To be or not to be real: fractal analysis of data streams from a regional climate change model

Santiago A. Nunes
University of São Paulo - Brazil
bynael@grad.icmc.usp.br

Luciana A. S. Romani
Embrapa Agriculture Informatics - Brazil
luciana@cnptia.embrapa.br

Priscila P. Coltri
University of Campinas - Brazil
pcoltri@cpa.unicamp.br

Ana M. H. Avila
University of Campinas - Brazil
avila@cpa.unicamp.br

Agma J. M. Traina
University of São Paulo - Brazil
agma@icmc.usp.br

Elaine P. M. Sousa
University of São Paulo - Brazil
parros@icmc.usp.br

ABSTRACT

This paper proposes a new analysis process aimed at discriminating the temporal behavior of the data generated by climate models from the real climate observations gathered from ground-based meteorological stations. Our approach combines fractal data analysis and the monitoring of the real and the model-generated data streams to detect deviations considering the intrinsic correlation among climate time series. Experimental studies showed that our approach can discriminate the data either as real or as generated by a model. Those results suggest that there are yet space to improve the climate change models, and that the fractal-based concepts may contribute in this improvement.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Data Stream, Fractal Analysis, Climate Data

1. INTRODUCTION

A major challenge posed to researchers in the 21st Century regards the world climate changes as a consequence of the global warming. Since it affects the whole planet, the Intergovernmental Panel on Climate Change (IPCC) was created to evaluate and analyze the data concerning such changes and to propose alternatives to deal with the current and upcoming problems caused by climate changes.

Among several challenges on climate changes, food security and sustainable development are certainly the most important. Overall, agriculture is strongly related to both issues, which makes this activity strategic in terms of economy and society to the whole world. In Brazil, the regional model Eta-CPTEC is under development in the Brazilian Center for Weather Forecasts and Climate Studies (CPTEC) since 1996 to provide weather forecasts over South America [1]. As scenarios show different changes in the future, consequences in agricultural production may cause different impacts for regions. Thus, the better performance of the model, the better the work of decision makers, as they can count on more reliable information. Knowing properly the performance of the Eta-CPTEC regional model may encourage important research on climate changes.

In this context, this paper proposes an analysis process to evaluate the quality of models output. Our approach deals with multiple time series as a multidimensional data stream. Therefore, it is possible to integrate multiple climate variables in a unified analysis process. In particular, we combine fractal-based data stream monitoring for change detection, considering the intrinsic correlation among time series defined by different climate variables.

2. BACKGROUND AND RELATED WORK

The Correlation Fractal Dimension D_2 [3] can be applied in data analysis to estimate the intrinsic dimension of real datasets exhibiting fractal behavior.The D_2 measures the non-uniform behavior of real data considering both linear and nonlinear attribute correlations [2].

In [4], the authors propose a technique and the corresponding algorithm, named *SID-meter*, to detect changes in evolving data streams, based on the information of intrinsic behavior provided by the fractal dimension D_2. The *SID-meter* continuously measures D_2 over time in order to monitor the evolving behavior of data, such that significant variations in successive measures of D_2 can indicate changes in the intrinsic characteristics of the data. *SID-meter* is used in the analysis process we propose.

3. THE ANALYSIS PROCESS

In order to assist domain specialists in studying climate data, mainly in evaluating the behavior of data from climate change models in comparison with the behavior of real data, we propose an analysis process based on the information provided by the fractal dimension, continuously calculated according to the *SID-meter* approach. The proposed process was implemented in a tool named *ClimFractal*, developed to support visual analysis of large climate datasets.

The analysis process works as follows: time series defined by different climate variables are combined to create a multidimensional data stream. Thus, each event e_i is defined as a set of climate measures of different variables collected (from meteorological stations) or generated (by climate models), considering specific location and instant of time. The *SID-meter* processes the stream and computes the successive values of the D_2 over time, considering the user-defined sliding window. By analyzing the variation of D_2 it is possible to identify the existence of correlations among the climate variables and how such correlations evolves, i.e., it is possible to evaluate the temporal behavior of the data. The fractal-based analysis can be applied to real climate time series and to time series generated by climate models as well, in order to compare them over time. From the specialist perspective, the whole process is realized through the *ClimFractal* interface, from the choice of the region and the period of interest to the visual analysis of the results.

