SAND2022-16549C

Sandia National Laboratories

Feature Pathway Detection using Random Forest Regressor Feature Importances

Matt Peterson, Meredith Brown, Jake Nicholas, Vitus Leung

OVERVIEW

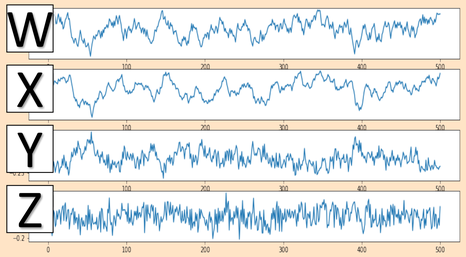
Goal
- Generate a weighted directed graph of source-impact pathways by using Feature Importance metrics from a Random Forest Regressor (RFR)

Approach
- Train a multi-variate RFR on a predetermined set of variables, time-lags and spatial dimensions
- Extract the feature importance weights between variable-variable pairs
- Convert weights into a weighted directed graph ("RFR feature pathways network")

Impact
- Identify source-impact pathways from simulation and/or observational data using a lightweight and relatively quick-to-train tool, RFR
- Once a RFR pathway network is built, it can queried to answer questions such as, "what is the most direct path from a source to impact?" and "what is the relative strength of a given pathway?"

EXAMPLE

$$W_t = 0.9W_{t-1} + \varepsilon_{W_t}$$
$$X_t = 0.8X_{t-1} + 0.5W_{t-1} + \varepsilon_{X_t}$$
$$Y_t = -0.9W_{t-1} + \varepsilon_{Y_t}$$
$$Z_t = 0.3X_{t-1} + 0.5Y_{t-1} + \varepsilon_{Z_t}$$



Synthetic Dataset

- Set of coupled equations
- Interdependences between variables
- Autocorrelation on some variables
- All dependences are 1 time step, but the algorithm will look for 1 and 2 time step dependences

	id	source	Target	Weight
0	1	W_-1	W	10.000000
9	10	W_-1	X	7.022222
10	11	W_-2	X	5.200000
11	12	X_-1	X	5.200000
18	19	W_-1	Y	9.066667
27	28	W_-1	Z	1.777778
28	29	W_-2	Z	5.111111
29	30	X_-1	Z	5.111111
30	31	X_-2	Z	2.000000
31	32	Y_-1	Z	2.000000

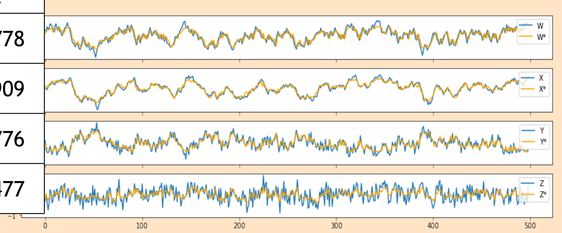
Feature Importances

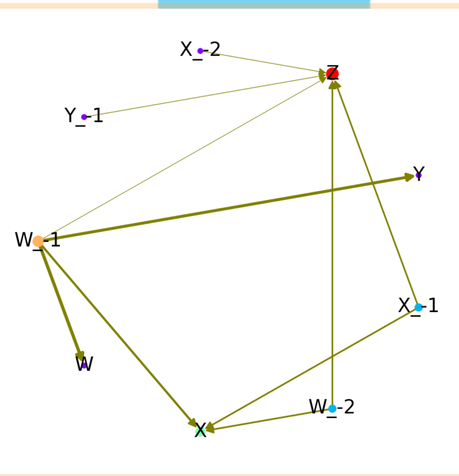
- Importance Weights are scaled from 0-10
- Some weights/connections are pruned using a random variable as a cutoff point
- The values are calculated using SHAP feature importance

Verification metrics

Variables	W	X	Y	Z
Existing edges	W_{t-1}	W_{t-1} W_{t-2}, X_{t-1}	W_{t-1}	X_{t-1}, W_{t-2} X_{t-2}, Y_{t-1} W_{t-1}

