University of Colorado
Department of Civil, Environmental and Architectural Engineering
Advanced Data Analysis Techniques
CVEN 6833

Homework Set 3 Date :11/16/2021
Due by:12/14/2021

Topics: Parametric/Nonparametric Time Series, Hidden Markov Model, Wavelet Spectral Analysis, Extreme Value Time Series and Copulas

Please present your work neatly. Organization of R-commands, functions will fetch 15% of points. Data, commands etc. at
http://civil.colorado.edu/~balajir/CVEN6833/HWs/HW-3/
The data is available at this google drive
https://drive.google.com/drive/folders/1Y2qelQNR-2Re0mJQzymYdoB7Y53pmoN4

K-NN Conditional Ensemble Forecast/Simulation

1. Use K-NN bootstrapping approach to simulate/forecast Spring streamflow on the Colorado River at Lees Ferry based on suite of predictors, also called, 'feature vectors', at two lead times – Mar 1 and Apr 1. The feature vector includes – EPS ensemble mean forecast, winter climate indices (AMO and PDO) and SWE. You can experiment with all or subset of these variables.
The steps for each lead time are:
(i) For each year, *t,* using the feature vector in that year, $\boldsymbol{x}_t$ *obtain* K-nearest neighbors from the historical feature vectors (of course, excluding the year, *t*)
(ii) Select one of the K neighbors using a weight metric. The selected neighbor corresponds to a historical year and with it the associated spring streamflow.
(iii) Repeat steps (ii) say 100 times, to obtain *ensemble simulation* of streamflow for each year t. Compute the mean or median of the ensemble to get a single value.
(iv) Repeat steps (I) – (iii) for all the years and similarly for the two lead times.
(v) Plot the historic flow vs ensemble mean forecast along with the 1:1 line for visual comparison. Compute skill scores for each lead time - correlation between the historic flow and the mean of the ensemble forecast, also compute RPSS. Comment on what you find.

Conditional Forecast/Simulation - Copulas

2. (i) Repeat problem 1 using Copula Regression – 'gcmr' package in R. There will not be an ensemble forecast, but a mean value.
(ii) Fit a Copula to the spring streamflow and April 1 SWE.

Support Vector Machines

3. Fit a Support Vector Machine (SVM) model for Spring flows using predictors for Mar 1 and Apr 1. Obtain estimates of flow from the models and compare their performance with historic flows.

Modeling Nonstationary Time Series - HMM & Forecast and simulation

4. Another way to model/simulate a time series is using HMM. For the spring streamflow in problem 1 for April 1st lead time fit a HMM and make the forecast. The steps are as follows:

- Fit a best HMM model for the spring streamflow
- Fit a best GLM (mostly logistic regression) to the state sequence as a function of predictors for April 1 and the state from the previous year – i.e.,

$$S\_t = f(S\_t\text{-}1, \boldsymbol{x}_t)$$

- Using this best GLM, for each year, *t,* based on the predictor vector obtain the probabilities of the *states* (i.e. the distribution of the HMM)
- Using these state probabilities, simulate flow from the corresponding *state* PDFs – to obtain an ensemble

- Plot the historic flow vs ensemble mean forecast along with the 1:1 line for visual comparison. Compute skill scores – RPSS and correlation between the historic flow and the mean of the ensemble forecast.
- Repeat this for March 1st lead time

5. Compare and comment on the results and methods (pros/cons/utility) employed in problems 1 ~ 4.

Modeling Space-Time Nonstationary Extreme Value Analysis (EVA)

6. Perform a space-time nonstationary EVA on the winter 3-day maximum precipitation.Below are the steps:

At each location fit a nonstationary GEV model to the 3-day winter maximum precipitation as function of the 3 covariates – the leading ~3 winter SST PCs. Make only the location parameterof the GEV non-stationary. This will result in 4 coefficients (intercept plus the three covariates)

- Fit a spatial model to each of the coefficients and to the scale and shape parameter
- For couple of wet and dry years (you can select the years based on the average spatial 3-day precipitation) – obtain the 2-year, 50-year and 100-year return levels on the spatial grid
    i. Using the spatial models obtain the GEV parameters at each grid point
    ii. Estimate the 2-year, 50-year and 100-year return levels at each grid point and map them. Compare with 2-year return levels with the observed values in the selected years
- For couple of representative locations plot the time series of 3-day precipitation maximumalong with the time varying return levels. Compare them with the stationary return levels

Bonus – Optional Questions

7. You can implement a Bayesian version of the above

Singular Spectrum Analysis – Diagnostics & Forecasting

8. For the spring season flow at Lees Ferry on the Colorado River perform SSA and makepredictions from it. The steps are as follows:

Diagnostics

b. Select a window size of about 10-20 years (feel free to experiment with the window size); create the Toeplitz matrix and perform SSA.
c. Plot the Eigen spectrum and identify the dominant modes.
d. Reconstruct the dominant modes (i.e. Reconstructed Components - RCs) and plot them. Infer from them the dominant periodicities.
e. Sum the leading modes and *plot them along with the original time series*. This will show the 'filtering' capability of SSA. Feel free to play with the number of RCs.
f. Plot the dominant modes and show their corresponding wavelet spectra.

Prediction

g. Use the 'feature vector' from problem 1 to simulate/project the leading modes and use them to make projections/simulations at April 1st lead time. Specifically Predict the last 5 year period 2014 - 2017
h. The steps are - apply the SSA to data for the pre-2014; make a prediction for 201d and repeat for each year through 2017.
i. *Plot the observed and predicted values; compute the median correlation.*