An introduction to machine learning and computational intelligence for time-series analysis: A case study for detection of climate tipping points

Seminar at Department of Meteorology

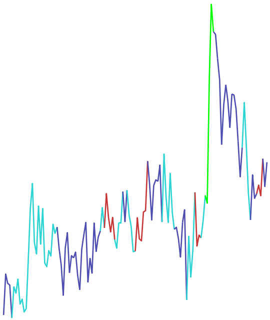
University of Reading

Javier Sánchez Monedero
jsanchez@uloyola.es

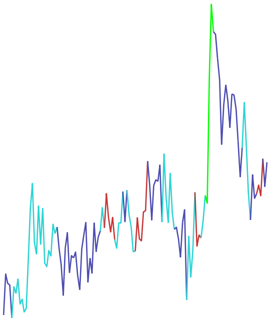
Dept. of Quantitative Methods, Universidad Loyola Andalucía.
Dept. of Computer Science and Numerical Analysis, University of Córdoba (former).
Learning and Artificial Neural Networks (AYRNA) research group.
http://www.uco.es/ayrna/

11th May 2017

Outline



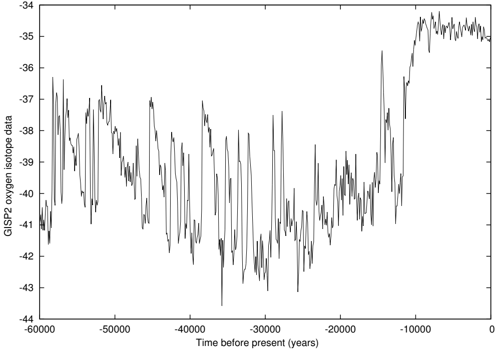
Outline



1. Introduction
2. Background: Machine Learning and Evolutionary Algorithms
3. Segmentation method
   - Evolutionary algorithm for time series segmentation
   - Detection of early warning signals in paleoclimate data
4. Prediction rules learning
5. Subpolar North Atlantic (SPG) temperature TPs
6. Conclusions and future work

Time series

- Time series data:
  - Ordered set of real-valued variables that are sampled or extracted on a continuous signal.
  - Many applications: finance, aerospace, entertainment, or climate.

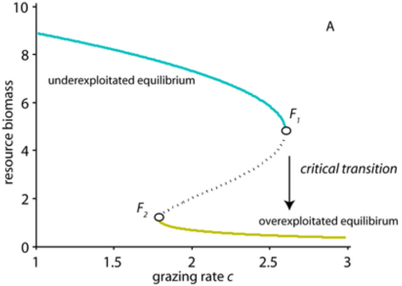


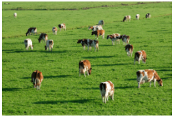
Tipping points

Tipping points Scheffer et al. [2009] and Lenton [2011]:

- A climate TP is a point where a dynamical system changes from one stable state to another stable state.
    - Small changes in a certain parameter of a time series can cause the transition.
- The detection of TPs in climate systems is currently a vividly discussed topic in the scientific community.

TP example

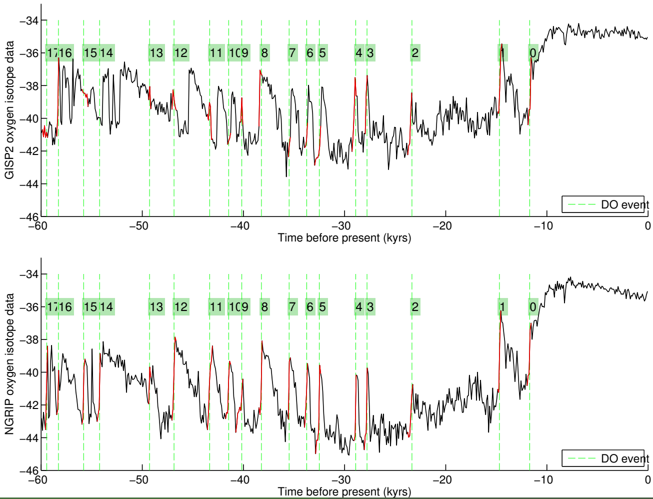




Source: Dakos et al. [2012].

Paleoclimate tipping points

Associated to Dansgaard-Oeschger (DO) events:



ESA-ACT project

Ariadna Project ID 13-9202: "Climate tipping points: Detection and analysis of patterns"

This project aimed at gaining knowledge about the typical statistical behaviour of paleoclimate data series before the occurrence of a tipping point to allow the detection of early warning signals (EWS) for upcoming abrupt climate transitions. Tools developped in the field of machine learning and soft computing were applied to Greenland ice core data to provide insights on the dynamical system under study.
http://www.esa.int/gsp/ACT/ess/projects/climate_tipping.html

