CS 229

Weekly Climate Indices: Generation and Prediction

Javier Echevarría Cuesta
Department of Mathematics
Stanford University
javierec@stanford.edu

J.K. Hunt
Department of Computer Science
Stanford University
jkhunt@stanford.edu

James Schull
Department of Computer Science
Stanford University
jschull@stanford.edu

1 Introduction

Climates and local climatic phenomena are enormously complex systems; understanding them precisely requires the granular measurement of an expansive set of variables. It is thus very often useful to be able to approximate different climatic states with indices: a deliberately limited set of variables that attempt to balance simplicity and representative power. Examples of such indices include daily or monthly surface temperature measurements from a finite set of global coordinates. Much work has been done with daily and monthly indices, but very little study has been carried out with regard to the production and usefulness of *weekly* indices. This presents an enticing avenue of investigation: while daily indices are more representative of weather than climate, and monthly indices capture information on a relatively long time-scale, weekly indices prove highly useful in approximating short-term climate patterns, and thus are valuable in a number of public and private sector contexts.

In this paper, we present a two-part investigation into weekly climate indices. In the first part, we explore the *generation* of weekly indices, employing a K-means clustering based approach to the generation of weekly surface temperature (SKT) indices, allowing us to visualize regions causally related to one-week-forward temperatures in California and Peru. In the second part, we focus our attention on the *prediction* of weekly indices, exploring the predictability of weekly averages of the North Atlantic Oscillation, a prominent climate pattern that strongly affects weather over the northeastern North America, Greenland, and Europe, and influences energy demand, crop yields, and fishery productivity, among many other climate-dependent human activities. [1] In this latter section of the project, we take two time series–a 70 year series of weekly 250 hPA pressure readings across a latitude-longitude grid of points, and a univariate time series of NAO readings–and train a variety of autoregressive and supervised models, including ARIMA and an LSTM, to predict the next week's NAO reading.

Our experiments yielded a number of interesting results. Our k-means clustering approach to generating SKT indices proved effective, reducing the dimensionality of the original dataset and identifying distinct regions of predictive power. We were able to improve upon a baseline forecast of weekly-averaged NAO indices, though we found it difficult to dramatically improve performance; we hypothesize that this level of unpredictability is a property of a time series composed of weekly averages of an oscillation that occurs on a non-weekly basis.

2 Related work

Our work on generation and prediction of climate indices builds upon a growing literature. In 2003, Steinbach et. al offered a "clustering-based methodology for the discovery of climate indices that overcome [the limitations of PCA and SVD] and is based on clusters that represent regions with relatively homogeneous behavior." [2] Evans and Singh extended from this methodology in their project from the Fall 2017 edition of this class, developing a novel pre-processing method and quantifying the representative power of their indices by predicting surface temperature one, six and nine months into the future in the region of Peru. [3] Given our interest in their work, we have been in communication with Evans and Singh, who have given us strategic guidance and provided access to some pre-processed data for our experiments. Notably, the work of Evans and Singh focused on generating monthly climate indices, while we focus on weekly climate indices, which exhibit some significantly different properties.

Modelling of the North Atlantic Oscillation has been explored, too: Scaife et al (2014) built an ensemble forecasting system used to predict the NAO index, and similar numerical models have been developed since. [4] More technically related to the models developed in this paper is the work done by Yuan et al (2019), who used a convolutional LSTM with ensemble empirical mode decomposition to make daily forecasts of the NAO index. [5] Our work extends loosely from Yuan et al's approach, applying a similar recurrent neural network to NAO predictions, but with the addition of potentially salient pressure information and a focus on weekly, rather than daily, predictions.

3 Dataset and Features

3.1 Surface Temperature

For the index generation part of our project, we use a reanalysis dataset produced by the National Center for Atmospheric Research, consisting of weekly surface temperature (SKT) readings from 1948 to today (3286 readings in total). Each time step consists of a 192×96 matrix representing SKT readings across the earth. Following Evans and Singh, we normalize the temperature data first by subtracting the mean and dividing by the standard deviation; we also account for the consistent increase of average temperature through time (climate change) by fitting a linear regression to the data and subtracting the slope.

