gCube Statistical Manager: the Algorithms

Gianpaolo Coro*, Leonardo Candela

Abstract
This technical report lists the algorithms and the processes running on the gCube Statistical Manager service. The Statistical Manager (SM) is a set of web services that aid in the application of statistical computing and data mining to a variety of biological and marine related problems. By means of the integration with the D4Science e-Infrastructure it distributes algorithms "as a Service". Furthermore, it relies on the D4Science computational resources also to execute processes on large datasets.

Keywords
Data Analytics — Data Mining — Ecological Modelling

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Consiglio Nazionale delle Ricerche, Via G. Moruzzi 1, 56124, Pisa, Italy
***Corresponding author**: gianpaolo.coro@isti.cnr.it

Contents

1 Introduction 1
2 Algorithms 1
2.1 Data Clustering and Anomalies Detection 2
2.2 Classification 2
2.3 Climate 2
2.4 Correlation Analysis 2
2.5 Filtering 2
2.6 Occurrences 2
2.7 Performances Evaluation 3
2.8 Species Simulation 3
2.9 Training 3
2.10 Time Series 3
2.11 Taxa 3
2.12 Maps 3
2.13 Geo Processing 4
2.14 Bayesian Methods 4
2.15 Obis Observations Trends 4
2.16 Records Extraction 5
3 Conclusion 5
Acknowledgments 5
References 5

1. Introduction

gCube [1, 2] is a software system specifically conceived to enable the creation and operation of an innovative typology of infrastructure, i.e., an *Hybrid Data Infrastructure* [3], that by leveraging Grid [4], Cloud [5], Digital Library [6] and Service-orientation [7] principles and approaches is delivering a number of data management facilities *as-a-Service*. One of its distinguishing feature is the orientation to serve the needs of diverse Communities of Practice [8] by providing each of them with a dedicated, flexible, ready-to-use, web-based working environment, i.e., a *Virtual Research Environment* [9, 10].

gCube is endowed with the *Statistical Manager*, a service that offers facilities for efficiently and effectively executing a rich array of statistical data processing algorithms [11, 12]. The service relies on the distributed and elastic computing capacities offered by the underlying infrastructure. It offers a set of off-the-shelf algorithms including clustering algorithms such as DBScan. Moreover, it enables a simple integration and execution of user-defined algorithms expressed in a number of programming and scripting languages including R. It currently embeds a rich array of different algorithms ranging from Anomalies Detection, Classification, Clustering, Simulation, Training, Bayesian Methods, Trends, and many more. These algorithms are then executed on a distributed infrastructure by completely hiding the complexity of such an execution while ensuring robustness, throughput, fault-tolerance, and privacy.

This report lists and briefly describes the algorithms that have been developed and deployed so far by using the Statistical Manager service.

2. Algorithms

A rich array of algorithms have been developed and endow the current version of the Statistical Manager. They include data clustering and anomalies detection (cf. Sec. 2.1), classification (cf. Sec. 2.2), climate changes impact on species distribution (cf. Sec. 2.3), correlation analysis (cf. Sec. 2.4), filtering (cf. Sec. 2.5), species occurrence data processing (cf. Sec. 2.6), performances evaluation (cf. Sec. 2.7), species distribution simulation (cf. Sec. 2.8), training (cf. Sec. 2.9), time series processing (cf. Sec. 2.10), taxonomic data processing (cf. Sec. 2.11), maps (cf. Sec. 2.12), geo-referenced data processing (cf. Sec. 2.13), bayesian methods (cf. Sec. 2.14), OBIS Observations Trends (cf. Sec. 2.15), and records

extraction facilities (cf. Sec. 2.16).

2.1 Data Clustering and Anomalies Detection

Dbscan A clustering algorithm for real valued vectors that relies on the density-based spatial clustering of applications with noise (DBSCAN) algorithm. A maximum of 4000 points is allowed.

Kmeans A clustering algorithm for real valued vectors that relies on the k-means algorithm, i.e., a method aiming to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. A Maximum of 4000 points is allowed.