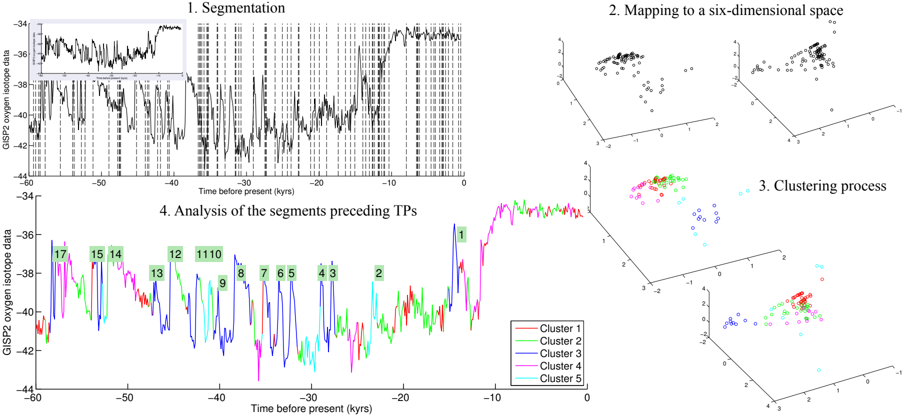
Objectives of the project

- Main objective: find and characterize patterns preceding TPs (Early Warning Signals) in paleoclimate data.
  - Detection of tipping points in climate system data series.
  - Obtain knowledge about the typical behaviour of climate systems before the occurrence of a tipping point.

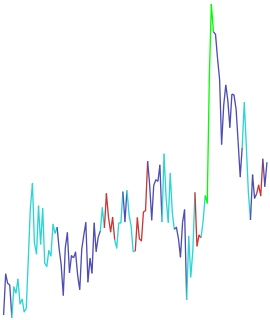
Time series segmentation

- As a first step, instead of directly designing prediction models, we apply a segmentation algorithm.
- Automatically, divides the time series into segments, and gives them a label.
- Visualises common patterns by attributing a (common) class label to such segments.

Proposal summary



Outline

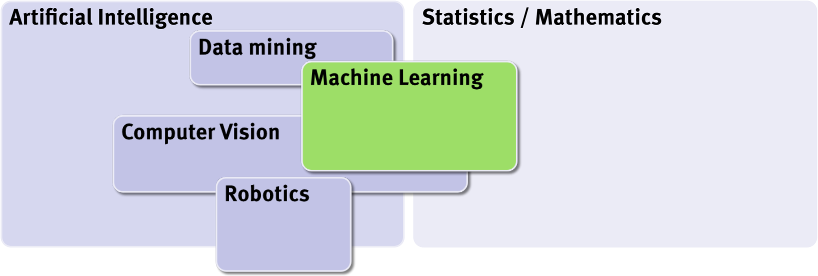


Machine Learning

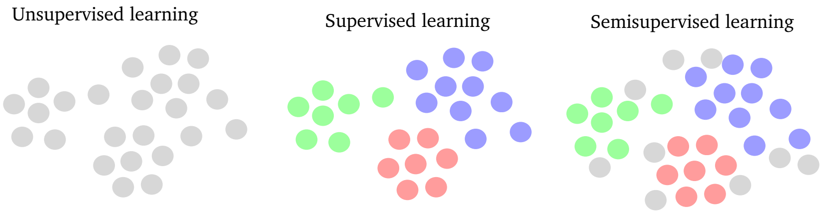
Machine Learning

In 1959, Arthur Samuel defined machine learning as a "Field of study that gives computers the ability to learn without being explicitly programmed"

Learn from data



Machine Learning Paradigms



Machine Learning methods

Wide collection of methods for regression/classification/clustering:

- Linear regression, logistic regresion, LDA, PCA... (classical statistics)
- Decision Trees, Neural Networks, Support Vector Machines, KLDA, KPCA...
- *k*-means, DBSCAN...
- ...

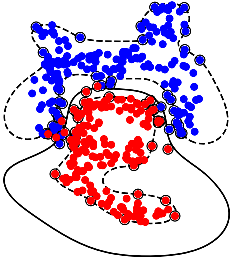


Figure: Kernel SVM example

Evolutionary Algorithms

Evolutionary Algorithms (EAs)

- Evolutionary Algorithms are non-determinism search algorithms with incorporates Darwin's theory of evolution by natural selection
- Solutions to a problem are evolved (mutation and crossover are applied) and the best individuals pass to the following generation
- EAs are widely used in optimisation problems.
- Online example http://www.boxcar2d.com/

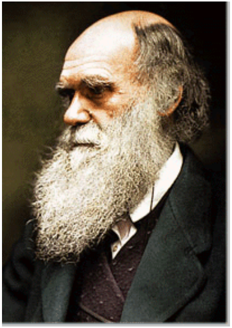


Figure: Charles Darwin

Evolutionary algorithm general flowchart

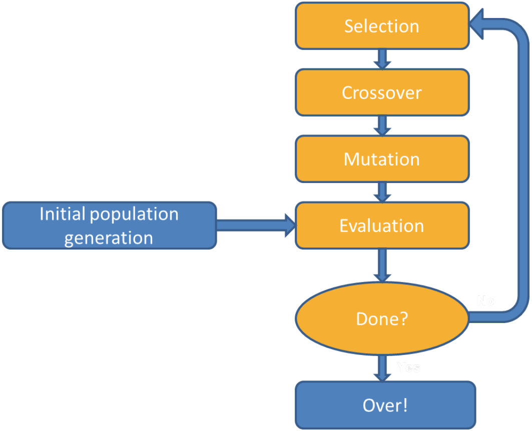


Figure: Source http://genetic.io/

Keys in evolutionary algorithms

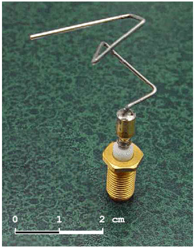
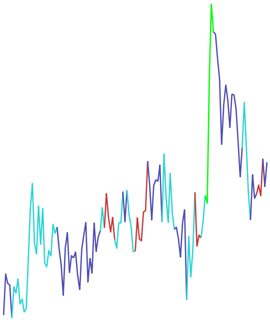


Figure: The 2006 NASA ST5 spacecraft antenna. This complicated shape was found by an evolutionary computer design program to create the best radiation pattern. It is known as an evolved antenna (Source Wikipedia).