Variable	R^2
W	0.778
X	0.909
Y	0.776
Z	0.477





RFR Feature Pathway Graph

- Thicker lines indicate stronger pathways/weights
- Transitive pathways exist such as (W_{t-2},X)
- These extra pathways are 'Not wrong, but not correct'

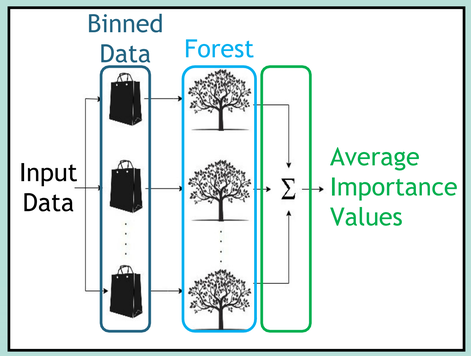
APPROACH

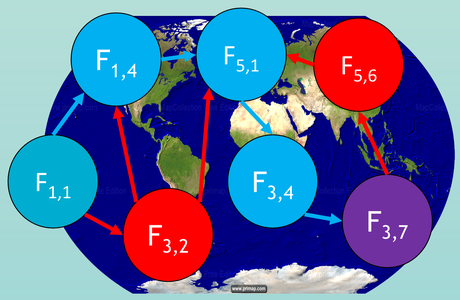
Random Forest (RF): ensemble of decision trees

Random Forest Regressors (RFRs): Machine Learning (ML) predictive models
- Require training data comprised of pairs of inputs and outputs
- Average predictions produced by decision trees in each RF
- Once trained, can predict the output given a set of input variables based on feature splits

Feature Importances: a measure of the predictive power that a particular input variable has on a given output variable
- If an input variable that has a high feature importance value were removed, the ML model would have a harder time predicting the output accurately

RFR is commonly used for and **well-developed** for **regression**, **classification** and **prediction** tasks.
- Our approach extends RFR for purpose of **discovering pathways** by using it to perform full pairwise analysis of input features and identify/rank dominant connections
- We can convert the Feature Importances into a weighted directed graph where the nodes are the a set of predetermined features of interest





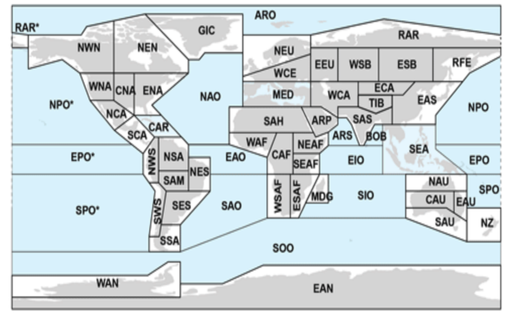
WORK IN PROGRESS

Expanding RFR-based approach to spatio-temporal data
- IPCC region maps
- signature-based clusters

Automatically Identify transitive pathways
- How do we identify them when we move away from ground truth problems?
- Are they incorrect or is it an alternative pathway?
- How does it affect individual feature importance weights?

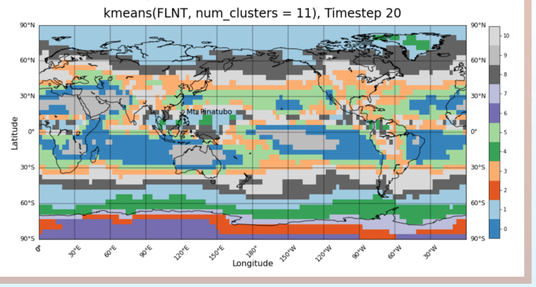
Develop way to query the graph for the best pathway chain from source to impact
- It is our ultimate goal to identify pathway chains
- Some pathway chains will be more relevant than others
- There may be choke points within the graph that all pathway chains go through

IPCC region maps



Iturbide et al., 2020. https://essd.copernicus.org/articles/12/2959/2020/

Signature-based Clusters



U.S. DEPARTMENT OF ENERGY
National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology &Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.
SAND No. _____

CONTACT INFO
Matt Peterson
mgpeter@sandia.gov

Sandia National Laboratories