University of Cordoba.
Department of Computer Science and Numerical Analysis.
Rabanales Campus, "Albert Einstein" building, 3rd floor.
14071 – Córdoba (Spain).
Ph. +34 957218349 Fax +34 957218630

Energies. Special Issue "Soft Computing Techniques in Energy System"

Cordoba, Spain, November 17, 2020.

Dear Editor,

Please find attached a manuscript, entitled *"Building Suitable Datasets for Soft Computing and Machine Learning Techniques from Meteorological Data Integration: A case study for predicting significant wave height"* submitted for possible publication in Energies, Special Issue "Soft Computing Techniques in Energy System".

We state that it is our original work, and it has not been submitted, nor it is under consideration in any other forum.

Declaration for any conflicts of interest

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Novelty of this work

This paper presents a new open source tool named SPAMDA for the creation, in an easy way, without the need of knowledge in programming languages and using a graphical interface, of datasets integrated by meteorological variables from two sources of information. These datasets can

later be used with Machine Learning (ML) and Soft Computing (SC) methodologies in classification and prediction studies, such as prediction of significant wave height and energy flux in coastal and ocean areas. This tool, among other functionalities (Section 3 of the paper describes concrete and comprehensive information of the developed software), allows managing and storing datasets with different configurations, in terms of predictive and objective variables and different temporal resolution, integrating two different sources of information available for the scientific community (NDBC and NNRP). Note that this data integration process has an extensive casuistry giving rise to incomplete datasets and it requires laborious pre-processing procedures, that implies a lot of time and effort by the researchers, that usually lead to errors (Appendix A of this work shows the casuistry and some problems of the mentioned data integration process).

SPAMDA has been developed to integrate these two sources of information by means of an user-friendly GUI, and it facilitates the management and storage of buoys and reanalysis data, time and geographical coordinates conversion of both kind of data, missing values handling (dates or measurements not recorded), pre-processing tasks, the generation of datasets to be used by ML and SC techniques in prediction tasks (classification or regression) including time horizon selection and output discretisation, among others. Therefore, the developed software tool will allow researchers focusing on the study of the meteorological aspects of the observations, providing a series of novelties and functionalities briefly described at the end of the Introduction section.

On the other hand, increasingly, ML and SC techniques are been applied to tackle wave characterisation, feasibility assessment of wave energy extraction and designing offshore structures, and this software allows the creation of datasets ready to use with the Weka Data Mining tool (integrated in the software itself) or in .csv format for being used with any other prediction tool. So, we hope that SPAMDA can be used by researchers when carrying out environmental modelling related to energy, atmospheric or oceanic studies or design of production systems as WECs (Wave Energy Converters, that need a proper prediction of waves in order to maximise the wave energy extraction), among others, using the two sources of information described. Moreover, it could manage other sources of reanalysis data with different spatial and temporal resolution with a reasonable development effort.

Should you have any doubts or question about this submission, please do not hesitate in contacting us,

Best regards,

Antonio M. Gómez-Orellana, Juan C. Fernández, Manuel Dorado-Moreno, Pedro A. Gutiérrez, César Hervás-Martínez

Department of Computer Science and Numerical Analysis, University of Cordoba, 14071, Córdoba, Spain.

email: {am.gomez@uco.es, jfcaballero@uco.es, manuel.dorado@uco.es, pagutierrez@uco.es, chervas@uco.es}