Design of genetic algorithms:

- Chromosome design.
- Crossover and mutation operators and replacement strategy.
- Fitness function to evaluate the quality of the solutions.
- Optionally: local optimisation (hybrid algorithms).

Outline



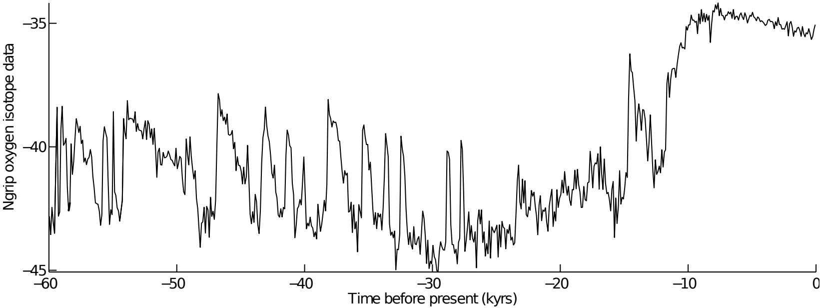
Approach

- Time-series segmentation: different alternatives (for a review, see Keogh et al. [2001]).
  - All of them try to provide a more compact representation of time series data dividing it into segments.
  - Some of them, try to match each segment with a predefined behaviour in order to obtain a high level representation.
- Inspired by Tseng et al. [2009], we propose a specific time series algorithm adapted to the climate TP detection, based on three main components:
  1. **A genetic algorithm (GA)** to define the cut points.
  2. **Mapping** each segment to six dimensional vector, based on some of the popular statistics considered for TP detection.
  3. A **clustering** algorithm to group these vectors.

  Also, we propose a method to evaluate the **quality** of the **segmentation**.

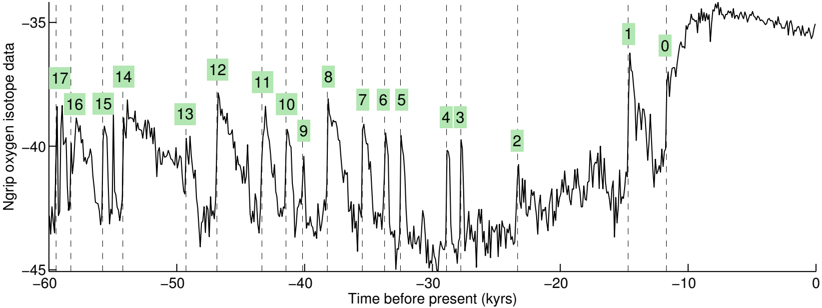
Approach

Original time series to be segmented:



Approach

DO events:



Genetic algorithm: pseudocode

Time series segmentation:

Require: Time series.
Ensure: Best segmentation of the time series.

```
Generate a random population of t time series segmentations.
Evaluate all segmentations of the initial population by using
the fitness function (mapping+clustering).
while not Stop Condition do
    Store a copy of the best segmentation.
    Select parent segmentations from current population.
    Generate offsping: apply crossover and mutation to
    construct new candidate segmentations.
    Evaluate the fitness of the offsping segmentations.
    Replace current population with offspring.
end while
return  Best segmentation from final population
```

Genetic algorithm: chromosome

a) Example chromosome. Each position represents an index of a time series value.

4	8	13	18

b) Segments of the time series resulting from the chromosome.

1	2	3	4

Segment 1

4	5	6	7	8

Segment 2

8	9	10	11	12	13

Segment 3

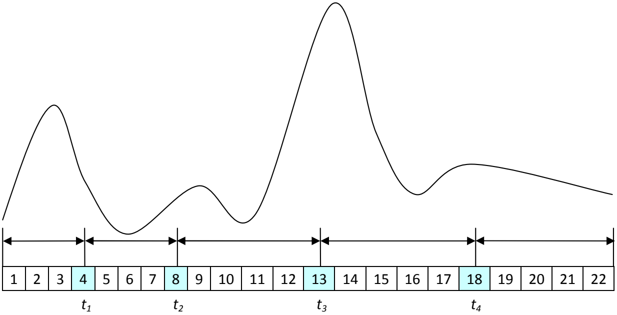
13	14	15	16	17	18

Segment 4

18	19	20	21	22

Segment 5

c) Corresponding segmentation and time series. The characteristics of each segment will be obtained for the corresponding part of the time series.



