Environmental Modelling & Software 100 (2018) 24–32

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

A system for airport weather forecasting based on circular regression trees

Pablo Rozas Larraondo [a, *], Iñaki Inza [b], Jose A. Lozano [b, c]

[a] National Computational Infrastructure, Building 143, Australian National University, Ward Road, ACT, 2601, Australia
[b] Intelligent Systems Group, Computer Science Faculty, University of the Basque Country, Paseo de Manuel Lardizabal, Donostia, 20018, Spain
[c] Basque Center for Applied Mathematics (BCAM), Mazarredo 14, Bilbao, 48009, Spain

ARTICLE INFO

Article history:
Received 18 March 2017
Received in revised form
29 August 2017
Accepted 8 November 2017

Keywords:
Circular variables
Weather forecasting
Meteorology
Regression trees
Machine learning

ABSTRACT

This paper describes a suite of tools and a model for improving the accuracy of airport weather forecasts produced by numerical weather prediction (NWP) products, by learning from the relationships between previously modelled and observed data. This is based on a new machine learning methodology that allows circular variables to be naturally incorporated into regression trees, producing more accurate results than linear and previous circular regression tree methodologies.

The software has been made publicly available as a Python package, which contains all the necessary tools to extract historical NWP and observed weather data and to generate forecasts for different weather variables for any airport in the world. Several examples are presented where the results of the proposed model significantly improve those produced by NWP and also by previous regression tree models.

Software availability

Name of software: AeroCirTree
Developer: Pablo Rozas Larraondo
Contact Address: National Computational Infrastructure, Building 143, Australian National University, Ward Road, ACT, 2601, Australia (pablo.larraondo@anu.edu.au)
Source: http://github.com/prl900/AeroCirTree
Programming Language: Python 3
Dependencies: Numpy, Pandas
Licence: GNU GPL v3

1. Introduction

Modern weather forecasting relies mostly on numerical models that simulate the evolution of the atmosphere, based on fluid dynamics and thermodynamics equations. These equations are solved for the discrete points of a regular grid covering the region of interest. Higher resolution models generate more detailed forecasts, but also require large computational resources and longer running times. Operational models trade off resolution quality for shorter processing times. The need for higher resolution forecasts has driven numerous methodologies to generate more detailed outputs, which is known as downscaling. Dynamic downscaling uses the output of a coarser model as the initial condition of a higher resolution local model, which better resolves sub-grid processes and topography (Carvalho et al., 2011). Another approach is statistical downscaling, where historical observed data are used to enhance the output of a numerical model. There are numerous methodologies for statistical downscaling based on different principles, such as analogues (Bannayan and Hoogenboom, 2008), interpolation (Plouffe et al., 2015) or machine learning models (Rozas-Larraondo et al., 2014; Salameh et al., 2009).

Aviation operations are highly affected by the weather and require the best quality meteorological information to maximise efficiency and safety. The International Civil Aviation Organization (ICAO) and the World Meteorological Organization (WMO) have established international standards to ensure high quality meteorological reports (WMO, 1995). To generate these reports, national weather services across the world employ highly qualified personnel who continuously observe and forecast conditions around the airport, such as visibility, direction and speed of the wind or proximity of storm cells. Aviation weather forecasters rely mainly on their knowledge of the airport and the quality of the NWP used.

* Corresponding author.
E-mail address: pablo.larraondo@anu.edu.au (P. Rozas Larraondo).

https://doi.org/10.1016/j.envsoft.2017.11.004
1364-8152/

There are a number of tools that facilitate the process of generating airport weather forecasts (Ghirardelli and Glahn, 2010; Jacobs and Maat, 2005), being an area of active research at the moment. Airports usually have long and regular series of high quality historical observation data that can be used to create statistical downscaling models to help forecasters in their work. The effect of non-resolved surrounding mountains, water bodies or local climate conditions can be incorporated by these models, by studying the local effects produced by weather patterns in the past.

Circular variables are present in any directional measurement or variable with an inherent periodicity. Weather data contain many parameters that are represented as circular variables, such as wind direction, geographical coordinates or timestamps. Most of the current regression machine learning algorithms focus on modelling the relationships between linear variables. Circular variables have a different nature to linear variables, so traditional methodologies are not able to represent their content thoroughly, leading to suboptimal results in most cases. The model presented in this article builds upon the concept of circular regression trees introduced by Lund (2002). Our model is computationally more efficient and generates contiguous splits for circular variables, which results in improved accuracy when compared to its precursor.