Lof Local Outlier Factor (LOF). A clustering algorithm for real valued vectors that relies on Local Outlier Factor algorithm, i.e., an algorithm for finding anomalous data points by measuring the local deviation of a given data point with respect to its neighbours. A Maximum of 4000 points is allowed.

Xmeans A clustering algorithm for occurrence points that relies on the X-Means algorithm, i.e., an extended version of the K-Means algorithm improved by an Improve-Structure part. A Maximum of 4000 points is allowed.

2.2 Classification

Feed Forward ANN Distribution A Bayesian method using a Feed Forward Neural Network to simulate a function from the features space (R^n) to R. A modeling algorithm that relies on Neural Networks to simulate a real valued function. It accepts as input a table containing the training dataset and some parameters affecting the algorithm behaviour such as the number of neurons, the learning threshold and the maximum number of iterations.

2.3 Climate

Bioclimate Hcaf A transducer algorithm that generates an Half-degree Cells Authority File (HCAF) dataset for a certain time frame, with environmental parameters used by the AquaMaps approach [12]. Evaluates the climatic changes impact on the variation of the ocean features contained in HCAF tables

Bioclimate Hspec A transducer algorithm that generates a table containing an estimate of species distributions per half-degree cell (HSPEC) in time. It evaluates the climatic changes impact on species presence.

Bioclimate Hspen A transducer algorithm that generates a table containing species envelops (HSPEN) in time, i.e., models capturing species tolerance with respect to environmental parameters, used by the AquaMaps approach [12]. Evaluates the climatic changes impact on the variation of the salinity values in several ranges of a set of species envelopes

Hcaf Interpolation Evaluates the climatic changes impact on species presence.

2.4 Correlation Analysis

Hrs An evaluator algorithm that calculates the Habitat Representativeness Score, i.e., an indicator of the assessment of whether a specific survey coverage or another environmental features dataset, contains data that are representative of all available habitat variable combinations in an area.

2.5 Filtering

Hcaf Filter An algorithm producing a HCAF table on a selected Bounding Box (default identifies Indonesia).

Hspen Filter An algorithm producing a HSPEN table containing only the selected species.

2.6 Occurrences

Absence Cells From Aquamaps An algorithm producing cells and features (HCAF) for a species containing absence points taken by an Aquamaps Distribution.

Occurrences Duplicates Deleter A transducer algorithm that produces a duplicate free table of species occurrence points where duplicates have been identified via user defined comparison thresholds. Works with up to 100,000 points.

Occurrences Intersector A transducer algorithm that produces a table of species occurrence points that are contained in both the two starting tables where points equivalence is identified via user defined comparison thresholds. Works with up to 10,000 points per table. Between two occurrence sets, it keeps the elements of the Right Set that are similar to elements in the Left Set.

Occurrences Marine Terrestrial A transducer algorithm that produces a table containing occurrence points by filtering them by type of area, i.e., by recognising whether they are marine or terrestrial. Works with up to 10,000 points per table.

Occurrences Merger A transducer algorithm that produces a duplicate-free table resulting from the union of two occurrence points tables where points equivalence is identified via user defined comparison thresholds. Works with up to 10,000 points per table. Between two Occurrence Sets, it enriches the Left Set with the elements of the Right Set that are not in the Left Set. Updates the elements of the Left Set with more recent elements in the Right Set. If one element in the Left Set corresponds to several recent elements in the Right Set, these will be all substituted to the element of the Left Set.

Occurrences Subtraction A transducer algorithm that produces a table resulting from the difference between two occurrence points tables where points equivalence is identified via user defined comparison thresholds. Works with up to 10,000 points per table. Between two Occurrence Sets, keeps the elements of the Left Set that are not similar to any element in the Right Set.

Presence Cells Generation An algorithm producing cells and features (HCAF) for a species containing presence points

2.7 Performances Evaluation

Discrepancy Analysis An evaluator algorithm that compares two tables containing real valued vectors. It drives the comparison by relying on a geographical distance threshold and a threshold for K-Statistic.