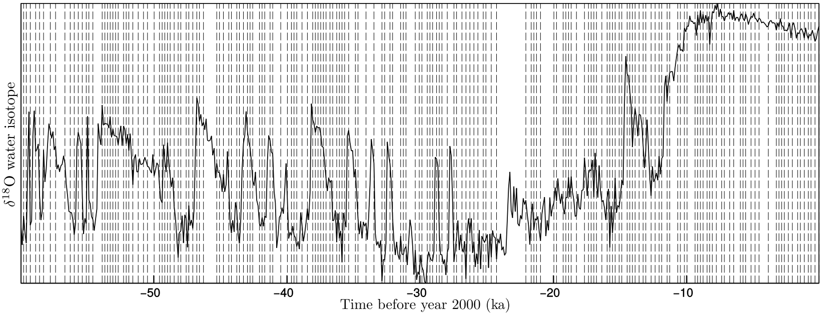
Genetic algorithm: fitness function

- Segments may have different length → map all the segments to the same dimensional space.
- Six statistical metrics are measured for all the segments of a segmentation:
  1. Variance (S_s^2)*.
  2. Skewness (γ_{1s}).
  3. Kurtosis (γ_{2s}).
  4. Slope of a linear regression over the points of the segment (a_s).
  5. Mean Squared Error (MSE_s).
  6. Autocorrelation coefficient (AC_s)*.

  *: Frequently found in TP detection literature.
- After mapping, similarities between segments can be obtained and we can search for groups of segments.

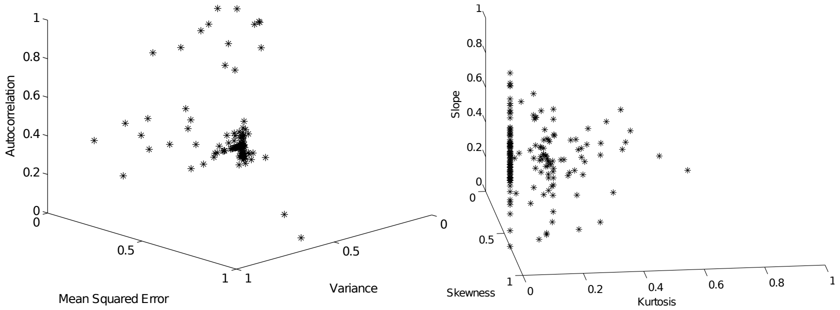
Genetic algorithm: fitness function

This process maps a segmentation (Cutting points established by the GA)...



Genetic algorithm: fitness function

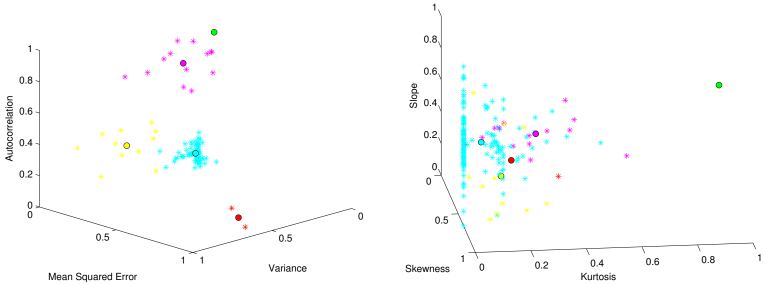
...to a six dimensional space (each segment is a point):



where standard clustering can be applied (K-means).

Genetic algorithm: fitness function

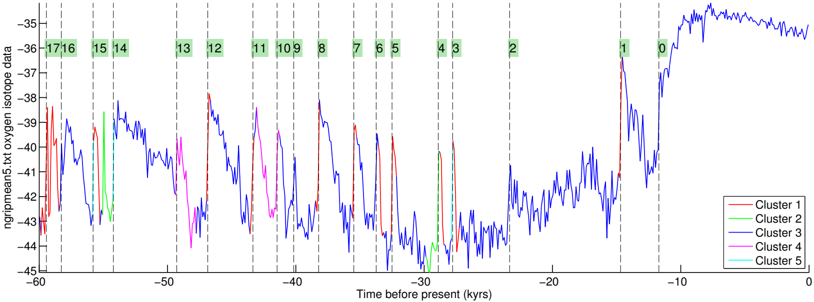
K-means groups points and finds a centroid for each group:



The fitness depends on the clustering quality (we test 10 fitness functions in Pérez-Ortiz et al. [2016]).

Approach

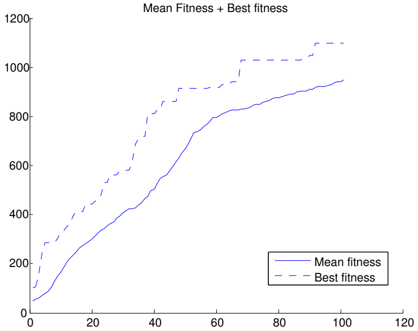
Clustering results (consecutive segments are grouped together):



Genetic algorithm: fitness function

Literature of clustering quality metrics. In Pérez-Ortiz et al. [2016], the Caliński and Harabasz index was found to be the most robust:

$$CH = \frac{\mathtt{Tr}(\mathbf{S}_B) \cdot (N-k)}{\mathtt{Tr}(\mathbf{S}_W) \cdot (k-1)} \approx \frac{\mathtt{Distance\ between\ clusters}}{\mathtt{Within\ cluster\ distance}}$$



Genetic algorithm: crossover

- Crossover: mix two parent segmentations to obtain a new one (hopefully, with the good properties of both) (≈ 0.8 of individuals).

10	15	18	26	33	36	47	52	59	62	68	75	80	84	88	92	95	99
15	20	23	27	32	36	45	48	55	65	71	77	81	86	91	96	98	99

a) Before applying crossover operator.

10	15	18	26	33	36	47	52	59	65	71	77	81	86	91	96	98	99
15	20	23	27	32	36	45	48	55	62	68	75	80	84	88	92	95	99

b) After applying crossover operator. The crossover point was randomly decided to be 60.