Whereas most papers first cluster climate data points using some notion of distance and then use these clusters to make predictions, we decided to follow the approach of [2], which is to cluster the points based on their predictive power, rather than cluster first and then predict. Given the $192 \times 96 \times 3286$ matrix of readings, we hence pre-process the data by flattening it into a two dimensional matrix of size 18432×3286 such that each row contains the weekly surface temperature reading for a single grid point since 1948. We use this to train a linear regression for every grid point, where the number of features is one and the objective is to predict the temperature a week after the given time step, for a chosen location. This extremely simple regression allows us to record the R squared value that results from each prediction; each point is associated to its own R squared value, which is supposed to approximate the predictive value of that particular point with respect to the chosen location.

3.2 Pressure and North Atlantic Oscillation

For the NAO prediction part of our project, we use two datasets. Firstly, we use a weekly time series of 250 hPA pressure readings over a 192×96 matrix representing earth, spanning 1948 to the present. For our target variable (and use of previous observations as features), we use the National Oceanic and Atmospheric Administration's NAO index dataset, comprised of daily readings from 1948 to the present. In order to convert this into a weekly index, we average each week's indices. We pre-process the pressure data via a number of steps. First, we employ min-max scaling to normalize the dataset and restrict each value to the [0,1] interval, limiting the impact of outliers and assisting with backpropagation. Min-max scaling comprises transforming each feature according to the following equation:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}.$$

Since the pressure dataset comprises 3730 time steps, each with $192 \times 96 = 18,432$ features, we perform PCA upon the dataset to reduce the number of features and prevent overfitting. Next, we set a lag variable and reshape our training matric such that for lag ℓ, we train and predict time step t's target using data from time steps $[t - \ell :]$ For training our LSTM, we reshape each example into a vector of dimension $(1, \ell, n\_features)$, where $n\_features$ is dependent on the number of principal components chosen during the preprocessing step.

4 Methods

4.1 Weekly SKT index generation with K-means clustering

Given the preprocessed SKT time series, we perform K-means clustering to generate SKT indices. In order to incorporate a notion of geographical proximity, we use the following distance metric:

$$d(x, y) = |R_1 - R_2| + \alpha((x_{\text{lat}} - y_{\text{lat}})^2 + (x_{\text{long}} - y_{\text{long}})^2)^{1/2},$$

where α is a hyperparameter. Setting $\alpha = +\infty$ will cause K-means clustering to simply cluster by region, while setting $\alpha = 0$ will separate the points into potentially very disperse groups of locations of similar predictive value. Once we have found k centroids, we define our k SKT indices to be the mean of the grid points associated to each centroid.

4.2 Weekly NAO prediction

In the second stage of our project, we train a variety of learning algorithms to predict NAO a week in advance. Though in our experiments we tried a wide range of models, in this section we will focus on the two types of models that we considered best suited to the task, and consequently invested the most time in optimizing: ARIMA, and an LSTM.

4.2.1 ARIMA

ARIMA (AutoRegressive Integrated Moving Average) is a linear forecasting algorithm; specifically, it is a univariate linear regression model that uses its own lags as predictors. ARIMA is characterized by 3 parameters: p, d, and q. p refers to the order of the autoregressive (AR) term, the number of lag weeks to be used as predictors; d refers to the order of differencing to make the series stationary; and q refers to the order of the moving average (MA) term, the number of lagged forecast errors to be used as predictors. To illustrate p we can consider a pure AR model that represents some value y at time step t as a linear combination of p lag weeks, including intercept and the error term. Similarly, to represent q, we can consider a pure MA model that represents some value y at time step t as a linear combination of q error terms (as mentioned in the above pure AR model). Ultimately, ARIMA is formulated as a linear combination of the above two models:

$$y_t = k + a_1 y_{t-1} + a_2 y_{t-2} + ... + a_p y_{t-p} + \epsilon_t + m_1 \epsilon_{t-1} + m_2 \epsilon_{t-2} + ... + m_q \epsilon_{t-q} \tag{1}$$

4.2.2 LSTM

An LSTM is a specific recurrent neural network architecture designed to capture long term dependencies in sequence data. Unlike a feedforward neural network, an LSTM consists of a single layer of hidden units, the outputs of which are fed back into itself in a loop. Unlike a traditional recurrent neural network, an LSTM's hidden layer 'remembers' and passes information through the loop via an internal component called a cell state, and the information that is remembered by the cell state is determined by interaction with three 'gates': the forget gate layer; the input gate layer; and an output gate layer. Briefly put, each of these layers outputs a value determined by both x_t, the tth input in the sequence, and h_{t-1}, a value output by the LSTM after the previous input. The forget, input, and output gates compute their value and determine modifications to the cell state via the following equations:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o).$$

5 Experiments and Results

5.1 SKT index generation: K-means clustering

5.1.1 Experiments and hyperparameter tuning

Since the intention of the earlier section of our project was to extract meaningful indices from global SKT data, most of our experiments revolved around visualization of the indices produced by K-means. We generated indices for both Northern California and Peru, experimenting with a range of different K's, and found that even at very low values of K, we could identify precise regions of predictive value.