Quality Analysis An evaluator algorithm that assesses the effectiveness of a distribution model by computing the Receiver Operating Characteristics (ROC), the Area Under Curve (AUC) and the Accuracy of a model

2.8 Species Simulation

Aquamaps Native Algorithm for Native Distribution by AquaMaps. A distribution algorithm that generates a table containing species distribution probabilities on half-degree cells according to the AquaMaps approach for Native (Actual) distributions.

Aquamaps Native 2050 Algorithm for Native 2050 Distribution by AquaMaps. A distribution algorithm that generates a table containing species distribution probabilities on half-degree cells according to the AquaMaps approach with native distribution estimated for 2050.

Aquamaps Native Neuralnetwork Aquamaps Native Algorithm calculated by a Neural Network. A distribution algorithm that relies on Neural Networks and AquaMaps data for native distributions to generate a table containing species distribution probabilities on half-degree cells.

Aquamaps Suitable Algorithm for Suitable Distribution by AquaMaps. A distribution algorithm that generates a table containing species distribution probabilities on half-degree cells according to the AquaMaps approach for suitable (potential) distributions.

Aquamaps Suitable 2050 Algorithm for Suitable 2050 Distribution by AquaMaps. A distribution algorithm that generates a table containing species distribution probabilities on half-degree cells according to the AquaMaps approach for suitable (potential) distributions for the 2050 scenario.

Aquamaps Suitable Neuralnetwork Aquamaps Algorithm for Suitable Environment calculated by Neural Network. A distribution algorithm that relies on Neural Networks and AquaMaps data for suitable distributions to generate a table containing species distribution probabilities on half-degree cells.

2.9 Training

Aquamapsnn The AquaMaps model trained using a Feed Forward Neural Network. This is a method to train a generic Feed Forward Artificial Neural Network to be used by the AquaMaps Neural Network algorithm. Produces a trained neural network in the form of a compiled file which can be used later.

Feed Forward Ann A method to train a generic Feed Forward Artificial Neural Network in order to simulate a function from the features space (R^n) to R. It uses the Backpropagation method. The algorithm produces a trained neural network in the form of a compiled file which can be used in the FEED FORWARD NEURAL NETWORK DISTRIBUTION algorithm.

Hspen The AquaMaps HSPEN algorithm. It is a modeling algorithm that generates a table containing species envelops (HSPEN), i.e., models capturing species tolerance with respect to environmental parameters, to be used by the AquaMaps approach.

2.10 Time Series

Hcaf Interpolation Evaluates the climatic changes impact on species presence.

2.11 Taxa

Bionym An algorithm implementing BiOnym, a flexible workflow approach to taxon name matching [13] . The workflow allows to activate several taxa names matching algorithms and to get the list of possible transcriptions for a list of input raw species names with possible authorship indication.

Bionym Biodiv An algorithm implementing BiOnym oriented to Biodiversity Taxa Names Matching with a predefined and optimized workflow. This version applies in sequence the following Matchers: GSay(thr:0.6, maxRes:10), FuzzyMatcher (thr:0.6, maxRes:10), Levenshtein (thr:0.4, maxRes:10), Trigram (thr:0.4, maxRes:10). BiOnym is a flexible workflow approach to taxon name matching. The workflow allows to activate several taxa names matching algorithms and to get the list of possible transcriptions for a list of input raw species names with possible authorship indication.

Bionym Local A fast version of the algorithm implementing BiOnym, a flexible workflow approach to taxon name matching [13] . The workflow allows to activate several taxa names matching algorithms and to get the list of possible transcriptions for a list of input raw species names with possible authorship indication.

Fin Gsay Match An algorithm for GSAy Matching with respect to the Fishbase database.

Fin Taxa Match An algorithm for Taxa Matching with respect to the Fishbase database.

2.12 Maps

Discrepancy Analysis An evaluator algorithm that compares two tables containing real valued vectors. It drives the comparison by relying on a geographical distance threshold and a threshold for K-Statistic.