Genetic algorithm: mutation

- Mutation: randomly modify a given segmentation of some individuals (≈ 0.2).

10	15	18	26	33	36	47	52	59

10	15	26	33	36	47	59

a) Mutation operator: remove two cut points (18 and 52).

10	15	18	26	33	36	47	52	59

10	15	18	26	33	36	39	47	52	59

b) Mutation operator: add a cut point: 39.

10	15	18	26	33	36	47	52	59

9	15	16	26	31	32	47	52	53

c) Mutation operator: randomly move cut-points to the left.

10	15	18	26	33	36	47	52	59

10	17	20	29	33	36	47	54	63

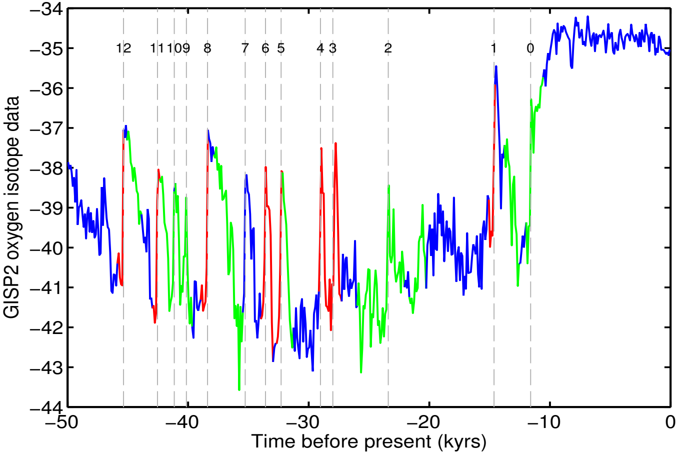
d) Mutation operator: randomly move cut-points to the right.

First experiment

- Two datasets (both from δ^{18}O ice core data):
  - GISP2 Greenland Ice Sheet Project Two.
  - NGRIP North Greenland Ice Core Project.
- Very important: no prior knowledge of TPs is given to the algorithm.
- Parameters of the algorithm are fixed for all the problems and the GA consistently returns good results ($k = 5$).
- The results can be different depending on the seed value, so the GA is run 10 times with different seeds to evaluate and remove the dependence on the seed value.
- 5-point running average reduced short-term fluctuations.
- For more information see Nikolaou et al. [2015].

Results of the first experiment

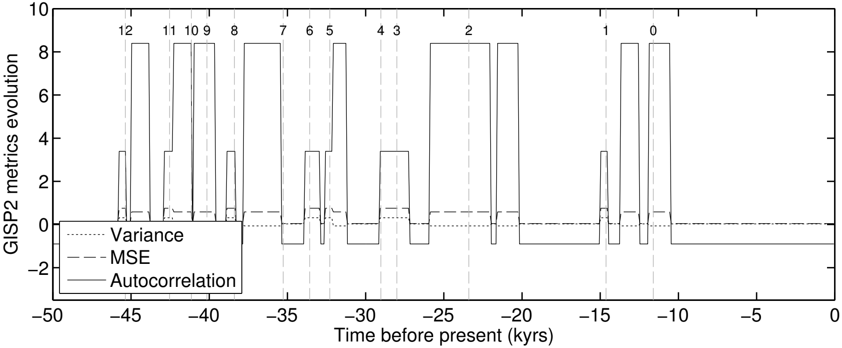
One given segmentation (GISP2, seed 10, DO events: $\mathcal{C}_1$ and $\mathcal{C}_4$):



	Variance	Skewness	Kurtosis	Slope LR	MSE	Autocorrelation
$\mathcal{C}_1$ (red)	2.734	0.313	-1.455	-0.004	0.753	3.391
$\mathcal{C}_2$ (green)	1.188	-0.068	-0.804	0.001	0.584	8.395
$\mathcal{C}_3$ (blue)	0.124	0.025	-1.716	0.001	0.045	-0.904
$\mathcal{C}_4$ (magenta)	2.189	0.445	-1.315	-0.001	2.159	-0.915
$\mathcal{C}_5$ (cyan)	0.843	-1.702	2.543	-9.5e-05	0.839	11.001

Results of the first experiment

DO events are consistently detected by the values of the centroids of their preceding segments:



Results of the first experiment

Detection accuracy (all seeds, all datasets):

DO event	Detectability success (%)	
	GISP2	NGRIP
End of Younger Dryas	80	100
1	90	90
2	70	50
3	30	20
4	90	90
5	30	50
6	60	20
7	70	50
8	90	100
9	0	0
10	50	70
11	70	70
12	80	80

Discussion of the first experiment

Findings:

1. DO events are usually grouped in two or three classes.
2. Early warning signals (EWSs) of DO events are found in the form of an increase in autocorrelation, variance, and MSE.
3. The increase in mean square error (MSE) and variance are suggested as an indicator of abrupt climate change (non-linear segment).
4. The increase in the autocorrelation coefficient (AC) cannot be solely used as EWS (AC, Variance and MSE have to analysed together).
5. Skewness and kurtosis are not useful to predict abrupt climate changes, which is in agreement with other results from previous studies.

