**Data Mining of Phanerozoic datasets**

**Andy Zehady; draft of 2 Feb 2019**

**Goal**

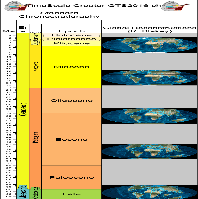
Find **periodicities** and other correlation, **cause-effect** relationships in *TSCreator* data since Cambrian period (541 Ma). (NSF project research lines, page-2)

**Testable Hypothesis**

1. H1 - “*Rates of evolution are correlated with rates of geochemical and sea-level change.*”
   1. Are there postulated long-term (2-5myr) astronomical cycles?
2. H2- “*The Earth has had semi-periodic episodes of unusual surface/biological change.*”
   1. What intervals are relatively eventful in Earth’s history and which intervals are anomalously “quiet”? Why?
   2. What are the correlation with catastrophic events? Do these oscillations support “Nemesis” or other periodic catastrophe models (20-60 myr)
   3. Which peaks in events correlate with peaks in volcanic activity, continental collisions and asteroid impacts?
3. H3- “*Pulses of biological evolution occur simultaneously with global changes in sediment facies*.”
   1. What are main trends with time and space? (i.e. trend over geologic time.. space..)
   2. Relationship between changes of sediments with evolution?
   3. What are the correlation with climate shifts and continental collisions?

**Project TimeLine**

|  |  |  |
| --- | --- | --- |
| Task | TimeLine | Comment |
| 1. Goal:    1. Marine Genera Range (loading + calculating speciation + extinction)    2. Which other curve will be relevant?       1. C-13, O-18 (Cenozoic), Sr    3. Correlation analysis    4. Factor analysis    5. Spectral analysis | Feb, 2019 |  |
|  |  |  |
|  |  |  |

**Available Data**

18

**Phanerozoic**

|  |  |
| --- | --- |
| **Dataset** | **Description** |
| **1. Marine genera ranges**  **(18,000 genera ranges )** | Through Phanerozoic based on the compendium by Jack Sepkoski that was published in 2002 |
| **2. Phanerozoic(0~541 Ma) sea level curve**  (Other phanerozoic sequence curves, e.g. Global sequence, Coastal onlap, T-R cycles (1st & 2nd order)) | Computed as mid-point of Coastal-onlaps.  (SEPM charts, 1998), CRET (2015) = Revised from: Haq, B.U., 2014.  [T-R cycle](https://en.wikipedia.org/wiki/Stratigraphic_cycles) ([Wiki](https://en.wikipedia.org/wiki/Stratigraphic_cycles))  Transgressive: Landward shoreline and facies shift  Regressive: Seaward shoreline and facie shift  **First-order cycles**  This cycle is most likely caused by the break-up and formation of super-continents. |
| **3. Stable Isotopes**   1. **Oxygen-18 curves and events**     1. **Cenozoic-Campanian marine Oxygen-18**    2. **Paleozoic-Mesozoic oxygen-18 and tropical sea-surface temperature** 2. **Carbon-13 curves and events** 3. **Strontium 87/86 ratio** | * 1. Detailed cenozic oxy-18 curve **Source:** Derived from Cramer 2009 Saltzman and Thomas (GTS2012 Carbon-isotope chapter); but only every 10th item from 9-point averaging of Benthic foraminifer compilation (29000 data points in original) is shown   2. Different sources for Cret-Jur; then Veizer-Prokoph 2015 for Cambrian-Triassic; but using Oxygen-to-Temp conversion from Veizer-Prokoph 2015  1. Triassic-lowermost Ladinian from Sun et al (2012) as cycle-adjusted by Mingsong Li. NO data for upper Anisian. Carboniferous = Saltzman, Geology 2003; Gzelian through Permian = Buggisch et al., PPP 2015. Devon = Average of the Max and Min envelope of Becker in GTS2012. Ordov = Bergstr\_m et al., 2009. Lethaia; Latest Ordov and early Silurian = Melchin et al (2014) with Hirnantian enhanced from Bergstr\_m et al (2014); rest of Silurian = Cramer, 2011; but fit to be consistent with base-Devon Klonk excursion of Becker (2012). CAMBRIAN = Zhu et al. (2006; provided by Loren Babcock, then rescaled by Peng-Babcock for Concise GTS in Dec07 2. Sr 87/86: Reflects the rate of oceanic spreading and continental erosion. **Source**: John McArthur 2004, Dovonian from Becker 2012, Silurian from Cramer 2011 |
| **4. Impacts**   1. Global Impacts (>50 km crater) 2. Regional Impacts (European, Russia and Asian, Australian, African, North American, South American) 3. Large Igneous Provinces 4. Passive Margin onsets and terminations (column (opening + closing of margins) | 1. Mainly from Earth Impact Database, 2008. [[Impacts](http://www.passc.net)]."   c) Mainly from Large Igneous Provinces Commission. (2008). Large Igneous Provinces (LIPs) Through Time.. For details click [[LIPs](http://www.largeigneousprovinces.org/) ].  d) Bradley, D.C., 2008. Passive margins through earth history. Earth-Science Reviews, 91: 1-26. doi:10.1016/j.earscirev.2008.08.001. (Especially the on-line supplement tables.) |
| **5. Regional biostratigraphy and basin lithostratigraphy**  **(More than 300 lithology columns)** | Regions include **Australia, New Zealand, China, India, Malaysia, Africa, Britain, Belgium, South America, USA-Alaska, Canada, Mexico, Russia**. |