5.1.2 Results and error analysis

Figure 12 depicts cluster assignments for grid points that were clustered using the aforementioned distance metric, with their predictiveness (as measured by R^2) calculated with respect to Peru.

We see that even at K=4, centroids distinctly highlight regions of causal association to Peru's temperature. Figure 2 depicts the centroids determined by clustering grid points based on their predictiveness with respect to Northern California's one-week-forward SKT index. As in the case of Peru, we find that the centroids determined by K-means identify a clearly defined region of predictive value; in fact, this region corresponds to a section of ocean slightly to the Northwest of Northern California, in accordance with our expectations regarding the movement of temperature patterns towards California across the Pacific.
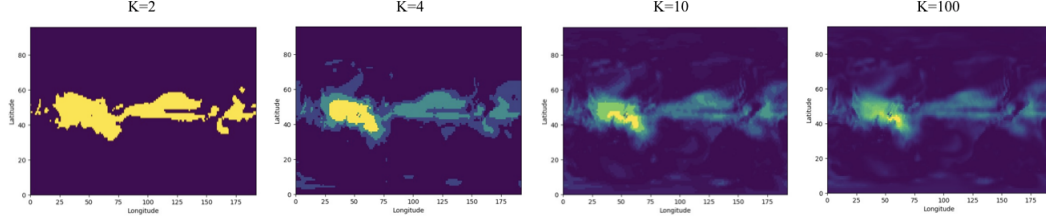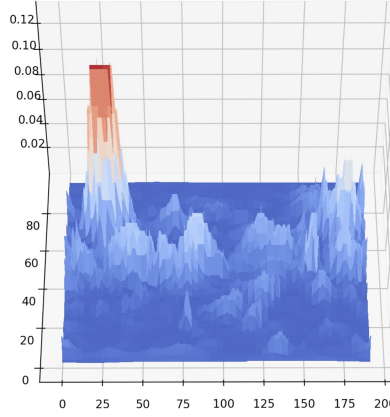
Figure 1: Centroid assignments (Peru)

Figure 2: Centroid assignments (Northern California)

5.2 NAO index prediction

5.2.1 Experiments and hyperparameter tuning

During the course of our supervised experiments, we chose to use root mean squared error (RMSE) as our metric, simple and common metric for regression tasks. A very common baseline method for supervised learning on time series datasets is the persistence algorithm (or naive forecast), which simply predicts the value at time step $t + 1$ to be the value at time step t. We find that the persistence algorithm yields a root mean squared error (RMSE) on the NAO index dataset of 106.6 (4 s.f.), which we consider our baseline.

In order to find the most appropriate model for NAO prediction, we tested a considerable number of hyperparameters. The parameters that were applied to our data preprocessing were the number of PCA components extracted from the pressure data, as well as the number of lag weeks that were fed to the LSTM or ARIMA. Specific to ARIMA were the aforementioned parameters p, d, q; specific to the LSTM were `num_epochs` and `num_hidden_units`. Since our training set was relatively small (approximately 3300 examples, depending on the lag parameter), an LSTM could be trained in a reasonable amount of time; consequently, we conducted a thorough grid search of parameters, including both the preprocessing parameters and the LSTM-specific parameters.