Discussion of the first experiment

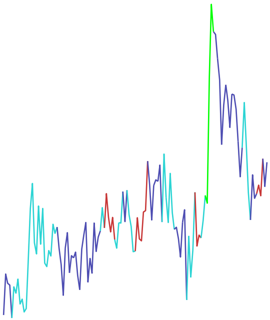
Low detectability for some DO events:

- They were triggered by noise close to a bifurcation.
- In agreement with the stochastic resonance model, where the climate system is responding to weak periodic forcing + noise.

Limitations:

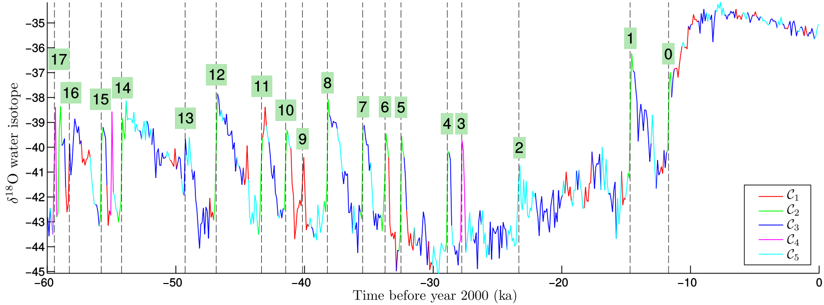
1. Small segments: easy to be clustered (skewness=0, MSE=0), good for fitness (m high).
2. Applying discovered EWSs for future data: length of the segment?.

Outline

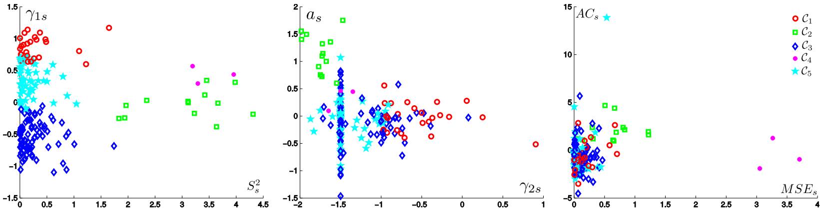


Improved algorithm

In Pérez-Ortiz et al. [2014, 2016] we improved the segmentation algorithm and presented an initial strategy for TP prediction.



Variables impact analysis



Dataset and experimental design

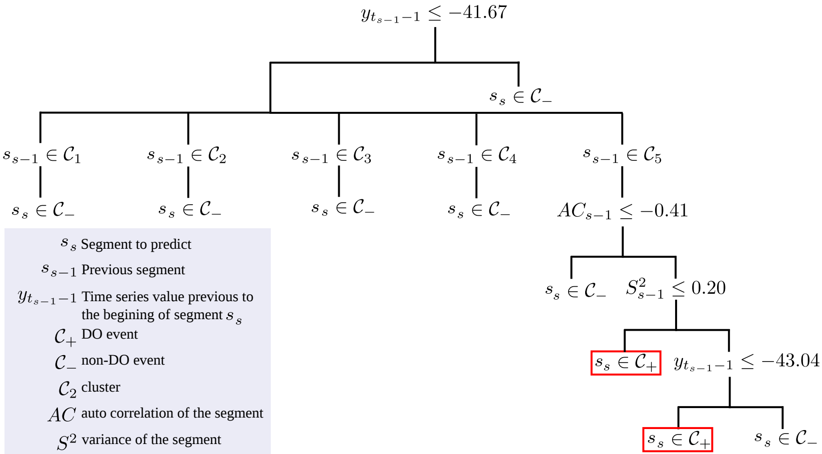
Dataset:

- Dependent/target variable is the class of next segment s_s ($\mathcal{C}_+$ for TP and $\mathcal{C}_-$ otherwise).
- 10 independent variables used for the prediction:
  - **Dynamic window** variables:
    - 6 statistical descriptors of s_{s-1}.
    - Class label of s_{s-1}.
  - **Fixed window** variables: three time instants before s_s (i.e., $y_{t_{s-1}-1}, y_{t_{s-1}-2}, y_{t_{s-1}-3}$).

Experimental design: leave-one-out approach due to the small number of TP segments.

Decision tree for NGRIP

Prediction phase: we use the C4.5 decision tree to learn a rule for predicting a future segment class:



- s_s Segment to predict
- s_{s-1} Previous segment
- $y_{t_{s-1}-1}$ Time series value previous to the begining of segment s_s
- $\mathcal{C}_+$ DO event
- $\mathcal{C}_-$ non-DO event
- $\mathcal{C}_2$ cluster
- AC auto correlation of the segment
- S^2 variance of the segment

Performance results for NGRIP

	actual class (observation)	
predicted class (expectation)	7 (true positive)	0 (false positive)
	11 (false negative)	165 (true negative)

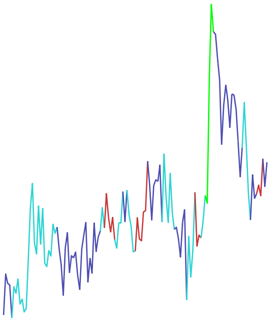
100% of non-TP detected and **38.9% TP** detected... but we are simultaneously addressing the detection of 18 events with the same method (and fixed method's parameters).

