Alexander Kowarik,
Angelika Meraner
Statistik Austria
April, 2017

# An automatic procedure for selecting weights in kNN imputation

# Disclaimer

- No solutions, just some ideas . . .

# kNN

```
kNN(data, variable = colnames(data), metric = NULL,
  k = 5, dist_var = colnames(data), weights = NULL,
  numFun = median , catFun = maxCat , makeNA = NULL,
  NAcond = NULL, impNA = TRUE, donorcond = NULL,
  mixed = vector(), mixed.constant = NULL,
  trace = FALSE , imp_var = TRUE, imp_suffix = "imp",
  addRandom = FALSE, useImputedDist = TRUE,
  weightDist = FALSE)
```

- *dist_var* - variables to be used for distance computation
- `weights` - weights for the distance computation
- *numFun*, *catFun* - funtion for aggregating (*sampleCat*, *maxCat*).
- *addRandom* - add a random value to the distance computation (with very low weight) to break ties
- *weightDist* - use the distances of the knns in the aggregation function

# Distance computation

- Extension of the Gower distance

$$d_{i,j} = \frac{\sum_{k=1}^{p} w_k \delta_{i,j,k}}{\sum_{k=1}^{p} w_k} \ , \tag{1}$$

- $w_k$ is the weight for the contribution of variable $k$ to the total distance.
- $w_k$ should represent the "importance" or the "dependency" of variable $k$ to the target variable
- If the number of target variables and/or (possible) distance variables is large $\Rightarrow$ automatic choice of weights

# Method 1 - Random forest

- Random forest is applied to the data set (or a subsample) and the importance is used as weight variable.
- $V1$: Directly apply RF to the data sets
- $V2$: Create binomial dummy variables for categorical variables and apply RF and then kNN to it

# Method 2 - Lasso shrinkage

- Lasso regression excludes variables, which are presumably not important for estimating the model.
- Can be applied to a subsample
- $V1$: within 1 standard deviation of minimal cross-validated error
- $V2$: minimum cross-validated error