Circular regression trees can better represent circular variables, as they consider more possibilities for splitting the space than linear regression trees do. Circular regression trees can define subsets of data around the origin $0, 2\pi$ radians point. For example, when predicting an event that shows a high correlation with the winter months in the northern hemisphere, a circular tree would be able to isolate the months from December to March in one group. On the other hand, a linear tree would most likely consider splits starting or ending at the beginning of the year, failing to create a group containing these months.

This paper introduces AeroCirTree, a system based on the described circular regression tree model, which is able to generate improved airport weather forecasts for any airport in the world. This software presents a general solution where all the necessary tools required to extract historical weather data, train models and generate new forecasts are made available. This system is intended to help aviation weather forecasters to produce better quality reports and for machine learning researchers to build upon more sophisticated models.

The paper is structured as follows: Section 2 contains the methodology used to create the model. Section 3 contains an introduction to the observed and numerical weather datasets used to develop and test the system. Section 4 presents results where the proposed model is compared with other regression tree methodologies. This section also contains a discussion of the results, providing the reader with deeper insight into the novelty of the proposed model. Section 5 provides a high level description of the model implementation, including its key components and their functionality as well as examples on how to use the software. Section 6 concludes this paper, revisiting the research highlights and proposing some ideas on future developments to carry this work forward.

2. Methodology

Because of their simplicity, training speed and performance, regression trees are a popular and effective technique for modelling linear variables. Classification and Regression Trees (CART) (Breiman et al., 1984) is one of the most popular versions of regression trees.

Linear regression trees recursively partition the space, finding the best split at each non-terminal node. Each split divides the space in two sets using a cost function, which is usually based on a metric for minimising the combined variance of the resulting children nodes.

Fig. 1 contains an example of a regression tree based on two linear variables x_1 and x_2. On the right side, there is a graphical representation of how the space is divided by creating splits on these two variables.

Circular variables are numerical variables whose values are constrained into a cyclical space - for example, a variable measuring angles in radians, spans between 0 and 2π, where both values represent the same point in space. Although these variables can be included in a linear regression tree, they have to be treated as linear variables, which is an oversimplification and normally leads to suboptimal results (Lund, 2002).

A circular variable defines a circular space. A circular space is cyclic in the sense that it is not bounded; for instance, the notion of a minimum and maximum value does not apply. The distance between two values in the space becomes an ambiguous concept, as it can be measured in clockwise and anticlockwise directions, yielding different results. Also, this space cannot be split in two halves by selecting a value, as the '$<$' and '$>$' operators are not applicable.

In order to split a circular variable, at least two different values need to be defined. These two values describe two complementary sectors, each containing a portion of the data. Circular regression trees use this splitting approach for incorporating circular variables into regression trees.

There are many examples of circular variables. Any variable representing directional data or a periodic event is circular. More specifically, in the field of airport weather forecasting, wind direction, the time of the day or the date are examples of circular variables.

Lund (2002) proposes a methodology that allows circular variables to be incorporated into regression trees. Fig. 2 contains a similar representation to the previous example, but considering one circular variable α and a linear one x_1. On the right side, there is a chart representing how the space is partitioned using polar coordinates.

The methodology presented in this work builds upon the concept of circular regression trees, presenting an alternative that improves computational performance and the accuracy of its results. Fig. 3 shows how the space is partitioned using the proposed methodology.

Visually comparing Figs. 2 and 3, it is evident that regions are split differently. The novelty of this methodology, when compared to the original version proposed by Lund, is that it always generates contiguous splits. In doing so, we avoid an excessive fragmentation of the space, and the splits provide a better generalisation for its child nodes. The original methodology uses the '$\in$' and '$\notin$' operators to generate all the splits for circular variables. This usually generates partitions in which the subsets defined by the $\in$ clause are surrounded by the complementary $\notin$ subset. Our methodology uses these operators to create just the first split of a circular variable and, after that, uses the '$<$' and '$>$' operators to create the subsequent splits. This change also results in a reduction of the search space for possible splits. The proposed algorithm for generating circular trees has, as a consequence, $\mathcal{O}(n)$ cost instead of $\mathcal{O}(n^2)$, when compared to Lund's original proposal. The only exception is when computing the first split of a circular variable, which has a computational cost of $\mathcal{O}(n^2)$, as it has to consider all the different splits around the circle.

3. Software and datasets

AeroCirTree is a collection of Python scripts which provides the tools to train and test the three previously described regression tree