Variables relevance in EWS

Variables' relevance to predict DO (i.e. EWS):

- The most discriminative one is the previous value of the time series: $y_{t_{s-1}-1}$.
- DOs usually preceded by $\mathcal{C}_5$.
- NGRIP: Autocorrelation and variance
- GISP2: kurtosis, skewness and slope

Outline



Abrupt cooling events in the SPG

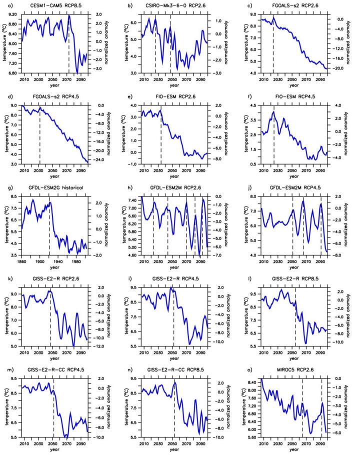
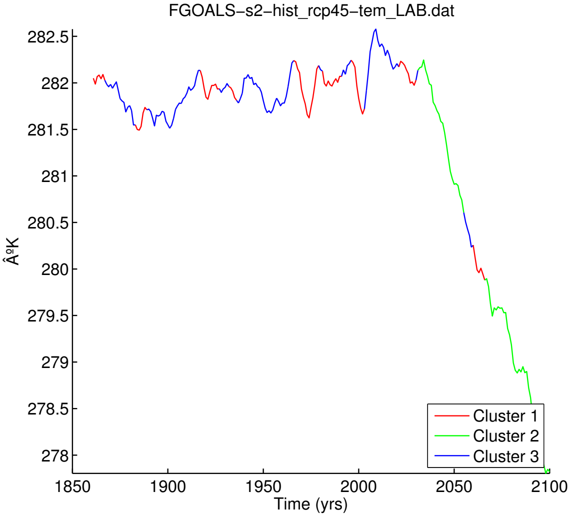
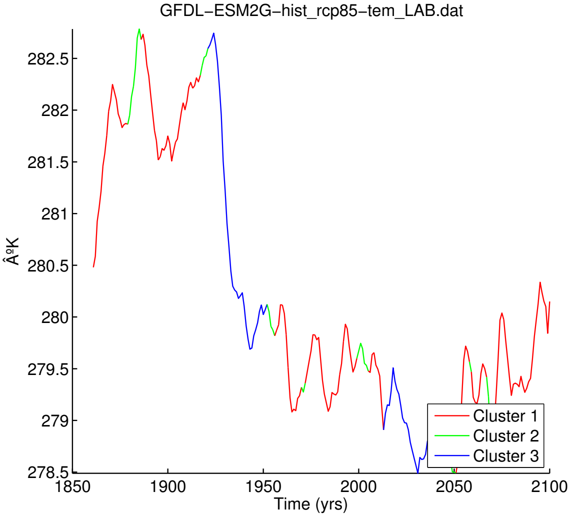


Figure: SST evolution (°C) in the SPG models showing an abrupt cooling (source Sgubin et al. [2017])

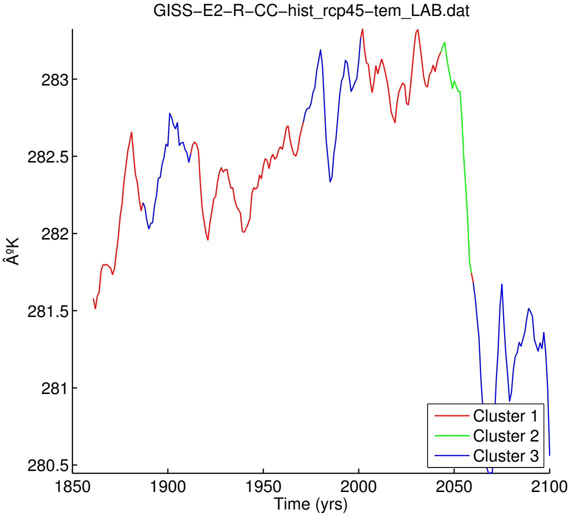
Premilinary segmentation



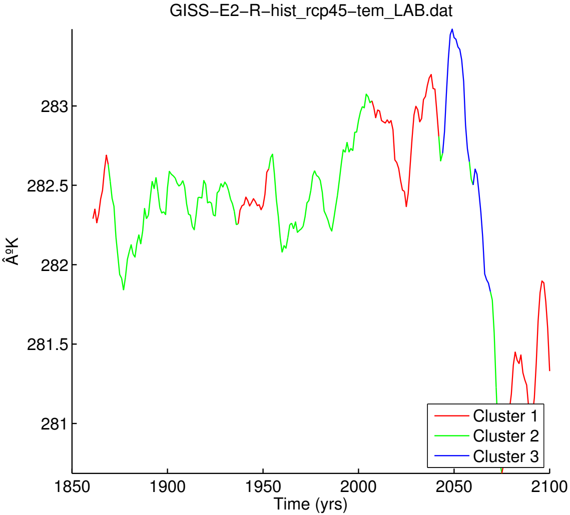
Premilinary segmentation



Premilinary segmentation



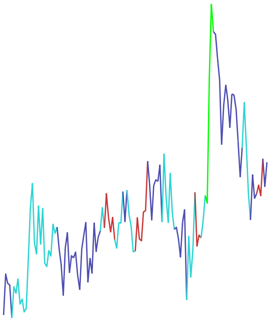
Premilinary segmentation



Premilinary segmentation conclusions

- We have run the algorithm of Pérez-Ortiz et al. [2016] reducing the number of clusters to 3.
- We have promising results in some series and random results in others.
- Some segments are too small:
  - Study how to restrict the minimum size of the segments: fitness, restrictions, group similar segments...
- We have to move from one time series with 18 TPs to several time series mainly with one TP:
  - This suggest to modify the algorithm to simultaneously segment several series: 1) individually segment the series 2) perform clustering with all the series.