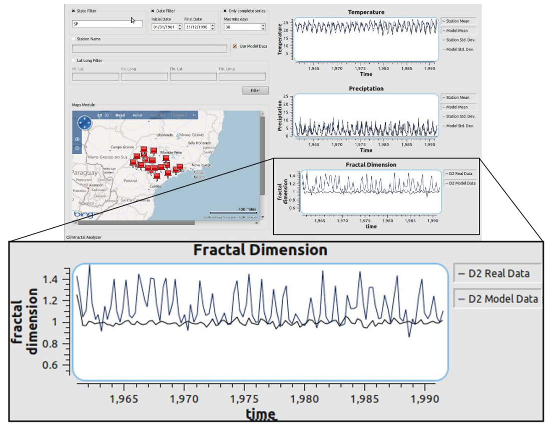

Figure 1: *ClimFractal* - D_2 measures.

4. EXPERIMENTAL RESULTS

We have applied our analysis process in two sets of climate time series from a climate database, as follows:

1. Real data - climate time series[1] containing daily measurements of precipitation and mean temperature obtained from 24 ground-based meteorological stations of the state of São Paulo - Brazil, from 1961 to 1990.

2. Data from regional climate change model - climate time series from Eta-CPTEC regional model containing daily measurements of precipitation and mean temperature from 1961 to 1990, considering the same study area.

A two-dimensional data stream composed of the attributes `precipitation` and `mean temperature` was defined for each dataset. We performed the experiments with two configurations of sliding windows: a 3-month window with D_2 updated each month, and a 6-month window with D_2 updated each two months. Due to lack of space, we only present the results from the second configuration.

The successive values of D_2 computed by the *SID-meter* for each stream as well as the statistical measurements are visualized in the *ClimFractal* interface (Figure 1). It can be observed that both datasets showed a similar pattern of general behavior, i.e., the D_2 increases in some periods for real data as well as for the Eta-CPTEC model outputs.

On the other hand, the values of D_2 are higher for the data estimated by the Eta-CPTEC model (D_2varying from 1 to 1.5). It indicates that there is less correlation between variables from the climate model than between the variables from real data. Even though, the similar patterns observed on both curves indicate that the model can represent the general climate in the period 1961 to 1990 for the study area, but it leaves out the finer details of the intrinsic characteristics of the climate system.

A model's output can also be analyzed by the method of Wilmot [5], that is an index of agreement (d) which reflects the degree to which the observed variable is accurately estimated by the predicted one. It is not a measure of correlation in the formal sense but a measure of the degree to which a model's predictions are error free, varying between 0 (complete disagreement) and 1 (perfect agreement).

The similarity of the D_2 graphs generated using real data and outputs from Eta-CPTEC can also be observed when the conventional method ([5]) is applied in both data. The index d presented a suitable similarity pattern for real and Eta-CPTEC's data. Most of indexes d for precipitation and temperature reached values greater than 0.6, indicating agreement between real and simulated data.

[1]Provided by Agritempo - www.agritempo.br

5. CONCLUSIONS

Initial results showed that our fractal-based analysis process can discriminate the data either as real or as generated by a model, as the intrinsic correlations between climate variables identified in real data (and confirmed by the specialists) are slightly different in data from the climate model. Those results suggest that there are yet space to improve the climate change models and that the fractal-based concepts may be useful. Analyzing the output of the climate change model can help the specialists to understand and improve the model, contributing to research on climate changes and their effects, such as the impacts on agriculture.

6. ACKNOWLEDGMENTS

Authors thank Embrapa, Fapesp, Microsoft Research, CNPq and Capes for the financial support and Agritempo and CPTEC/INPE for the climate data used in this work.

7. ADDITIONAL AUTHORS

Caetano Traina (University of São Paulo - Brazil, email: `caetano@icmc.usp.br`), Sin C. Chou (CPTEC/INPE - Brazil, email: `chou@cptec.inpe.br`).

8. REFERENCES

[1] S. C. Chou et al. Downscaling of south america present climate driven by 4-member hadcm3 runs. *Springer - ClimDyn*, 25:33–59, 2007.

[2] C. Faloutsos and I. Kamel. Beyond uniformity and independence: Analysis of r-trees using the concept of fractal dimension. In *ACM PODS*, pages 4–13, 1994.

[3] M. Schroeder. *Fractals, Chaos, Power Laws.* W. H. Freeman and Company, 1991.

[4] E. P. M. Sousa, C. Traina, A. J. M. Traina, and C. Faloutsos. Measuring evolving data streams' behavior through their intrinsic dimension. *New Generation Computing Journal*, 25:33–59, 2007.

[5] C. Willmott et al. Statistics for the evaluation and comparison of models. *JOURNAL OF GEOPHYSICAL RESEARCH*, 90:8995–9005, Sept. 1985.