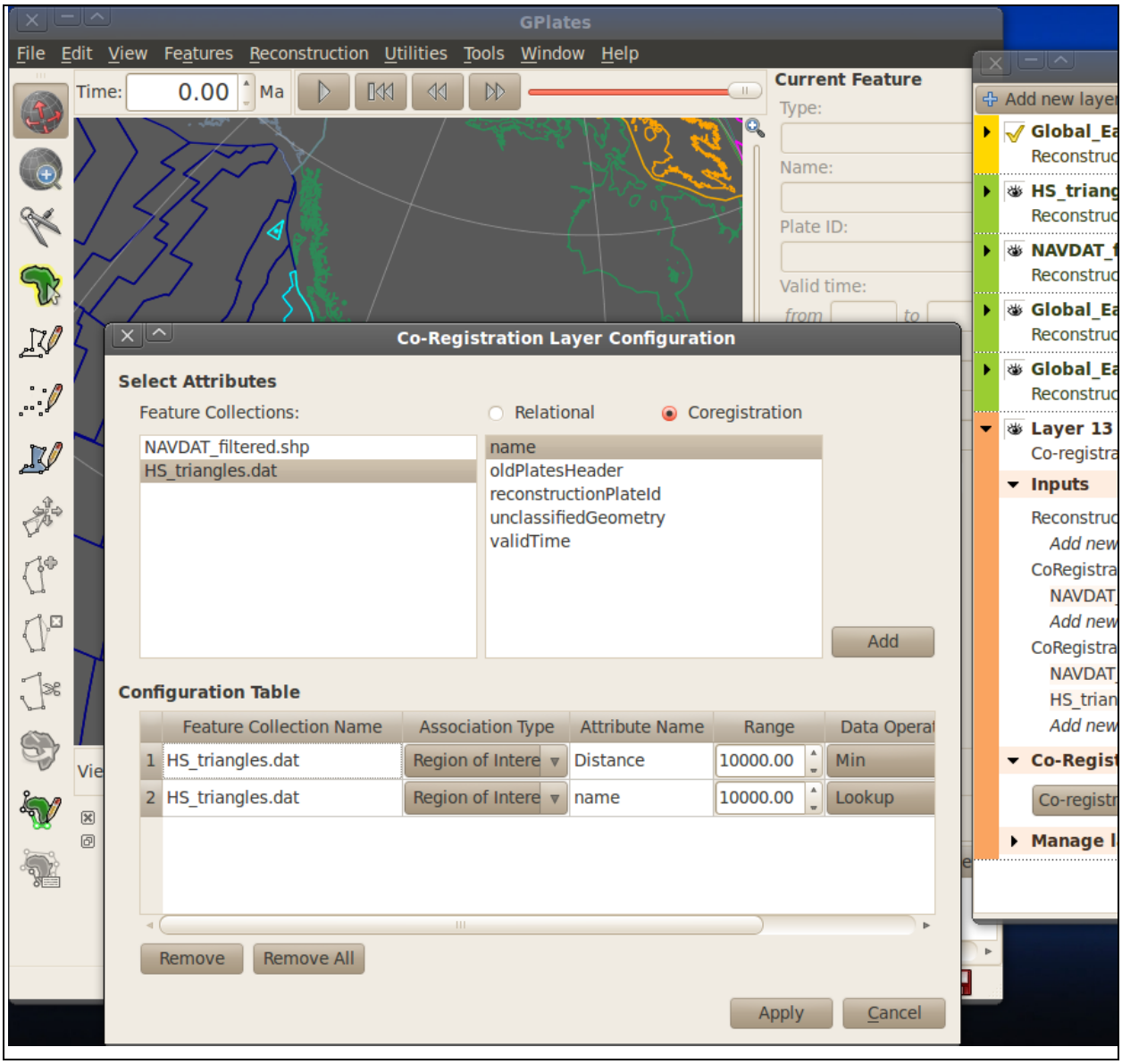
Step 3: Data coregistration

The palaeo-association to be computed is to calculate the distance between all rock samples (the seeds), and their closest hot spot for each time, storing also the hot spot name. The

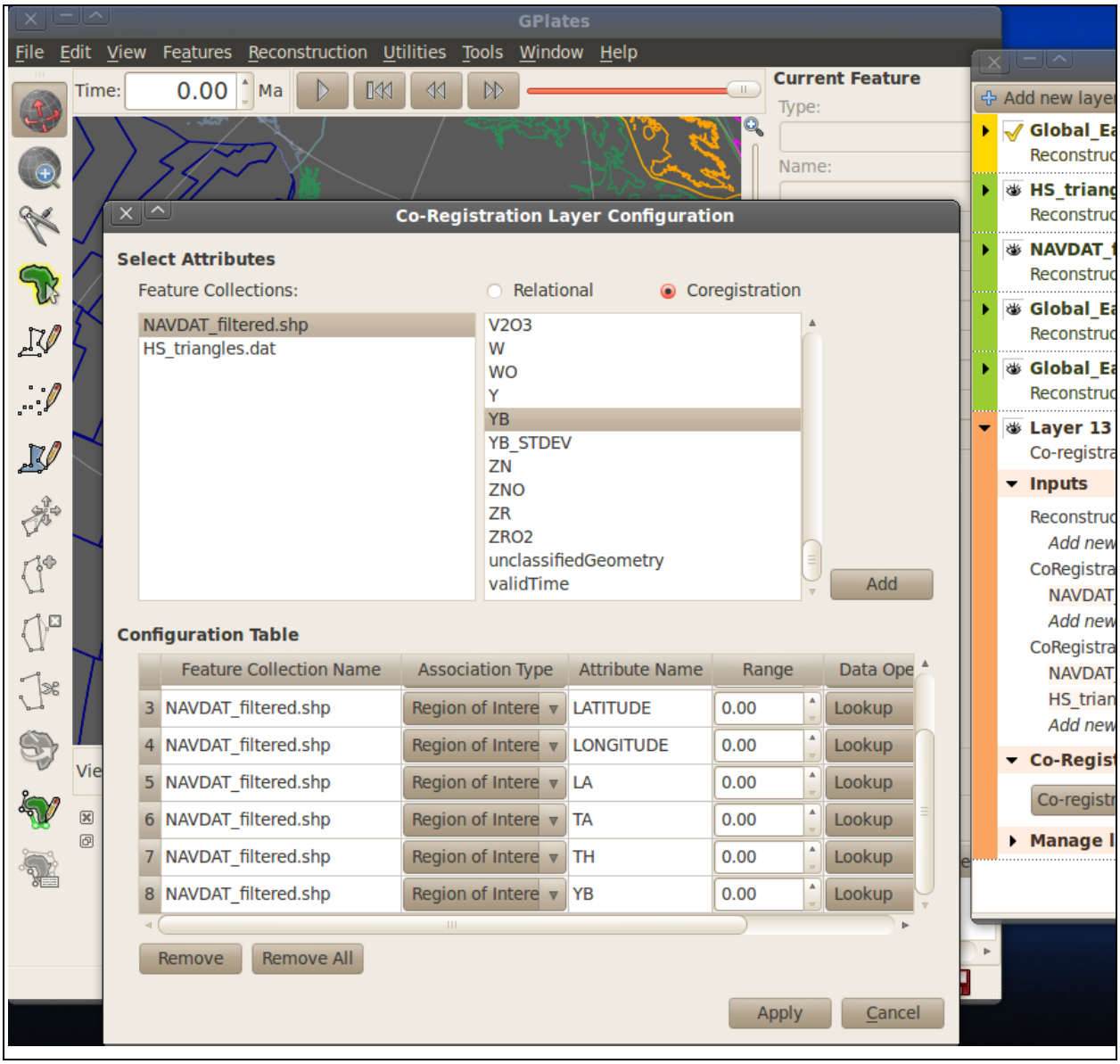
coregistration must thus be set up to compute two relational associations, followed by selection of the full geochemistry metadata corresponding to the rock samples for subsequent analysis. The following coregistration procedure should be carried out:

1. In a newly added coregistration layer choose the NAVDAT dataset, i.e. the volcanic rocks, as the *seed*
2. Both the NAVDAT and Hot Spot datasets are used as coregistration datasets
3. In the first coregistration step, choose a relational association between the seed and hot spot dataset: the distance within a 10000km radius to the nearest hot spot (a large radius ensures we always get an association, implying that the first valid distance in the time series will be the one at the time of its inception).
4. In the next association, the name of the nearest hot spot is extracted within the same radius ("Lookup" yields the nearest feature).
5. Next all the attributes with respect to the seed dataset should be coregistered to - zero radius required because these belong to the seed. The coregistration tool allows multiple attributes to be selected by dragging of the mouse.
6. The coregistration should be computed for the selected time-period and exported.

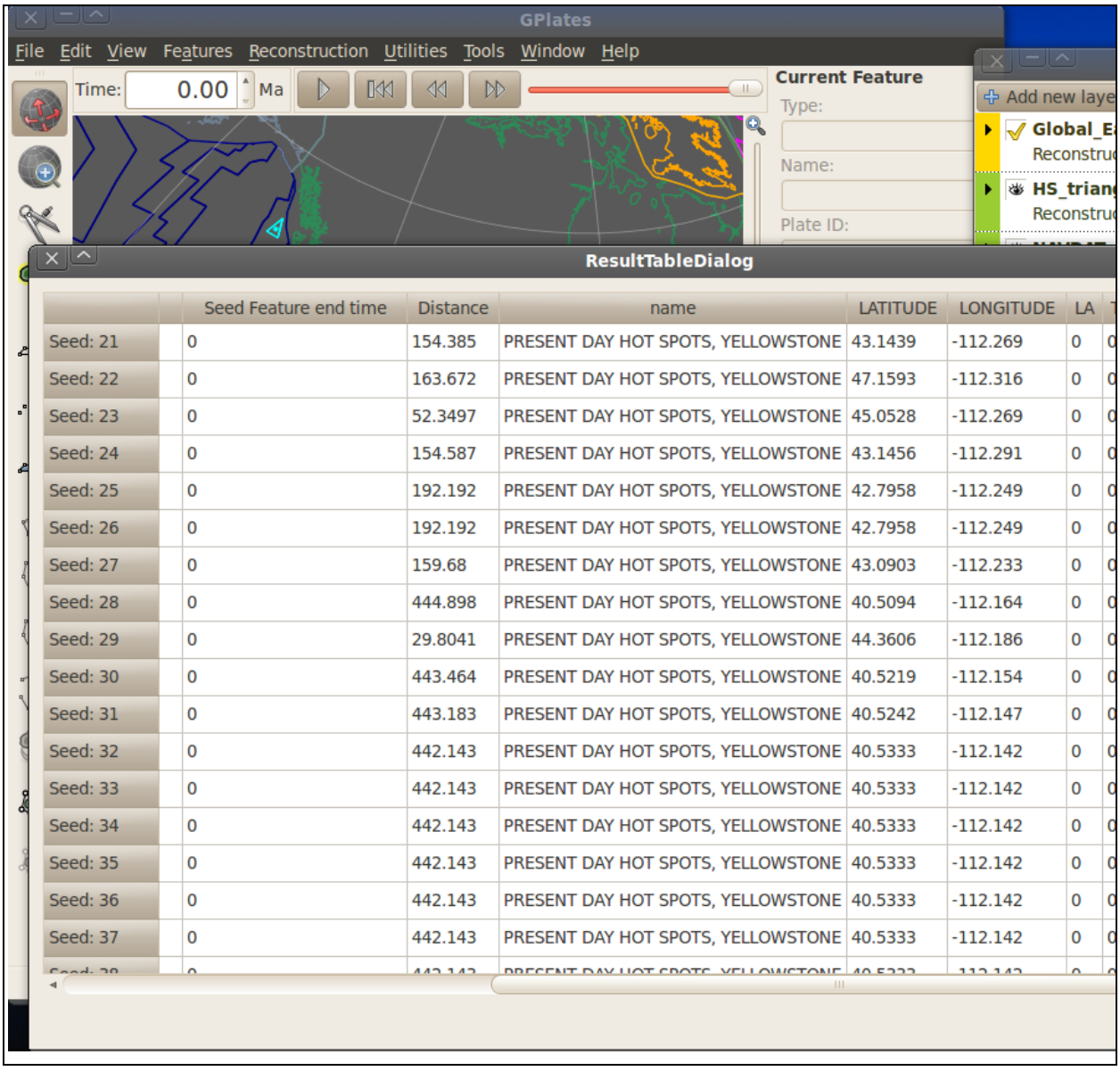
The first associations are depicted below (note that the red disk symbol in the bottom right corner of the main window indicates that the changes made to the NAVDAT dataset have not been saved yet - this, of course , only applies if step 2 has been skipped):



