# Data properties, tests, cleaning

# ML approaches

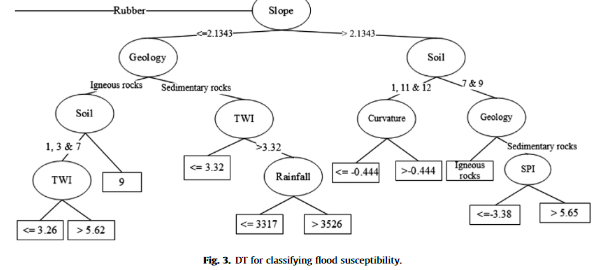
## Rule based decision trees

### Tehrany etal 2013 Spatial prediction

“…of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS”

section 4.1.3 p73&74

“There are different ways to perform the DT modeling such as Chi-squared Automatic Interaction Detection (CHAID), Exhaustive CHAID, Classification and Regression Trees (CRT), Quick, Unbiased, and Efficient Statistic Tree (QUEST). (Roe et al., 2005). In this paper CHAID was used and at each step where the conditioning (predictor) factor has the strongest relationship with the dependent variables are chosen. The classes of each conditioning factor will be merged if they are not significantly different between them with respect to the dependent variable (Berry and Linoff, 1997). Exhaustive CHAID is made by the modification of CHAID algorithm that examines all possible splits for each conditioning factor (Biggset al., 1991). CRT first applies the segmentation on the dataset to dividing the homogenous parts based on their interaction with dependent variable. And the results of this process will extract the pure nodes (Kusiak et al., 2010). CHAID method is a proper choice for hazard modeling among others, as its performance is fast and it has the ability of multi-way node splitting (Kusiak et al.,2010). The CHAID algorithm was implemented in SPSS V.19 and the criteria were selected based on the literature (Lee and Park,2013; Ture et al., 2009). Splitting and merging categories can be selected between the ranges of 0–1. The value of 0.9 and 0.001 were set for splitting and merging parameters which achieved through the many times trial and error. The next criterion was selected for Chi-square statistic.”



Creates a nice DT plot. IDK if it’s as powerful as e.g. xgBoost, but might be a nice preliminary step to generate these plots (unless they’re completely unrepresentative of the situation)

## NARX Neural Network

### Ahmed etal 2019 Forecasting GRACE

“Nonlinear autoregressive with exogenous input (NARX) neural network is used in this study to predict and forecast TWSGRACE time series over ten African watersheds. Given our previous knowledge of the spatiotemporal variabilities in TWSGRACE over the African watersheds and how they are controlled by natural and anthropogenic interventions [9,74], NARX model was selected because it provides effective, efficient, and powerful (1) nonlinear systems modelling and predictive tool, (2) learning algorithm that discovers, and is not affected by, the long temporal dependence in the model outputs and/or inputs [75,76], and (3) faster convergence in reaching the optimal weights of the connections between neurons and/or inputs [77–80].”

### Hyndman & Khandakar 2008 Forecast

ARIMA & exponential smoothing timeseries models

“There is a widespread myth that ARIMA models are more general than exponential smoothing. This is not true. The two classes of models overlap. The linear exponential smoothing models are all special cases of ARIMA models—the equivalences are discussed in Hyndman et al. (2008a). However, the non-linear exponential smoothing models have no equivalent ARIMA counterpart. On the other hand, there are many ARIMA models which have no exponential smoothing counterpart. Thus, the two model classes overlap and are complimentary; each has its strengths and weaknesses. The exponential smoothing state space models are all non-stationary. Models with seasonality or non-damped trend (or both) have two unit roots; all other models—that is, non-seasonal models with either no trend or damped trend—have one unit root. It is possible to define a stationary model with similar characteristics to exponential smoothing, but this is not normally done. The philosophy of exponential smoothing is that the world is non-stationary. So if a stationary model is required, ARIMA models are better. One advantage of the exponential smoothing models is that they can be non-linear. So time series that exhibit non-linear characteristics including heteroscedasticity may be better modelled using exponential smoothing state space models. For seasonal data, there are many more ARIMA models than the 30 possible models in the exponential smoothing class of Section 2. It may be thought that the larger model class is advantageous. However, the results in Hyndman et al. (2002) show that the exponential smoothing models performed better than the ARIMA models for the seasonal M3 competition data. (For the annual M3 data, the ARIMA models performed better.) In a discussion of these results, Hyndman (2001) speculates that the larger model space of ARIMA models actually harms forecasting performance because it introduces additional uncertainty. The smaller exponential smoothing class is sufficiently rich to capture the dynamics of almost all real business and economic time series.”

# Evaluation

### Ahmed etal 2019 Forecasting GRACE

“The performance of the NARX model was evaluated using several standard coefficients; Root mean square error (RMSE), scaled RMSE (R\*), correlation coefficient (r), Nash–Sutcliffe efficiency (NSE) coefficient, and the seasonal adjusted NSE (NSE\*) that are commonly used in the analysis and evaluation of statistical model results, in our case the NARX model outputs [114]. For example, NSE was found to be the best objective function for evaluating the overall fit between the predictive and observed values [115]. Likewise, r values were found to be sensitive to the differences between modeled and observed data including the extreme values (i.e., outliers) [116].”

### Servat & Dezetter 2009 selection

P315 “objective functions” = model evaluation metrics?

“There is a very large number of objective functions in the literature. It was not the intention to carry out an exhaustive study thereof which is, in practice, not possible. Five different objective functions were therefore studied, three of which were the subject of numerous uses in hydrological modelling. The remaining two were set up considering elements which seemed important.”

Crec

CrecBi

Fortin

Nash / Nash-Sutcliffe Efficiency NSE

SExpER

### (Elith et al., 2008) Working guide

Where BRT models are developed with CV, *statistics on predictive performance can be estimated from the subsets of data excluded from model fitting (see Fig. 4 and Supplementary material*). For the model presented previously, the CV estimate of prediction error was close to that on independent data, although slightly overoptimistic (Fig. 5, compare solid and open circles; Table 3, see estimates on independent data compared with CV). This is a typical result, although the ability of CV to estimate true performance varies with data set and species prevalence. In small data sets, CV estimates of predictive performance may be erratic, and repeated and/or stratified cross-validation can help stabilize them (Kohavi 1995). Predictive performance should not be estimated on training data, but results are provided in Table 3 to show that BRT overfits the data, regardless of careful model development (Table 3; see difference between estimates on training and independent data). While overfitting is often seen as a problem in statistical modelling, our experience with BRT is that prediction to independent data is not compromised – indeed, it is generally superior to other methods (see e.g. comparisons with GLM, GAM and multivariate adaptive regression splines, Elith et al. 2006; Leathwick et al. 2006). The flexibility in the modelling that allows overfitting also enables an accurate description of the relationships in the data, provided that overfitting is appropriately controlled.