An Environmental Knowledge Engine for Model-Based Geostatistics

- Can we predict the presence of a taxon ? -

Juan M. Escamilla Molgora
j.escamillamolgora@lancaster.ac.uk
Supervisors: Peter Atkinson, Luigi Seda and Peter Diggle
Lancaster Environment Center | FST | FHM

CONACYT · Lancaster University · GBIF Global Biodiversity Information Facility · CHICAS centre for health informatics, computing, and statistics

Aims

* Formalise a framework for exploring, analysing and modelling relationships of ecological and environmental spatio-temporal data.

* Implement and derive network traversal algorithms for extracting latent patterns of ecological phenomena.

* Implement a graph-based spatial regression of acyclic networks (taxonomic trees) based on the spatial co-occurrences of species (presence-only).

* Explore structural properties of the graph structures within the geographic space (e.g. resilience,connectedness, average degree, etc).

* Develop a methodology for modelling absences in presence-only data.

* Provide the enviromental comunity with an Open Source tool for handling heterogeneous data descentralised and scalable.

.

keywords: ecology, complex networks, spatial analysis, graph-based regression, complex dynamics.

Motivations

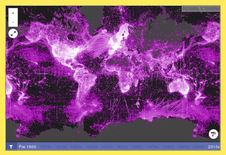
Ecology is an integrated science that studies the *relationships*(interactions) of living beings with the environment. The science has evolved radiating into different specialisations that can be grouped by different aspects; spatial scale (soil microbial ecology to continental scale (Macroecology)); organisational level (from population ecology to ecosystem ecology); functional properties (e.g. pollination ecology, behavioural ecology, reproduction ecology), etc.
Evolution is the process that underlies these biological patterns. It is through this light that complex adaptations and strange conspicuous coincidences can be **explained scientifically**.

How to do it:

Requirements:
Massive Data handling of at least 200 GB.
This is achieved using Postgis as the OLAP engine. Includes geoprocessing functions and indexes.
Efficient computing
The processeses are implemented in a RESTfull interface that can be parallelised using containers or virtualization technologies.
Scalable
The environmental data has a history of being open access. New methodologies, protocols and common metadata have been developed to be analysed in decentralized semantic queries. Data matching will come not only by joining spatial and temporal components but also by joining objects and relationships under the concept of ontologies. This implies that in the near future the data will grow exponentially and a common framework should support it. Migration to **BigData** infrastructures should be easy.

In this work I make use of the world biggest repository of biodiversity data (GBIF) to aggregate biological meaningful relationships in space. I want to understand how these structures transform the environment and dialectically how does the environment affects them**.**

Data source and processing techniques
The Global Biodiversity Information Facility (GBIF) is an international effort for merging different datasets from around the world in a unified, heuristically curated and open access platform. The database comprises approximately **400,000,000** records of presence (occurrences) of species. Each record has a collect date-time stamp, geographic coordinates and the full taxonomic classification by: species, genus, family, class, order, phylum and kingdom.