More selections, with geochemical attributes coregistered (here only the spherical coordinates and a selection of four chemical attributes of the volcanic rock data have been chosen for coregistration):



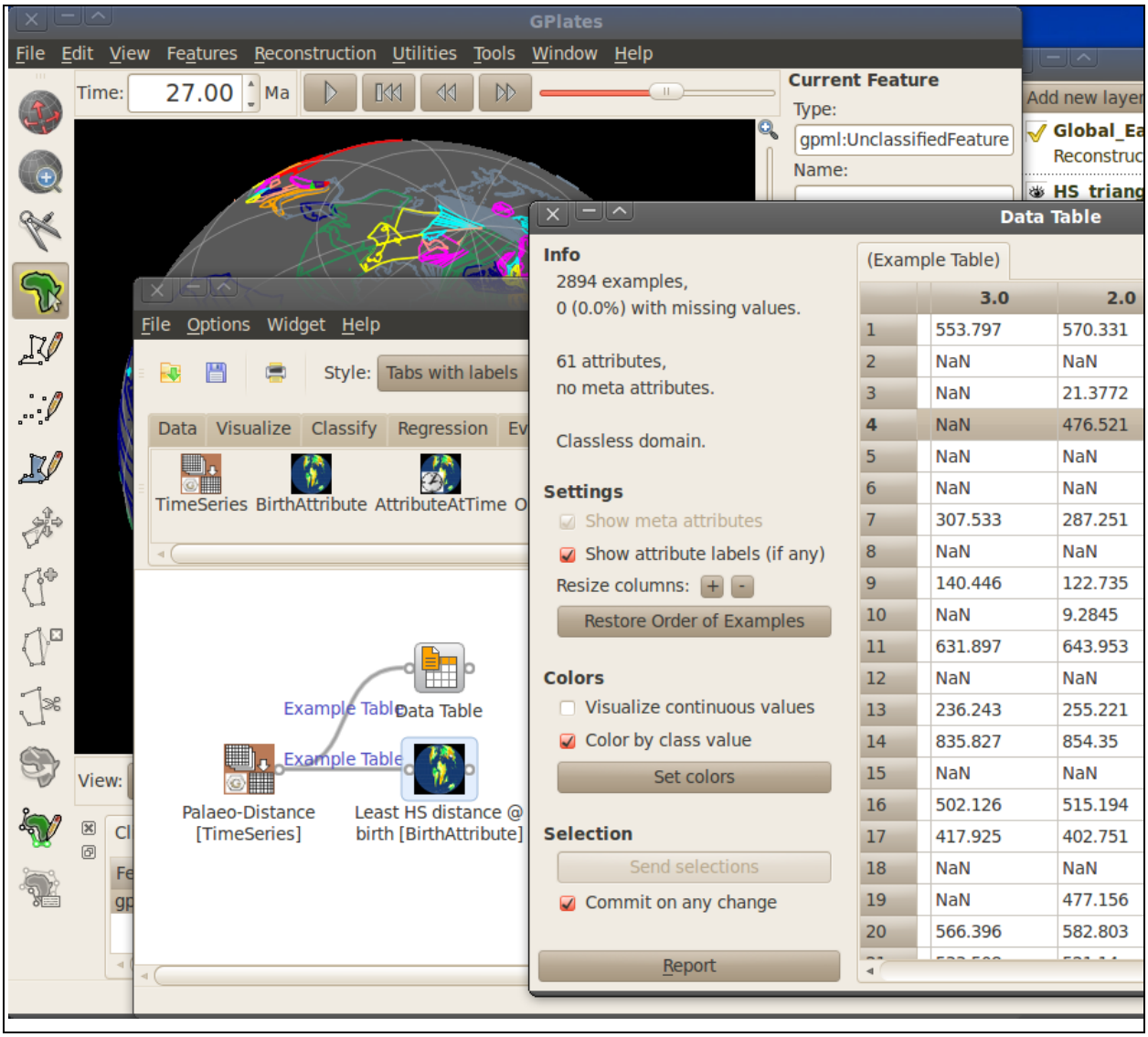
GPlates' coregistration table result (most probably there will be no coincidence in the numbering of the seeds of the picture due to the way GPlates' internally achieves the results):