5.2.2 Results and error analysis

Our grid search over the preprocessing and RNN parameters revealed that the optimal model for this task was an LSTM with 50 hidden units and a tanh activation function, trained for 30 epochs with 10 weeks of lag and a pressure dataset reduced to 10 dimensions via PCA. The train and validation loss are shown in the below figure. This model yielded an RMSE of 93.57 (4 s.f.), representing a 12.2% improvement upon our baseline. Our hyperparameter search for ARIMA revealed that the optimal parameters were a 4 week autoregressive lag (p) and a 7 week moving average (q), with differencing (d) of 0, yielding an RMSE of 90.12 (4 s.f.), a 15.5% improvement upon our baseline.
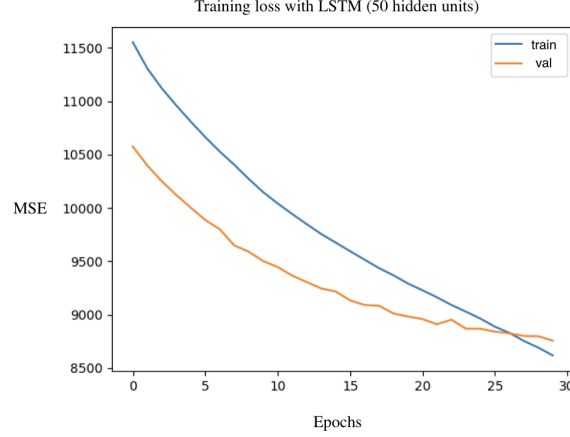
4

Figure 3: LSTM loss

Model	RMSE
Baseline (persistence)	106.6
LSTM	93.57
ARIMA	90.12

While our models outperformed the baseline by a reasonable amount, we were intrigued by how difficult it was to improve their performance. We believe that this is a strong reflection of the difficulty of predicting the weekly average of a climate phenomenon that oscillates on a non-weekly basis. The weekly-averaged NAO time series appears far noisier than the daily series, without the sinusoidal structure visible in the latter; and consequently, train set performance may well not correlate strongly with test set performance.
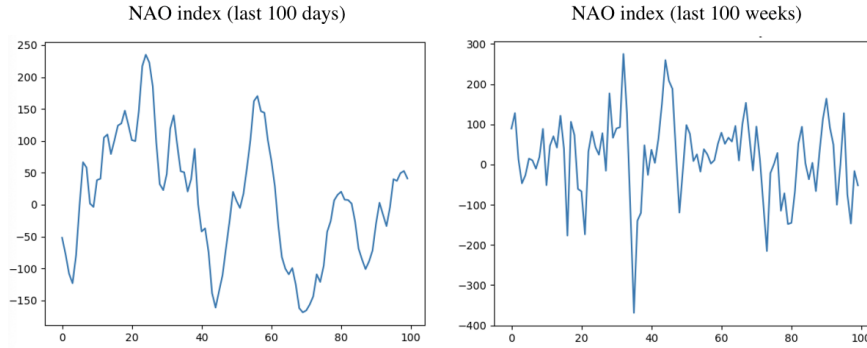

Figure 4: Daily versus Weekly time series

6   Conclusion and Future Work

In this paper, we explored methods for both generating and predicting weekly climate indices. We chose to focus on weekly indices because they have utility in a range of domains, and yet are under-studied. In the first part of our project (generation), we found that K-means clustering with an uncommon pre-processing step identified distinct regions of predictive power. In the second part of our project (prediction), we trained a variety of models to forecast a weekly index tracking the North Atlantic Oscillation, and found that the best-performing model was ARIMA, an autoregressive model. Interestingly, we faced difficulties in dramatically improving performance beyond our baseline; we suggest that this illustrates the difficulties associated with modelling climate oscillations whose periodicity is incongruous with the periodicity of the index being modelled. Given more time, we would experiment with different approaches to week-forward predictions, using daily as well as weekly input data, and integrating other climate features alongside pressure.

Link to code

`https://github.com/jameskschull/climate-indices`

References

[1] https://www.climate.gov/news-features/understanding-climate/climate-variability-north-atlantic-oscillation

[2] Maximilian Evans, Jasdeep Singh (Dec. 2017), *Unsupervised Machine Learning for Long Range Climate Prediction*, Stanford University.

[3] Michael S Steinbach et al. (Dec. 2003), *Discovery of climate indices using clustering*, University of Minnesota.

[4] Scaife, A.A. et al. (2014), *Skilful Long Range Prediction of European and North American Winters*, Geophys. Res. Lett., 41, 2514–2519.

[5] Shijin Yuan et al. (May 2019), *Prediction of North Atlantic Oscillation Index with Convolutional LSTM Based on Ensemble Empirical Mode Decomposition*, School of Software Engineering, Tongji University, Shanghai.

[6] Fereday, D.R. et al. (2012), *Seasonal forecasts of northern hemisphere winter 2009/10*. Environ. Res. Lett., 7, 34031–34037.

[7] Lin, H. et al (2011), *Impact of the North Atlantic Oscillation on the forecast skill of the Madden-Julian Oscillation*, Geophys. Res. Lett., 37, 96–104.