Maps Comparison An algorithm for comparing two OGC/NetCDF maps in seamless way to the user. The algorithm assesses the similarities between two geospatial maps by comparing

them in a point-to-point fashion. It accepts as input the two geospatial maps (via their UUIDs in the infrastructure spatial data repository - recoverable through the Geoexplorer portlet) and some parameters affecting the comparison such as the z-index, the time index, the comparison threshold.

Points To Map A transducer algorithm to produce a GIS map of points from a set of points with x,y coordinates indications. A maximum of 259,000 is allowed

Polygons To Map A transducer algorithm to produce a GIS map of filled polygons associated to x,y coordinates and a certain resolution. A maximum of 259,000 is allowed.

Species Map From Csquares A transducer algorithm to produce a GIS map from a probability distribution associated to a set of csquare codes. A maximum of 259,000 is allowed.

Species Map From Points A transducer algorithm to produce a GIS map from a probability distribution made up of x,y coordinates and a certain resolution. A maximum of 259,000 is allowed.

2.13 Geo Processing

Occurrence Enrichment An algorithm performing occurrences enrichment. It takes as input one table containing occurrence points for a set of species and a list of environmental layer, taken either from the e-infrastructure GeoNetwork (through the GeoExplorer application) or from direct HTTP links. The algorithm produces one table reporting the set of environmental values associated to the occurrence points.

Timeextraction An algorithm to extract a time series of values associated to a geospatial features repository (e.g., NETCDF, ASC, GeoTiff files). The algorithm analyses the time series and automatically searches for hidden periodicities. It produces one chart of the time series, one table containing the time series values and possibly the spectrogram.

Timeextraction Table An algorithm to extract a time series of values associated to a table containing geospatial information. The algorithm analyses the time series and automatically searches for hidden periodicities. It produces one chart of the time series, one table containing the time series values and possibly the spectrogram.

Xyextractor An algorithm to extract values associated to an environmental feature repository (e.g., NETCDF, ASC, GeoTiff files). A grid of points at a certain resolution is specified by the user and values are associated to the points from the environmental repository. It accepts as one geospatial repository ID (via their UUIDs in the infrastructure spatial data repository - recoverable through the Geoexplorer portlet) or a direct link to a file and the specification about time and space. The algorithm produces one table containing the values associated to the selected bounding box.

Xyextractor Table An algorithm to extract values associated to a table containing geospatial features (e.g., Vessel Routes, Species distribution maps). A grid of points at a certain resolution is specified by the user and values are associated to the points from the environmental repository. It accepts as one geospatial table and the specification about time and space. The algorithm produces one table containing the values associated to the selected bounding box.

Zextraction An algorithm to extract the Z values from a geospatial features repository (e.g., NETCDF, ASC, GeoTiff files). The algorithm analyses the repository and automatically extracts the Z values according to the resolution wanted by the user. It produces one chart of the Z values and one table containing the values.

Zextraction Table An algorithm to extract a time series of values associated to a table containing geospatial information. The algorithm analyses the time series and automatically searches for hidden periodicities. It produces one chart of the time series, one table containing the time series values and possibly the spectrogram.

2.14 Bayesian Methods

Lwr An algorithm to estimate Length-Weight relationship parameters for marine species, using Bayesian methods. Runs an R procedure. Based on the Cube-law theory.

2.15 Obis Observations Trends

Most Observed Species An algorithm producing a bar chart for the most observed species in a certain years range (with respect to the OBIS database)

Most Observed Taxa An algorithm producing a bar chart for the most observed taxa in a certain years range (with respect to the OBIS database)

Species Observations Per Area An algorithm producing a bar chart for the distribution of a species along a certain type of marine area (e.g. LME or MEOW)

Species Observation Lme Area Per Year Algorithm returning most observed species in a specific years range (data collected from OBIS database).

Species Observation Meow Area Per Year Algorithm returning most observed species in a specific years range (data collected from OBIS database).

Species Observations Trend Per Year An algorithm producing the trend of the observations for a certain species in a certain years range.