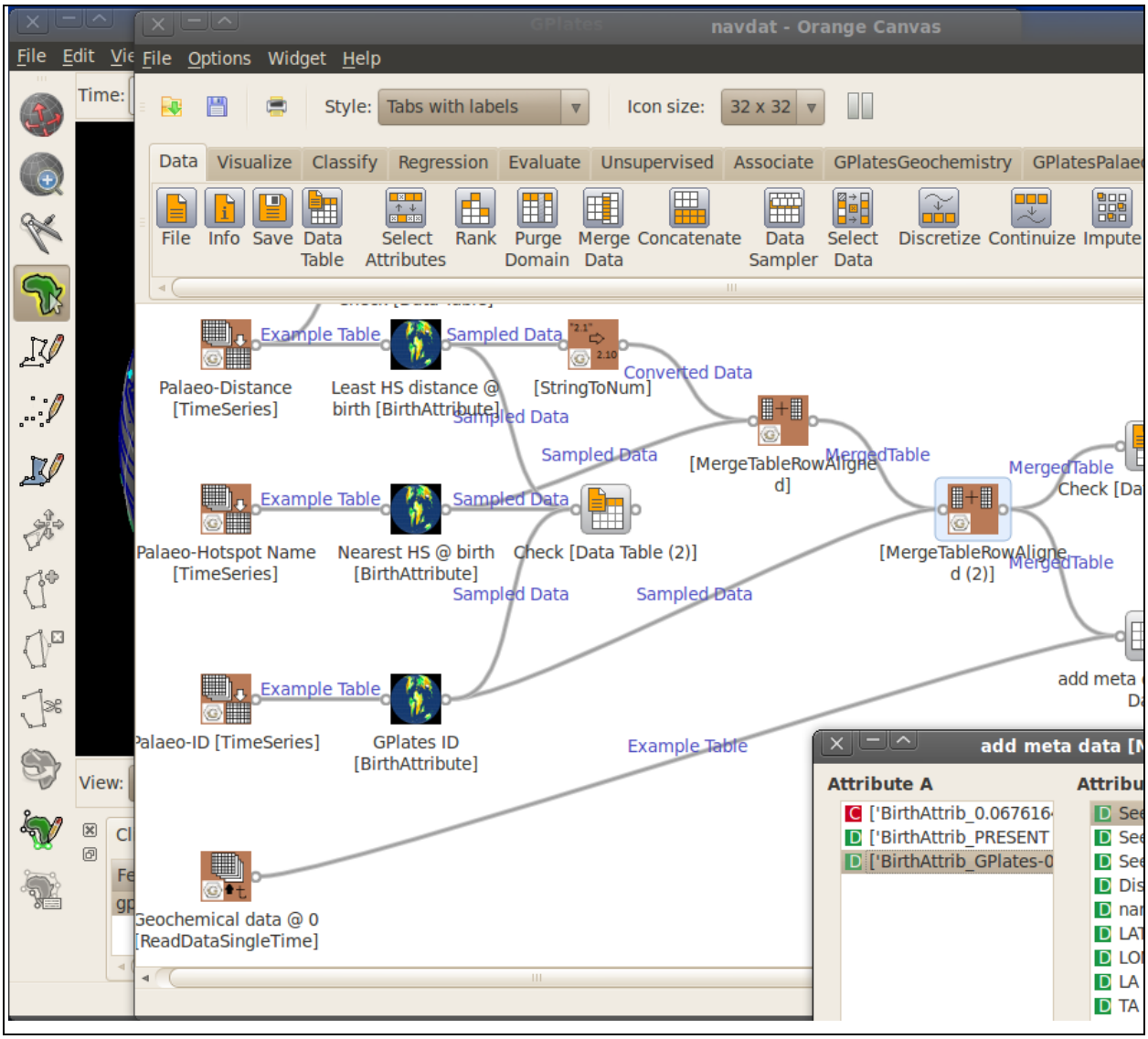
	Seed Feature end time	Distance	name	LATITUDE	LONGITUDE	LA
Seed: 21	0	154.385	PRESENT DAY HOT SPOTS, YELLOWSTONE	43.1439	-112.269	0
Seed: 22	0	163.672	PRESENT DAY HOT SPOTS, YELLOWSTONE	47.1593	-112.316	0
Seed: 23	0	52.3497	PRESENT DAY HOT SPOTS, YELLOWSTONE	45.0528	-112.269	0
Seed: 24	0	154.587	PRESENT DAY HOT SPOTS, YELLOWSTONE	43.1456	-112.291	0
Seed: 25	0	192.192	PRESENT DAY HOT SPOTS, YELLOWSTONE	42.7958	-112.249	0
Seed: 26	0	192.192	PRESENT DAY HOT SPOTS, YELLOWSTONE	42.7958	-112.249	0
Seed: 27	0	159.68	PRESENT DAY HOT SPOTS, YELLOWSTONE	43.0903	-112.233	0
Seed: 28	0	444.898	PRESENT DAY HOT SPOTS, YELLOWSTONE	40.5094	-112.164	0
Seed: 29	0	29.8041	PRESENT DAY HOT SPOTS, YELLOWSTONE	44.3606	-112.186	0
Seed: 30	0	443.464	PRESENT DAY HOT SPOTS, YELLOWSTONE	40.5219	-112.154	0
Seed: 31	0	443.183	PRESENT DAY HOT SPOTS, YELLOWSTONE	40.5242	-112.147	0
Seed: 32	0	442.143	PRESENT DAY HOT SPOTS, YELLOWSTONE	40.5333	-112.142	0
Seed: 33	0	442.143	PRESENT DAY HOT SPOTS, YELLOWSTONE	40.5333	-112.142	0
Seed: 34	0	442.143	PRESENT DAY HOT SPOTS, YELLOWSTONE	40.5333	-112.142	0
Seed: 35	0	442.143	PRESENT DAY HOT SPOTS, YELLOWSTONE	40.5333	-112.142	0
Seed: 36	0	442.143	PRESENT DAY HOT SPOTS, YELLOWSTONE	40.5333	-112.142	0
Seed: 37	0	442.143	PRESENT DAY HOT SPOTS, YELLOWSTONE	40.5333	-112.142	0

Step 4: Data Mining

After exporting the coregistration result open Orange, and extract the three necessary time-series from the coregistration, i.e. the palaeo-distance, the palaeo-hotspot name, and the palaeo-ID (this last one will be unnecessary in a future release). The following depicts the resultant time-series for a few exemplar rocks (with the widget renamed appropriately):