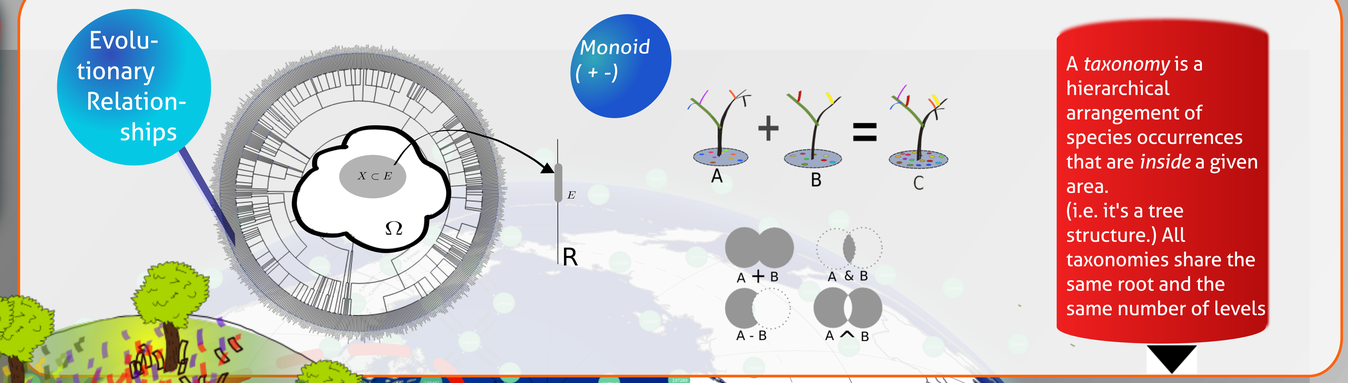


The GBIF visualization:
http://gbif.org

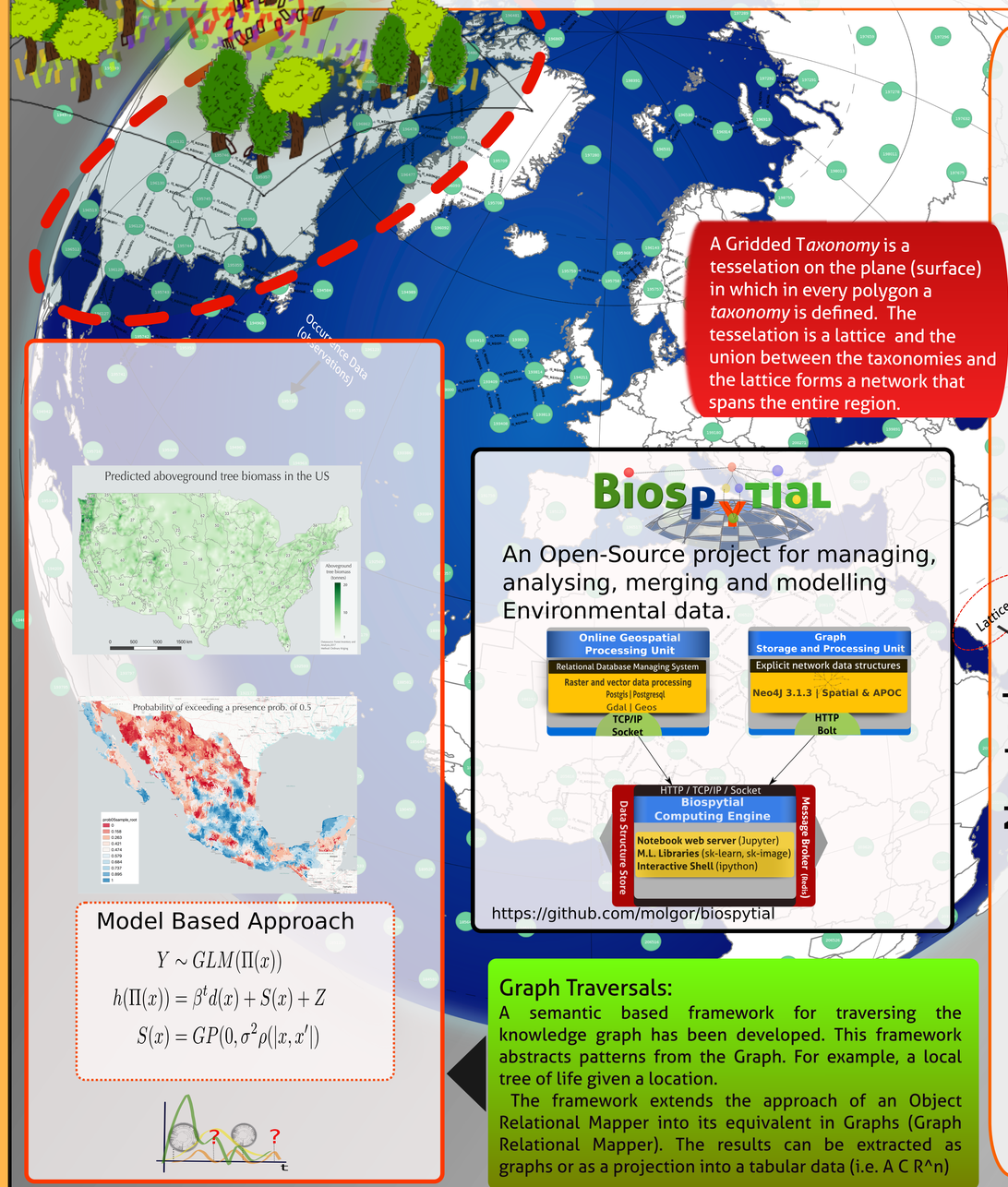
I make use of the systematic classification of species to construct a hierarchical structure of organisms based on the relationship "*being a member of..* ". The structure is a tree *(acyclic graph)* and it is based on the evolutionary features (genes, morphology, etc) that reconcile better the phylogenetic differences.

