**Final task: Predicting the GDD of a given day using an improved K-NN model**

Nearest neighbor methods have been intensively used in the field of statistics and in pattern recognition

procedures. Despite they are simple, nearest neighbor algorithms are considered robust. Although many

other sophisticated techniques have been developed, nearest neighbor methods remain very popular. A

K-NN algorithm involves selecting K data vectors similar to the vector of interest. In the context of weather data simulation, the nearest neighbor approach involves simultaneous sampling, with replacement, weather variables, like precipitation and temperature, from the observed data. To generate weather variables for a new day, t+1, days with similar characteristics to those simulated for the previous day t are first selected from the historical record. One of these nearest neighbors is then selected according to a defined probability distribution or kernel and the observed values for the day subsequent to that nearest neighbor are adopted as the simulated values for day t+1. Therefore, once we select a nearest neighbor, we can take this neighbor GDD as GDD for day t+1.

The spatial dependencies should be preserved because the same day’s weather is adopted as the weather for all stations. Apart from the spatial dependencies, temporal dependence is preserved as the simulated values for day t +1 are conditioned on the values for the previous day t.

Consider that the daily historic weather vector consists of *p* variables. Suppose the number of stations considered in the model is *q* and data are available for *N* years. Letdenote the vector of weather variables for day *t* and station *j*, where *t*=1 , . . , *T,* and *j*=1, . . . , *q*; *T* being the total number of days in the observed time series. The feature vector for day t can be expressed in expanded form as  whererepresents the value of the weather variable i for station j and day t.

**It is important** that because we want to predict the GDD of a given day, so the p variables should contain variables that have strong relationship with GDD (i.e. feature selection should be implemented first). For example, the Max Temp and Min Temp should be considered.

The algorithm cycles through various steps to obtain the weather for day *t*+1 and the steps of the algorithm are as follows:

**1.** Compute regional means of the *p* variables across the *q* stations for each day of the historical record

 (1)

Where

 i=1,…, p, and t = 1,…,T (2)

For example, for the vector , if t =1, 1 represents precipitation, then the represents means of the precipitation across q stations for day 1.

**2.** Determine the size, *L*, of the data block that includes all potential neighbors to the current feature vector. A temporal window of width *w* is chosen and all days within the window are considered as potential candidates to the current feature vector. Yates et al. (2003) used a temporal window of 14 days, which implied that if the current day is January 20 then the window of days consists of all days between January 13 and January 27 for all *N* years but excluding January 20 for the given year. Although the value of w is 14, actual number of all days between January 13 and January 27 is 15. Therefore every year in N years has (w+1) potential neighbors. And the current day January 20 should not be consider as a neighbor of itself. Thus, the day block of potentials consists of L =(w+1) N1 days.

**3.** Compute mean vectors across q stations for each day in the data block consisting of potential neighbors using the expressions given in step 1.

**4**. Compute the covariance matrix, Ct , for day t using the data block of size L× p .

**5.** The weather on the first day t (e.g., 1 January) comprising all p variables at q stations is randomly chosen from the set of all January 1 values in the historic record of N years. The algorithm cycles through the following steps to select one of the nearest neighbors to represent the weather for day t+1 of the simulation period.

**6.** Compute Mahalanobis distances (Davis 1986) between the mean vector of the current day’s weather and the mean vector for day i, where i = 1,……,L. day i means a potential neighbor of the current day t . The distance metric can be given through

 (3)

Where *T* represents the transpose operation; and =inverse of the covariance matrix. Yates et al. (2003) used the Mahalanobis distance metric to determine the closeness of any given neighbor to the current vector as it does not require explicit weighting and standardization of the variables.

**7.** Determine the number of *K* nearest neighbors from all potential neighbors. Lall and Sharma (1996) suggested the use of the generalized cross validation score (GCV)for choosing *K*. Rajagopalan and Lall (1999) and Yates et al. (2003) recommended the use of a heuristic method for choosing *K* according to which *K* =.

8. Sort the Mahalanobis distances in ascending order and retain the first *K* nearest neighbors. A discrete probability distribution that gives higher weights to the closer neighbors is used for resampling from the *K* nearest neighbors. Weights are assigned to each neighbor of j neighbors according to the metric given by

 (4)

The cumulative probabilities, , are given by

 (5)

The neighbor with the smallest distance is assigned the highest weight, while the neighbor with the largest distance (i.e. the neighbor) gets the least weight.

9 Determine the nearest neighbor of the current day by using the cumulative probability metric given by Eq. (5):

Firstly, generate a random number, .

If <r<, then the day *j* for which *r* is closest tois selected.

If , then the day corresponding to  is selected.

If , then the day corresponding to is selected.

The selected neighbor is adopted to represent the day t+1. Finally, once we select the neighbor, we can take its GDD as GDD of our predicting day t+1.

So far, we can already predict the GDD of future days. Based on this situation, we can do some other

predictions. For example, showing the expected date for spring flowers to emerge.

**Reference**

1. Mohammed Sharif and Donald H. Burn “Improved K-Nearest Neighbor Weather Generating Model”