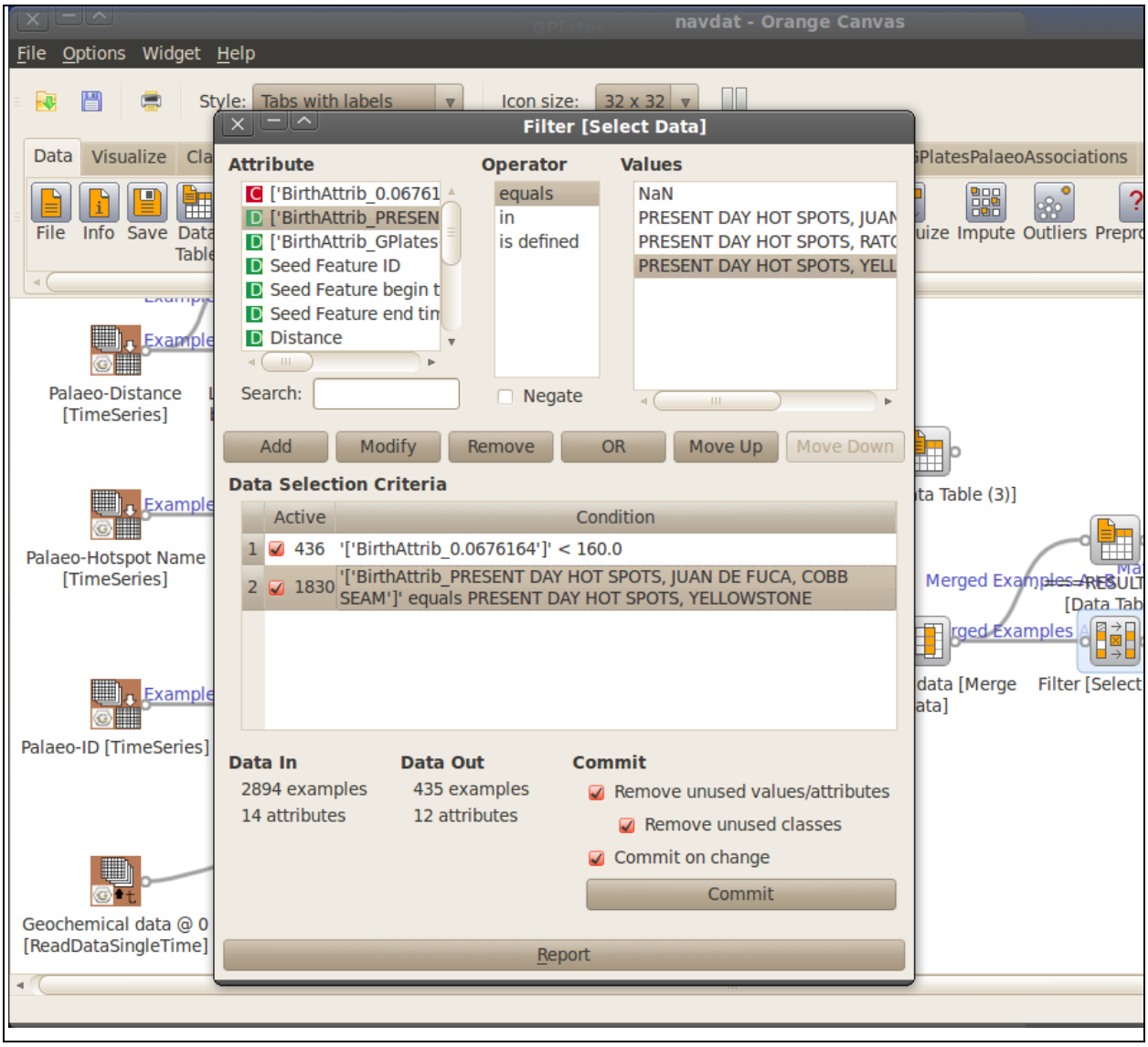


The three time-series should be extracted, and their values at time of birth computed (this is how we localise rocks in close proximity to the hot spots at their inception). These three vectors should then be combined. The full geochemical metadata is extracted from the present-day. A special merge component is then used to append the metadata to the palaeo-features - this widget allows the GPlates-IDs to be selected for both datasets, thereby obtaining row-wise correspondence (database tables operate in the same fashion for table joins). The following schema depicts the data processing phase, resulting in the final merged dataset (tables denoted with "Check" are just for checking intermediate results):