Outline



1. Introduction
2. Background: Machine Learning and Evolutionary Algorithms
3. Segmentation method
   - Evolutionary algorithm for time series segmentation
   - Detection of early warning signals in paleoclimate data
4. Prediction rules learning
5. Subpolar North Atlantic (SPG) temperature TPs
6. Conclusions and future work

Conclusions

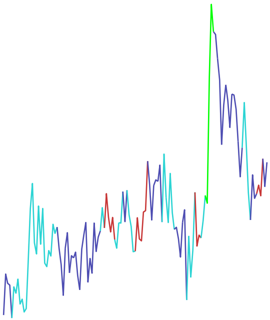
- Novel GA for TS segmentation with differentiating characteristics.
- Tool for visualization of TS from a segment-based perspective.
- Increasing MSE as an EWS.
- Proposal for prediction method.

Future work

Future research:

- Comparison to state-of-the-art recent approaches (mainly based on statistical methods).
- Consider other models for the prediction.
- New restrictions for the genetic algorithm (e.g. minimum segment length, fusion of consecutive segments of the same classes...).
- SPG abrupt cooling case:
  - Simultaneous segmentation of several time series.
  - Specific statistical descriptors.
  - Evaluate different fitness functions.
  - Incorporate the prediction accuracy in the fitness function.

Outline



7 References

Associated publications I

Pérez-Ortiz, M., Gutiérrez, P., Sánchez-Monedero, J., Hervás-Martínez, C., Nikolaou, A., Dicaire, I. and Fernández-Navarro, F. *Time Series Segmentation of Paleoclimate Tipping Points by an Evolutionary Algorithm* Hybrid Artificial Intelligence Systems (HAIS 2014) Springer International Publishing, 2014, Vol. 8480, pp. 318-329

Pérez-Ortiz, M., Durán-Rosal, A., Gutiérrez, P., Sánchez-Monedero, J., Nikolaou, A., Fernández-Navarro, F. and Hervás-Martínez, C. *On the use of evolutionary time series analysis for segmenting paleoclimate data* Neurocomputing, 2016, Vol. Accepted

Nikolaou, A., Gutiérrez, P. A., Durán, A., Dicaire, I., Fernández-Navarro, F. and Hervás-Martínez, C. *Detection of early warning signals in paleoclimate data using a genetic time series segmentation algorithm* Climate Dynamics, 2015, Vol. 44(7), pp. 1919-1933

References I

Vasilis Dakos, Stephen R Carpenter, William A Brock, Aaron M Ellison, Vishwesha Guttal, Anthony R Ives, Sonia Kefi, Valerie Livina, David A Seekell, Egbert H Van Nes, et al. Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data. *PLoS One*, 7(7):e41010, 2012.

Eamonn Keogh, Selina Chu, David Hart, and Michael Pazzani. An online algorithm for segmenting time series. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 289–296. IEEE, 2001.

Timothy M. Lenton. Early warning of climate tipping points. *Nature Climate Change*, 1:201–209, 2011.

Athanasia Nikolaou, Pedro Antonio Gutiérrez, Antonio Durán, Isabelle Dicaire, Francisco Fernández-Navarro, and César Hervás-Martínez. Detection of early warning signals in paleoclimate data using a genetic time series segmentation algorithm. *Climate Dynamics*, 44(7):1919–1933, 2015. ISSN 1432-0894.

M. Pérez-Ortiz, P.A. Gutiérrez, J. Sánchez-Monedero, C. Hervás-Martínez, Athanasia Nikolaou, Isabelle Dicaire, and Francisco Fernández-Navarro. Time series segmentation of paleoclimate tipping points by an evolutionary algorithm. In *Hybrid Artificial Intelligence Systems*, volume 8480 of *Lecture Notes in Computer Science*, pages 318–329. Springer International Publishing, 2014. ISBN 978-3-319-07616-4.

M. Pérez-Ortiz, A. Durán-Rosal, P.A. Gutiérrez, J. Sánchez-Monedero, A. Nikolaou, F. Fernández-Navarro, and C. Hervás-Martínez. On the use of evolutionary time series analysis for segmenting paleoclimate data. *Neurocomputing*, Accepted, 2016.

Marten Scheffer, Jordi Bascompte, William A Brock, Victor Brovkin, Stephen R Carpenter, Vasilis Dakos, Hermann Held, Egbert H Van Nes, Max Rietkerk, and George Sugihara. Early-warning signals for critical transitions. *Nature*, 461(7260):53–59, 2009.

Giovanni Sgubin, Didier Swingedouw, Sybren Drijfhout, Yannick Mary, and Amine Bennabi. Abrupt cooling over the north atlantic in modern climate models. *Nature Communications*, 8(14375):1–12, 2017.

Vincent S Tseng, Chun-Hao Chen, Pai-Chieh Huang, and Tzung-Pei Hong. Cluster-based genetic segmentation of time series with dwt. *Pattern Recognition Letters*, 30(13):1190–1197, 2009.

Questions?

Thanks!