**Methods – How?**

**Handling multiple timeseries data.**

Some timeseries have no data points in certain intervals, some have more data points in certain time period. For example, we have oxy-18 data points for mainly Cenozoic era. We can smooth taking average with sliding window to match datapoint for multiple timeseries. Need to be careful with whether to take overlapping sliding windows or not, depending on the data. Some curve may also have unevenly sampled data point. In such cases, it may require interpolation of data points.

We can also use Rate of change (difference) of certain curves like sea level, temperature, carbonate trends, oxy-18(temperature) to see the effect.

Conduct analyses on three different time intervals: 1) entire Phanerozoic; 2) Meso-Cenozoic; and 3) Paleozoic. Examine the Meso-Cenozoic records first, examine the Paleozoic records in the context of the Meso-Cenozoic; and compare results to assess the entire Phanerozoic.

**Frequency count per bin.**

For evolution data, count million-year level origination and extinction event. Determine the interval with high frequency event.

Relate with global event count, super LIP/major LIP and other stable isotope curve point indicating temperature, sea level, continental weathering. See whether high frequency events align with eventful interval. Think about how physical phenomena can influence evolution?

**Correlation-coefficient**

Find cross-correlation co-efficient using Pearson correlation (or others).

**Lag analysis**

Correlation can be lagged too. We can do lagged autocorrelation to see to find the lag between two timeseries.

**Coherence**

In time series analysis, and particularly in spectral analysis, it is used to describe the strength of association between two series where the possible dependence between the two series is not limited to simultaneous values but may include leading, lagged and smoothed relationships

**Spectral analysis**

For the data points covering same time period (like whole Phanerozoic, whole Cenozoic or Mesozoic etc.) perform spectral power analysis to determine frequency/periodicity of each curve. Does the same periodicity occur on multiple curves?

For analyzing underlying factors between two or more datasets, the following analysis can be done. Each has its strength and weaknesses and depends on the relationship between the datasets (needs caution). Reference: <https://atmos.washington.edu/~dennis/552_Notes_4.pdf>

**Principal Component Analysis (PCA), Singular Value Decomposition (SVD)**

**Empirical Orthogonal Function (EOF) Analysis**

**Maximum Covariance Analysis (MCA)**

Identifies the correlation patterns of two data matrices that are examples of different structures, or state vectors, but which share a common sampling dimension. For example, the fields of water temperature and surface chlorophyll content, measured at the same set of times (does not have to be the same locations). It does so by applying SVD to the covariance matrix of the two fields. Example: <https://menugget.blogspot.com/2011/12/maximum-covariance-analysis-mca.html#more>

**Canonical Correlation Analysis (CCA)**

[**http://www.ocgy.ubc.ca/~william/EOSC510/Ch3/Ch3.pdf**](http://www.ocgy.ubc.ca/~william/EOSC510/Ch3/Ch3.pdf)

Canonical correlation analysis (CCA) looks for patterns in two spacetime datasets with maximum temporal correlation coefficient. CCA does not necessarily pick patterns which explain much covariance and can be severely affected by random sampling fluctuations. To minimize these issues, the two fields should be prefiltered by projection onto a subset of their leading EOFs sufficient to explain most (e. g. 90%) of their variance. See Bretherton et al. (1992). Ref: <https://atmos.washington.edu/~breth/classes/AS552/lect/lect22.pdf>