A dual hidden Markov model for the prediction of intermittent streamflow in desert regions

Catherine Wilcox, Taha B.M.J. Ouarda, Wei Lee Woon
Masdar Institute, Earth Observation and Hydro-Climate Lab
PO Box 54224, Abu Dhabi, UAE
Email addresses: cwilcox@masdar.ac.ae, touarda@masdar.ac.ae, wwoon@masdar.ac.ae

1. Introduction

- Arid and semi-arid environments frequently feature streams with intermittent streamflow
- Within "flow" state additional patterns may exist, driven by hidden climactic variables
- Markov and hidden Markov models provide statistically-based tools for evaluating and predicting the states which drive hydrological phenomena [1]
- Objective: to evaluate the effectiveness of multiple Hidden Markov model (HMM) methods in the characterization and prediction of streamflow in desert regions
- Data from 35 intermittent USGS streamflow sites in southwest USA used to test methods [2]
    - Sites selected primarily based on length (10+ years) , % intermittency

2. Markov and hidden Markov chains

- First-order Markov chains operate on the principle that the probability of the current value in a series depends only on the value directly before it. [3]

$$P(X_n = x_n | X_{n-1} = x_{n-1} \dots X_0 = x_0) = P(X_n = x_n | X_{n-1} = x_{n-1})$$

- HMMs assume a Markovian series of hidden state transitions. Each state has a given distribution for the probabilities of emitting actual data values (Fig. 1).
- HMMs depend largely on Bayesian statistics, interpreted as follows:

$$P(state|emission) = \frac{P(emission|state)P(state)}{\sum_{states} P(emission)}$$

3. Methodology

Non-hidden "outer" Markov, hidden "inner" Markov (NHoHi)

- All zero-flow points in training data initially assigned to outer state 1, nonzero points to outer state 2 (Fig. 2)
- Two-state HMM trained only on nonzero data (Fig. 3)
- Algorithms used:
    - Baum Welch [5]
    - Viterbi [6]
- Trained HMM used to predict test data (Fig. 4)

Hidden "outer" Markov, hidden "inner" Markov (HoHi)

- Initial HMM run on training data without prior knowledge of states
- Data is sorted into two sets based on state
- Additional two-state HMM trained on each data set
- Other methods same as NHoHi

- Initial "outer" HMM was tested alone as a control for comparison with other methods

4. Continuous vs. discrete analysis

Primary methods featured a discretized HMM by binning data. The optimal number of bins for each data set varied strongly according to Bayesian information criterion (BIC) [7].

To remedy, a continuous Gaussian mixture model (GMM) was introduced [8]. Instead of the bin frequency, the mean and standard deviation of each state were optimized.

GHMM predicted the most likely distribution for the data, but additional methods are required to assign values to the prediction.

5. Significant findings

- NHoHi model convergence depended on percent of intermittency. Generally the model was not able to make predictions on data series with >50% intermittency (Table 1).
- When NHoHi converged, its prediction accuracy averaged at 94.6%%.
- The HoHi model was found to always converge to an absorbing Markov chain due to the over-complexity of the model. Adaptations removed the problem, but skewed results.
- Prediction accuracy increased with data length, decreased with variance (Fig. 5,6)
- NHoHi produced more accurate and consistent results than standard HMM (Table 1)

6. Results and figures

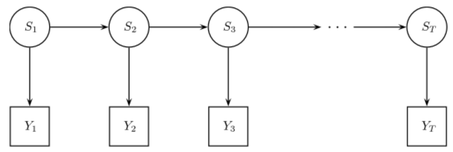

Fig. 1: Hidden Markov Model with hidden states (circles) and emissions (squares) [4]

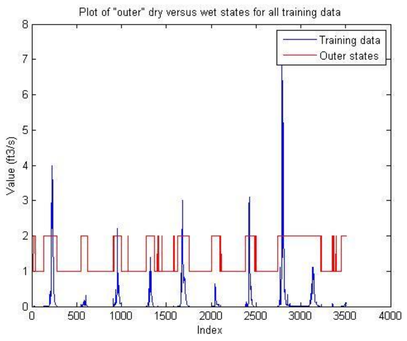

Fig. 2: Plot of training data and outer, pre-assigned state data (NHoHi case)