Graph Database System as the processing engine

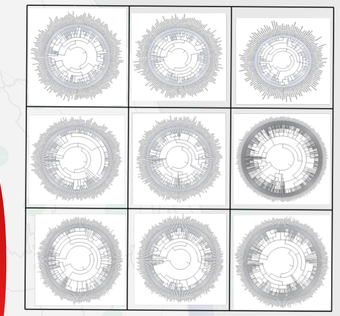
Use graph structures for representing data and performing semantic queries. Data is represented by nodes, edges (relationships) and properties (key-values lists). The use of links instead of joins in gives a tremendous advantage in performance when compared to RDMS (Have, et.al. 2013). Query processing languages like *Gremlin*, and *Cypher* implement the most common network based operations. Traversing and analysing network-based BigDataBases can't be easier.



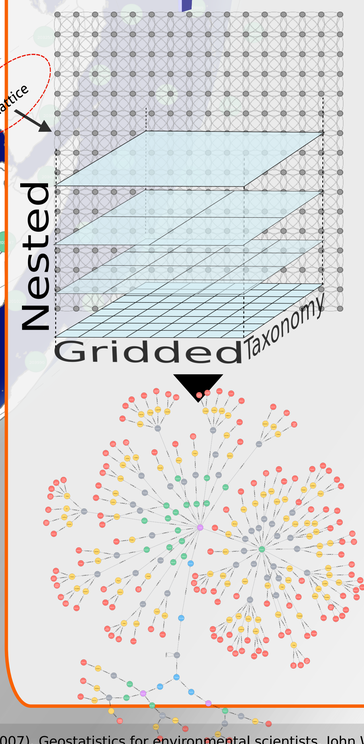
A *taxonomy* is a hierarchical arrangement of species occurrences that are *inside* a given area.
(i.e. it's a tree structure.) All taxonomies share the same root and the same number of levels



A Gridded T*axonomy* is a tesselation on the plane (surface) in which in every polygon a *taxonomy* is defined. The tesselation is a lattice and the union between the taxonomies and the lattice forms a network that spans the entire region.



The Nested structure induces a Quad-tree



Model Based Approach

$$Y \sim GLM(\Pi(x))$$
$$h(\Pi(x)) = \beta^t d(x) + S(x) + Z$$
$$S(x) = GP(0, \sigma^2 \rho(|x, x'|)$$

BiospYtiaL

An Open-Source project for managing, analysing, merging and modelling Environmental data.

https://github.com/molgor/biospytial

Graph Traversals:

A semantic based framework for traversing the knowledge graph has been developed. This framework abstracts patterns from the Graph. For example, a local tree of life given a location.
The framework extends the approach of an Object Relational Mapper into its equivalent in Graphs (Graph Relational Mapper). The results can be extracted as graphs or as a projection into a tabular data (i.e. A C R^n)

References:
* Gelfand, Alan E., et al., eds. Handbook of spatial statistics. CRC press, 2010
* Diggle, P.J., Tawn, J.A., Moyeed, R.A., 2002. Model-based geostatistics. J. R. Stat. Soc. Ser. C (Applied Stat. 47, 299–350)
* Robinson, Ian, Jim Webber, and Emil Eifrem. Graph Databases: New Opportunities for Connected Data. " O'Reilly Media, Inc.", 2015.
* Have, Christian Theil, and Lars Juhl Jensen. "Are graph databases ready for bioinformatics?." Bioinformatics 29.24 (2013): 3107-3108
* Webster, R., & Oliver, M. A. (2007). Geostatistics for environmental scientists. John Wiley & Sons.
* Newman, M. (2010). Networks: an introduction. OUP Oxford

Aknowledgements to the Open Source and Free Software community specially: GDAL, GEOS, Python, Django, Numpy, QGIS, Neo4J, PostGIS, PostgreSQL and R.

Made with Inkscape and other FOSS.