Taxonomy Observations Trend Per Year Algorithm returning most observations taxonomy trend in a specific years range (with respect to the OBIS database).

2.16 Records Extraction

Get Occurrences Algorithm An Algorithm that retrieves the occurrences from a data provided based on the given search options.

Get Taxa Algorithm An Algorithm that retrieves the taxon from a data provided based on the given search options.

3. Conclusion

This technical report lists the current set of algorithms endowing the gCube Statistical Manager.

The implementation of these algorithms is publicly available through the gCube web page hosted by the Ohloh service[1]. In particular, the Statistical Manager algorithms are in the data-analysis package.

Acknowledgments

The work reported has been partially supported by the *iMarine* project (FP7 of the European Commission, FP7-INFRASTRUCTURES-2011-2, Contract No. 283644), the *EUBrazilOpenBio* project (FP7 of the European Commission, FP7-ICT-2011.EU-Brazil, Contract No. 288754), and the *ENVRI* project (FP7 of the European Commission, FP7-INFRASTRUCTURES-2011-2.3.3, Contract No. 283465).

References

[1] gCube Development Team. gCube Website. `https://www.gcube-system.org`, 2008.

[2] Leonardo Candela, Donatella Castelli, and Pasquale Pagano. gCube v1.0: A Software System for Hybrid Data Infrastructures. Technical Report 2008-TR-035, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", CNR, 2008.

[3] Leonardo Candela, Donatella Castelli, and Pasquale Pagano. Managing big data through hybrid data infrastructures. *ERCIM News*, (89):37–38, 2012.

[4] Ian Foster and Carl Kesselman. *The Grid: Blueprint for a Future Computing Infrastructure*. Morgan-Kaufmann, 2004.

[5] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz, Andy Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, and Matei Zaharia. A view of cloud computing. *Communications of the ACM*, 53(4):50–58, April 2010.

[6] Leonardo Candela, Donatella Castelli, Nicola Ferro, Yannis Ioannidis, Georgia Koutrika, Carlo Meghini, Pasquale Pagano, Seamus Ross, Dagobert Soergel, Maristella Agosti, Milena Dobreva, Vivi Katifori, and Heiko Schuldt. *The DELOS Digital Library Reference Model - Foundations for Digital Libraries*. DELOS: a Network of Excellence on Digital Libraries, February 2008. ISSN 1818-8044 ISBN 2-912335-37-X.

[7] M. N. Huhns and M. P. Singh. Service-oriented computing: key concepts and principles. *IEEE Internet Computing*, 9:75–81, 2005.

[8] E. Wenger. *Communities of Practice: Learning, Meaning and Identity*. Cambridge University Press, 1998.

[9] Leonardo Candela, Donatella Castelli, and Pasquale Pagano. Virtual research environments: an overview and a research agenda. *Data Science Journal*, 12:GRDI75–GRDI81, 2013.

[10] Leonardo Candela, Donatella Castelli, and Pasquale Pagano. Making Virtual Research Environments in the Cloud a Reality: the gCube Approach. *ERCIM News*, (83):32–33, October 2010.

[11] Gianpaolo Coro, Pasquale Pagano, and Leonardo Candela. Providing statistical algorithms as-a-service. TDWG 2013 - Taxonomic Database Working Group 2013 (Firenze, 28-31 October 2013), 2013. Abstract.

[12] Leonardo Candela, Donatella Castelli, Gianpaolo Coro, Pasquale Pagano, and Fabio Sinibaldi. Species distribution modeling in the cloud. *Concurrency and Computation: Practice and Experience*, n/a:n/a, 2013. Article first published online: 11 July 2013 `http://onlinelibrary.wiley.com/doi/10.1002/cpe.3030/abstract`.

[13] Edward Vanden Berghe, Nicolas Bailly, Gianpaolo Coro, Fabio Fiorellato, Casey Aldemita, Anton Ellenbroek, and Pasquale Pagano. Bionym: a flexible workflow approach to taxon name matching. Technical Report 2014-TR-022, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", CNR, 2014.

[1] `https://www.openhub.net/p/gCube`