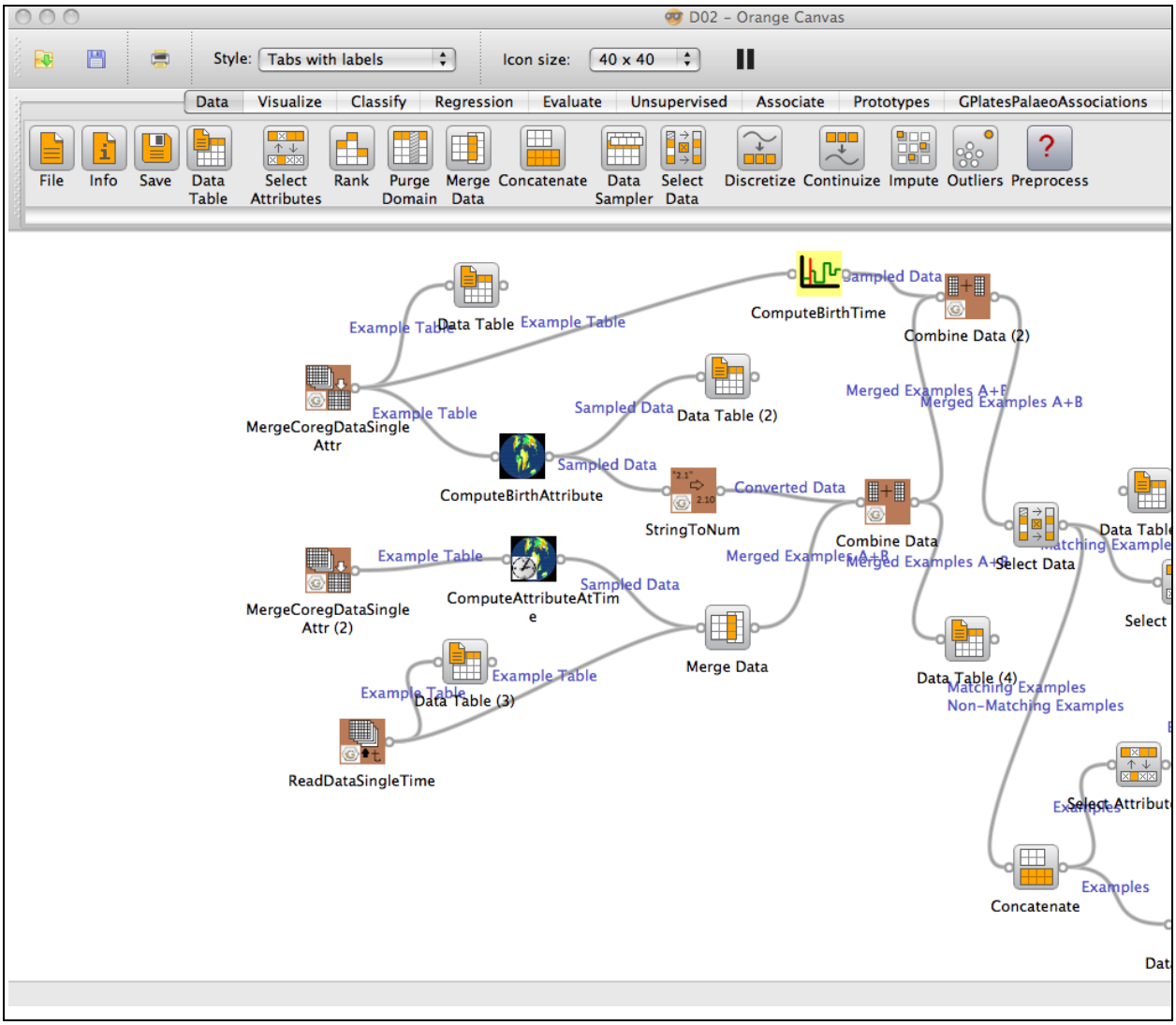
Step 5: Filter result

The next step is to filter out rocks that have formed close to the hot spot, and those that have formed close to Yellowstone in particular. The *Select Data* widget allows for defining of this filtering operation. Firstly for the palaeo-distance, a threshold distance should be chosen to suit the problem e.g. 200km. Next the palaeo-name attribute should be filtered using the same widget to extract data corresponing to Yellowstone only. The following snapshot captures the filtering step:

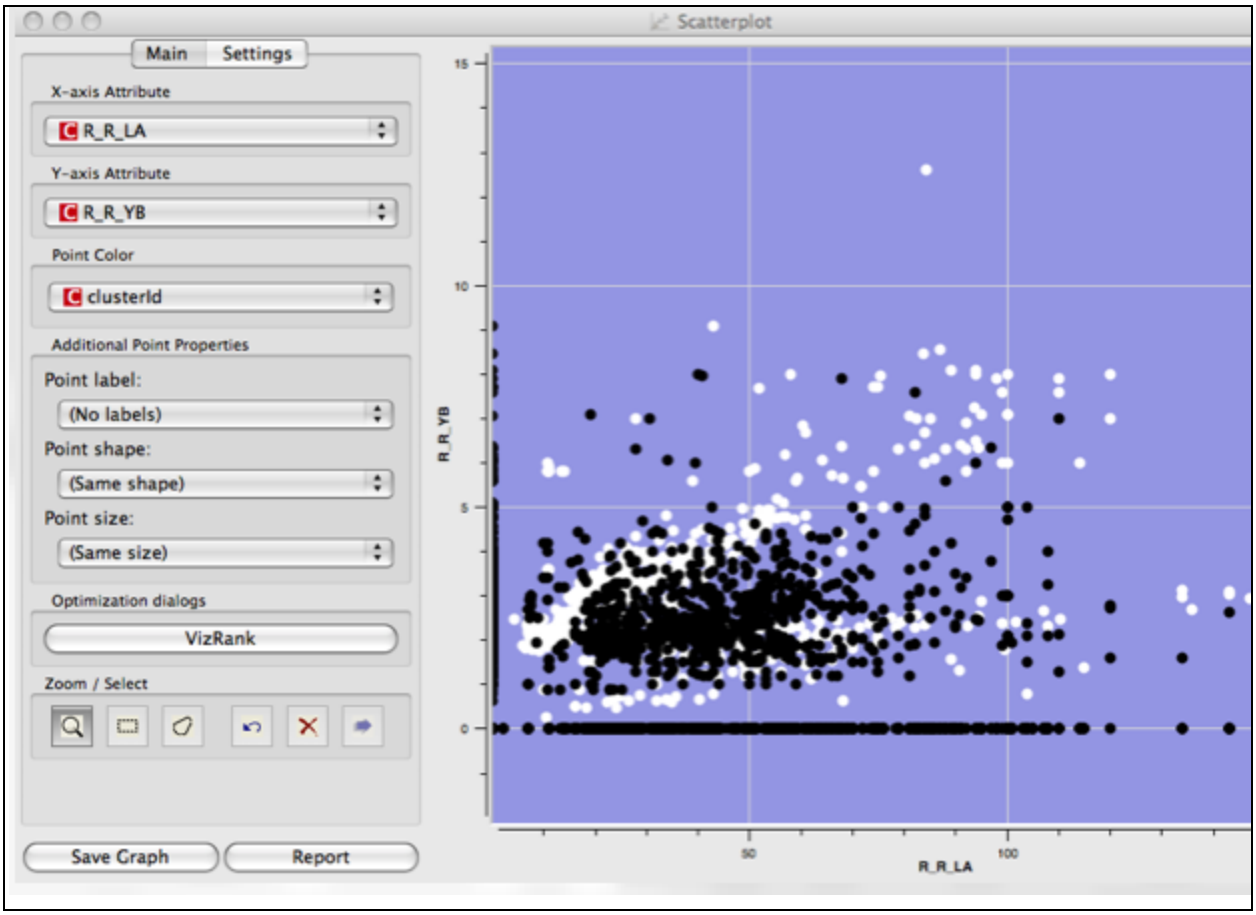


Step 6: Analysis

The final part of the data mining is to extract the lat-long coordinates of the segmented rock samples and to analyse the geochemistries of the two resultant populations, i.e. rocks associated with Yellowstone, and those not associated. The former task is computed by selecting the Latitude and Longitude attributes, and exporting. Several geochemical studies are possible. In comparing plume-based magmatism to margin-based magmatism, rare earth elements such as Yb, La, Ta and Th are good candidates. Coming soon: direct access to decorating selections in GPlates.
An important hint: The Select Data widget filters data and provides the complement as output, too, and automatically appends an identity for subsequent processing - in this study we need to identify both samples associated with hot spots, and those not. And the *Select Attribute* widget requires one of the two attributes to be chosen as a class attribute for saving to work. The following plot depicts a possible schema to complete the data mining task:



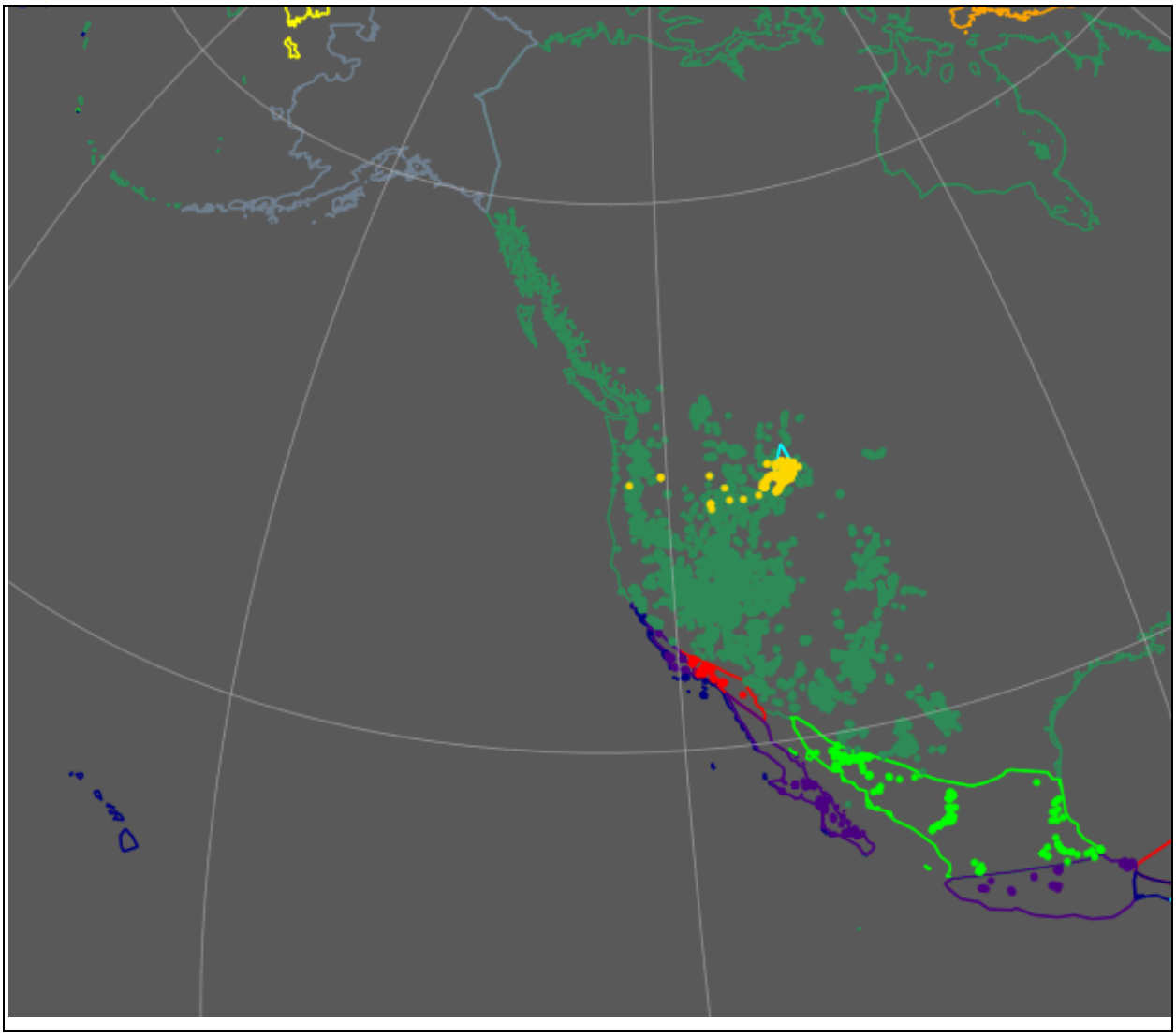
The following plot depicts Yb vs La for the two populations:



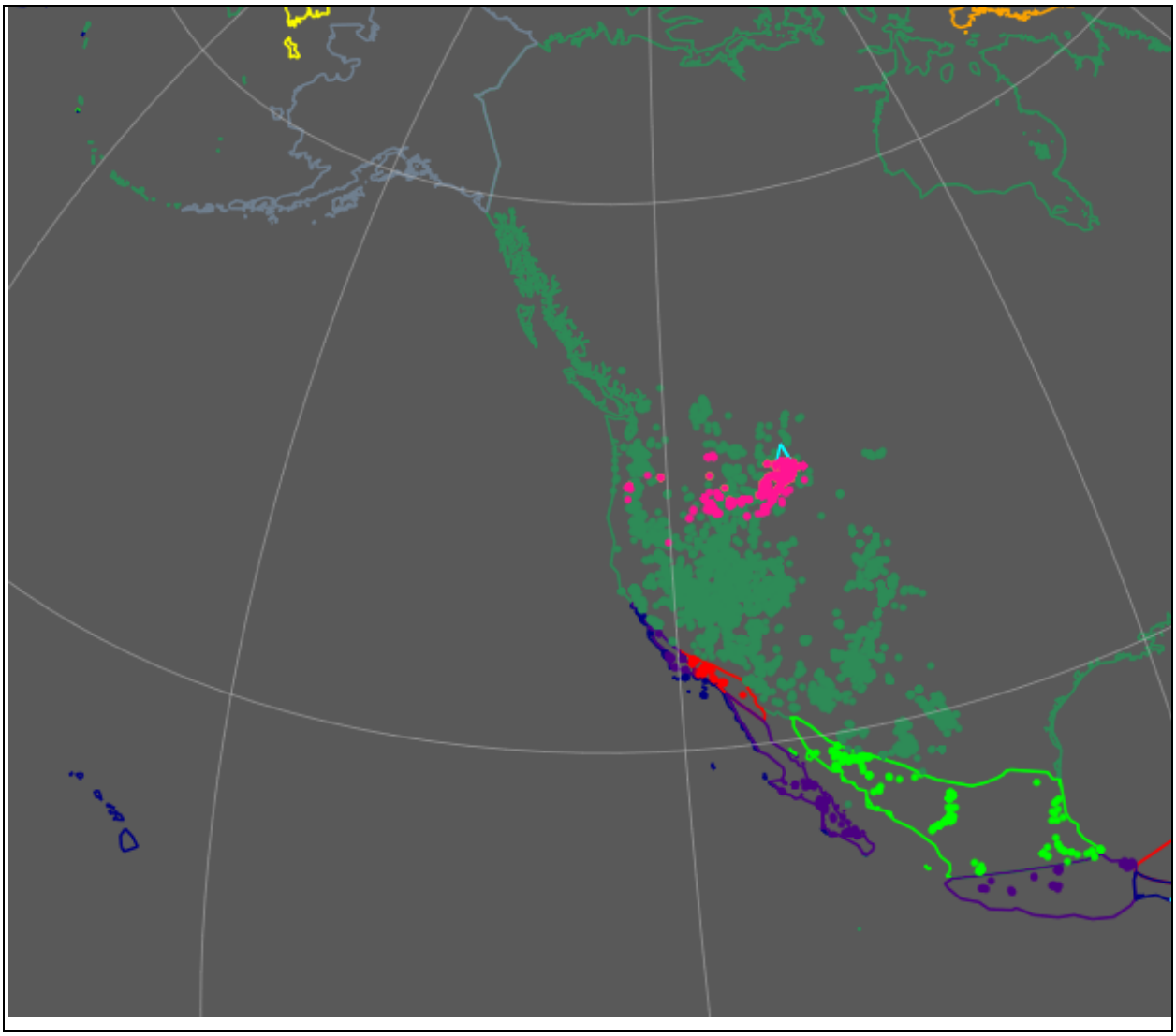
Step 7: Outlook: Interactive analysis

Coming soon: changing the palaeo-distance threshold in the data mining tool will be fed back to GPlates, providing graphical feedback. The following three palaeo-distance thresholds depict the segmentation implication:

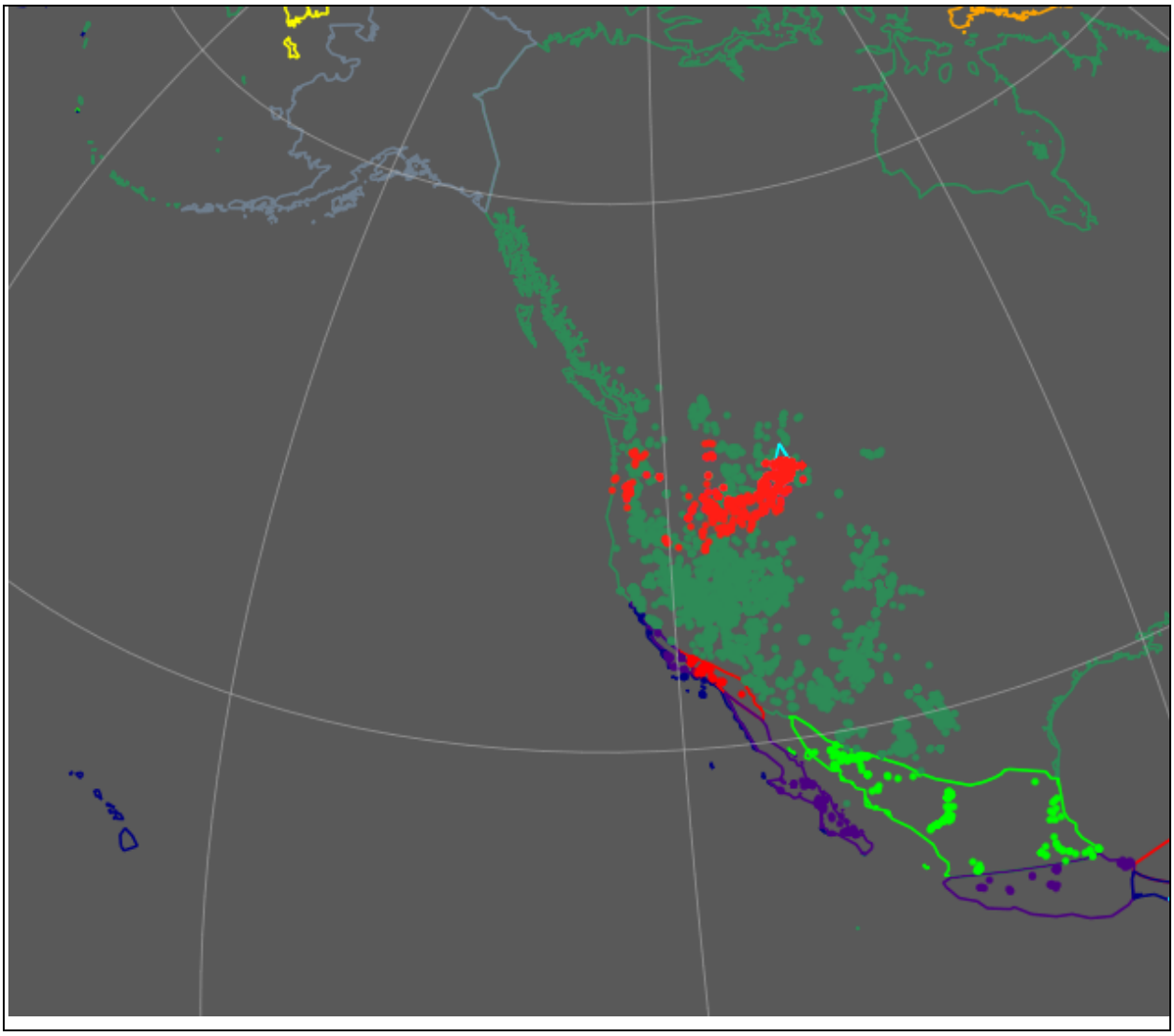
Palaeo-distance threshold 160km:



Palaeo-distance threshold 260km:



Palaeo-distance threshold 360km:



The segmented result can also be viewed through time in GPlates, for example consider the following snapshots through 10, 4, 2 and 0 Ma respectively: