A spatiotemporal quantification of the relative importance of indicator inputs for drought estimation

Soni Yatheendradas (UMD/ESSIC & NASA/GSFC)
Christa Peters-Lidard (NASA/GSFC; PI)
Sujay Kumar (NASA/GSFC)
David Mocko (SAIC & NASA/GSFC)

18th February, 2021

The 2nd NOAA Workshop on Leveraging AI in Environmental Sciences
Some project objectives

- **DATA**: Provide comprehensive database of drought indicators and the United States Drought Monitor (USDM)

- **ANALYSIS**: Use the database to quantify the relative importance of each indicator for the USDM drought by location and time of year

- **UTILITY**: Per user requirements and purpose, provide an importance-ordered list of indicators, either the full set or a reduced one that best and adequately represents the USDM
USDM weekly maps

U.S. Drought Monitor
Continental U.S. (CONUS)

October 15, 2019
(Released Thursday, Oct. 17, 2019)
Valid 8 a.m. EDT

D0: Abnormally Dry 70th < Percentile < 80th
D1: Moderate Drought 80th < Percentile < 90th
D2: Severe Drought 90th < Percentile < 95th
D3: Extreme Drought 95th < Percentile < 98th
D4: Exceptional Drought 98th < Percentile

Intensity:
- None
- D0 Abnormally Dry
- D1 Moderate Drought
- D2 Severe Drought
- D3 Extreme Drought
- D4 Exceptional Drought

The Drought Monitor focuses on broad-scale conditions. Local conditions may vary. See accompanying text summary for forecast statements.

Author:
Richard Helm
NCE/NOAA

droughtmonitor.unl.edu
Pilot study: Using inputs indicators of CPC objective blends

- CPC objective drought blends use 5-6 indicators, for short-term or for two different long-term regions
- Pilot study question: “What are the relative importances of the CPC objective blend inputs in reproducing the USDM?”
USDM considered for the period 2006-2019 (input data-based)

We try mutual information (MI), and machine learning techniques: Neural Network (NN) and Random Forest (RF)

Relative importances of inputs calculated in each technique are similar to the CPC blend "weights"

MI and RF techniques have their own inherent technique for calculating these importances
Mutual Information-based analysis

- **Mutual information (MI)**
  - calculates **common information entropy** between USDM and input
  - calculated **independently** for each input indicator
  - **uses only the data** – unlike machine learning that also involves model/s

- **Fractional Information (FI)** shown is the MI normalized by the entropy of USDM

- For comparison against ML, next slides show the **Normalized FI (NFI)** that is the FI normalized by the sum of FIs of the considered set of inputs
Example spatial maps: fractional info & most important input

Legend:

PMDI: Palmer (Modified) Drought Index
Normalized Fractional Information vs. CPC Blend weights

CPC Blend is a linear model plus a mapping to categories, while the mutual information-derived NFI considers all nonlinearity.
Example NN with one 16-neuron intermediate layer:
Training confusion Matrices for example inputs and spatial domains

- High misclassification levels, especially at extreme drought categories, likely due to class imbalance and remediable by category-weighted loss functions
- Can a technique unaffected by class imbalance (e.g., random forests) provide acceptable confusion matrices?
Random Forests & its implementation

• About Random Forests:
  • Random forest technique considers an ensemble of decision trees
  • Each decision tree considers a random bootstrapped set of training examples
  • For classification, ensembling done through majority vote

• Our implementation:
  • Initial Stratified Shuffle Split of all data into cross-validation and out-of-sample sets
  • A stratified K-Fold cross-validation training with 3 folds
  • Further out-of-sample evaluation
  • Each tree considers ALL input features
Cross-validation Confusion Matrices for Random Forests

Perfect simulation of exact drought category: see “one-to-one” line!
Out-of-sample Normalized Confusion Matrices for Random Forests

For majority of true labels, correct prediction occurs and seen along the "one-to-one" line

But adding this out-of-sample into previous slide’s training set to create new bigger training set again gives a perfect prediction!
Random Forest-based importances vs. CPC Blend weights
Summary & ongoing work

- Mutual information-based importance measures are importance reference points for machine learning technique-based measures
  - We have developed a library to obtain combined mutual information of any set of indicators with USDM for ongoing investigation of the incremental utility of inputs
- Class imbalance hampers training in neural network, and potentially remedied by category-weighted loss functions
  - We are testing out these coded category-weighted functions and their predictions
- Excellent prediction of random forests before pruning generalization is highly encouraging for obtaining similar prediction levels using category-weighted loss functions in hyperparameter-tuned neural network (and its input relevance calculation)
- Input indicators: We are acquiring and processing additional inputs (100+ that go into the USDM, modeled products, remotely sensed products etc.)