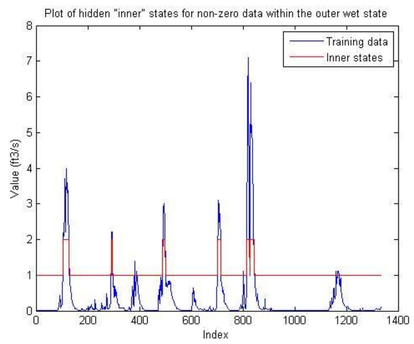

Fig. 3: Plot of HMM inner states (NHoHi case)

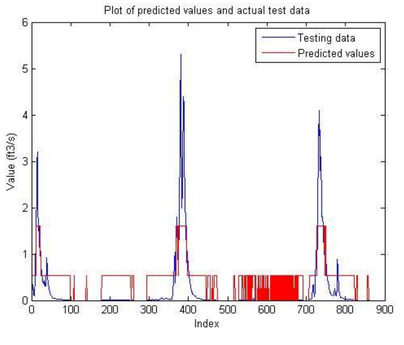

Fig. 4: Sample prediction (NHoHi case)

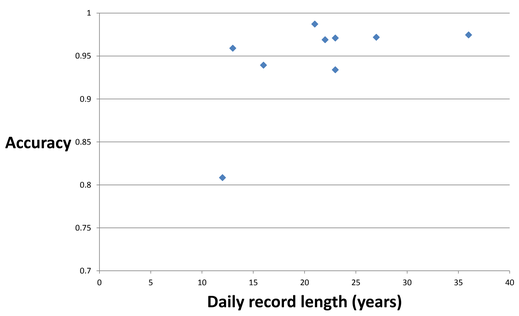

Fig. 5: Prediction accuracy vs. length of data record

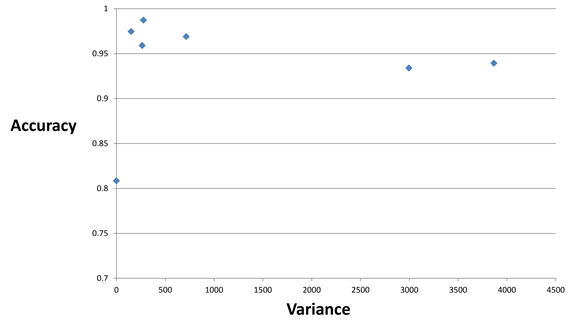

Fig. 6: Prediction accuracy vs. variance

7. Data/Results details

Table 1: Summary of results for NHoHi

Number of sites tested	35
Minimum daily record length (years)	12
Maximum daily record length (years)	36
Average daily record length (years)	20.8
Average intermittency	74.20%
Average intermittency w/ model convergence	34.60%
Average intermittency w/ model non-convergence	87.90%
Average accuracy of prediction	94.60%
Minimum improvement over standard HMM	5.61%
Average improvement over standard HMM	19.94%

References

[1] Thyer, K., & Kuczera, G. (2003). A hidden markov model for modelling long-term persistence in multi-site rainfall time series 1. model calibration using a bayesian approach. Journal of Hydrology, 275(1), 12-26.
[2] United States Geological Survey (USGS). (2012). USGS Surface-Water Daily Data for the Nation. Retrieved from http://waterdata.usgs.gov/nwis/dv/?referred_module=sw
[3] Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257-286.
[4] Visser, I. (2011). Seven things to remember about hidden markov models: A tutorial on markovian models for time series. Journal of Mathematical Psychology, 55(6), 403-415.
[5] Welch, L. R. (2003). Hidden markov models and the baum-welch algorithm. IEEE Information Theory Society Newsletter, 53(4), 1-14.
[6] Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory, 13(2), 260-269.
[7] Priestley, M.B. (1981) Spectral Analysis and Time Series, Academic Press
[8] Bilmes, J. A. (1998). A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. International Computer Science Institute, 4(510), 126.