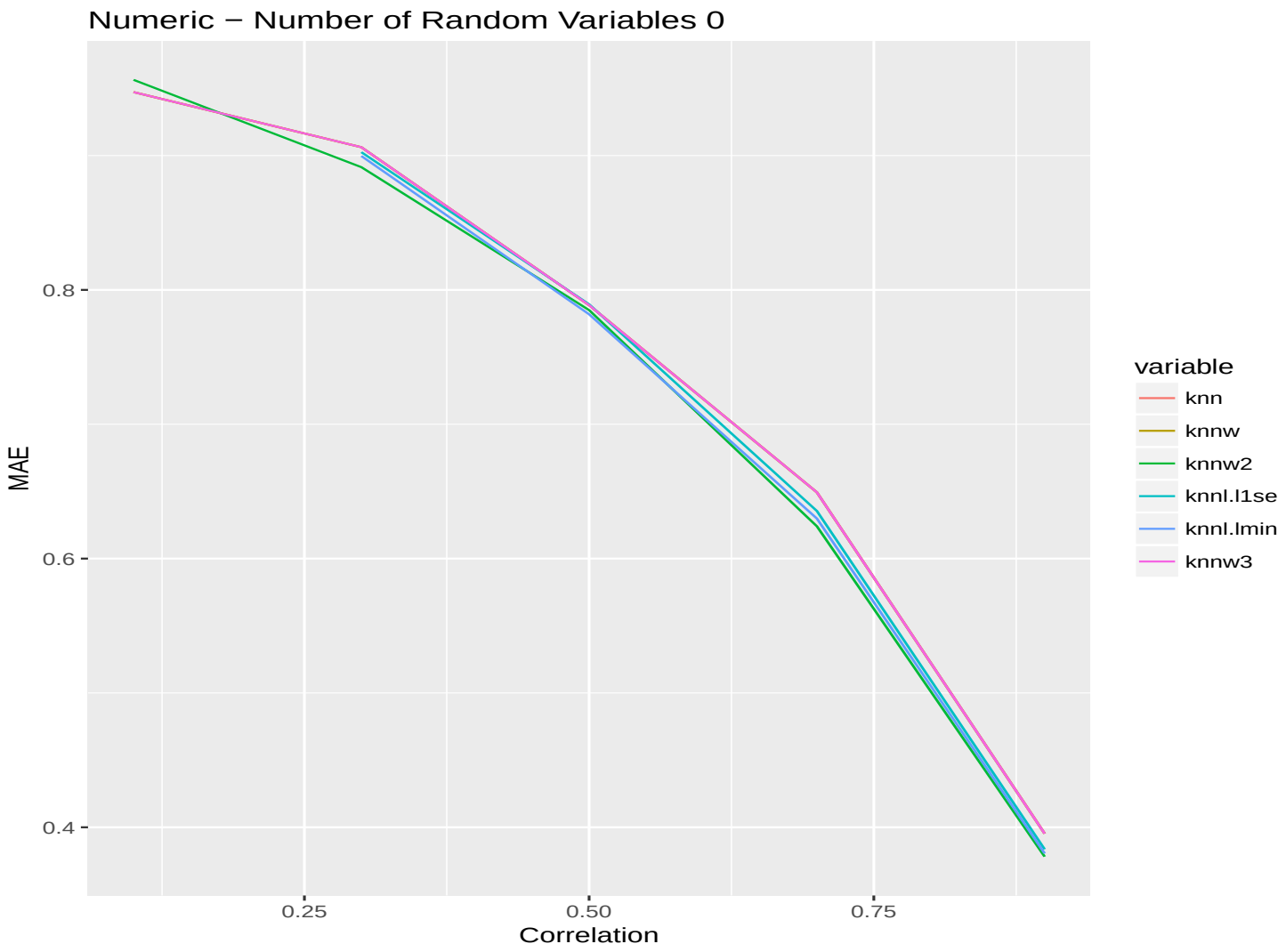
# Method 3 - Single feature accuracy

- Apply kNN to the observed data points (or a subsample) with only one distance variable at a time
- Use the inverse of the error (classification error or mean absolute error depending on the type of target variable) as weight for the kNN

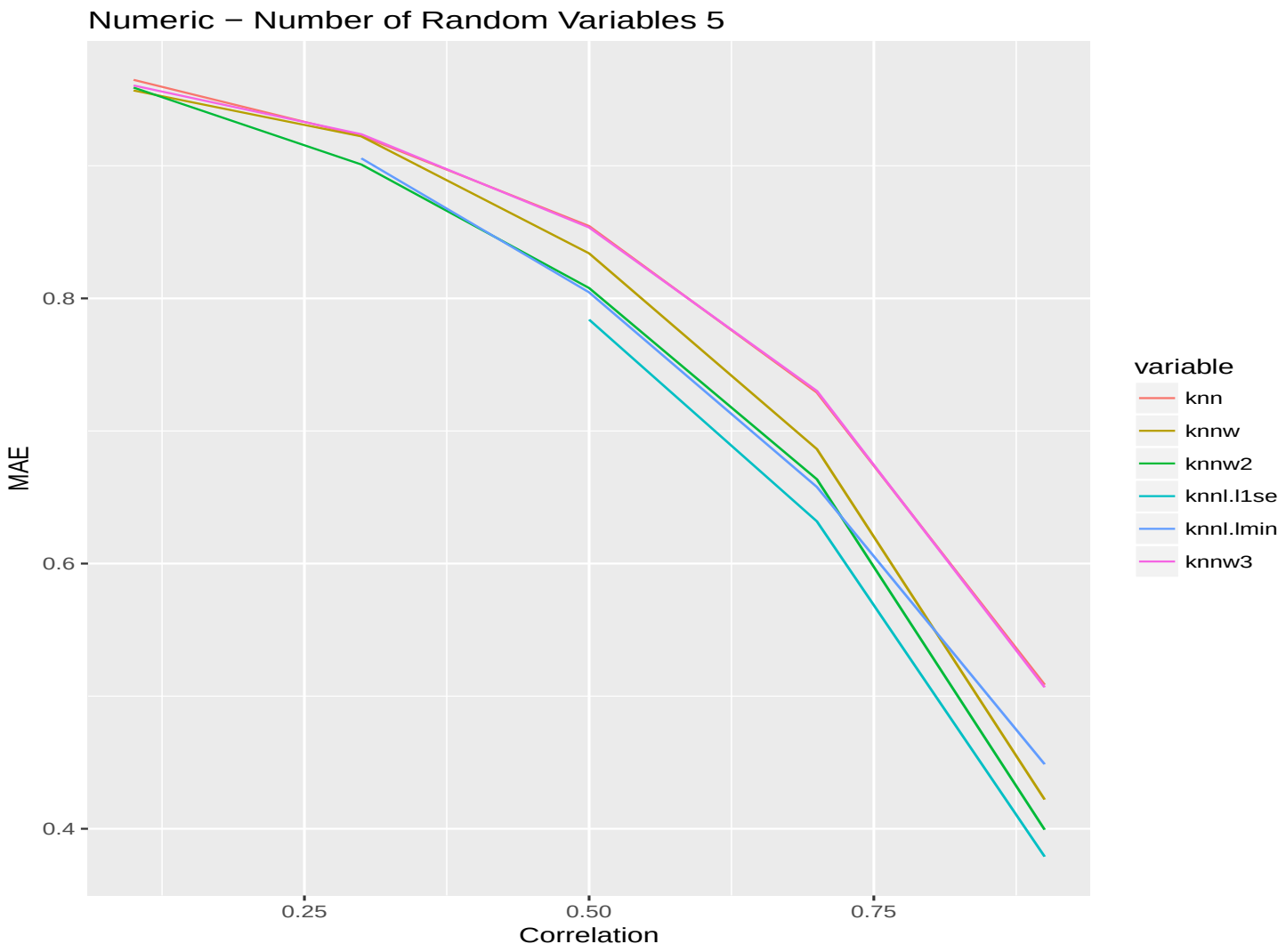
# (small) Simulation study

- Simulated data set from a normal distribution with varying correlation
- 2 numerical variables
- 2 mixed variables
- 2 categorical variables
- and a varying number of uncorrelated random numerical variables

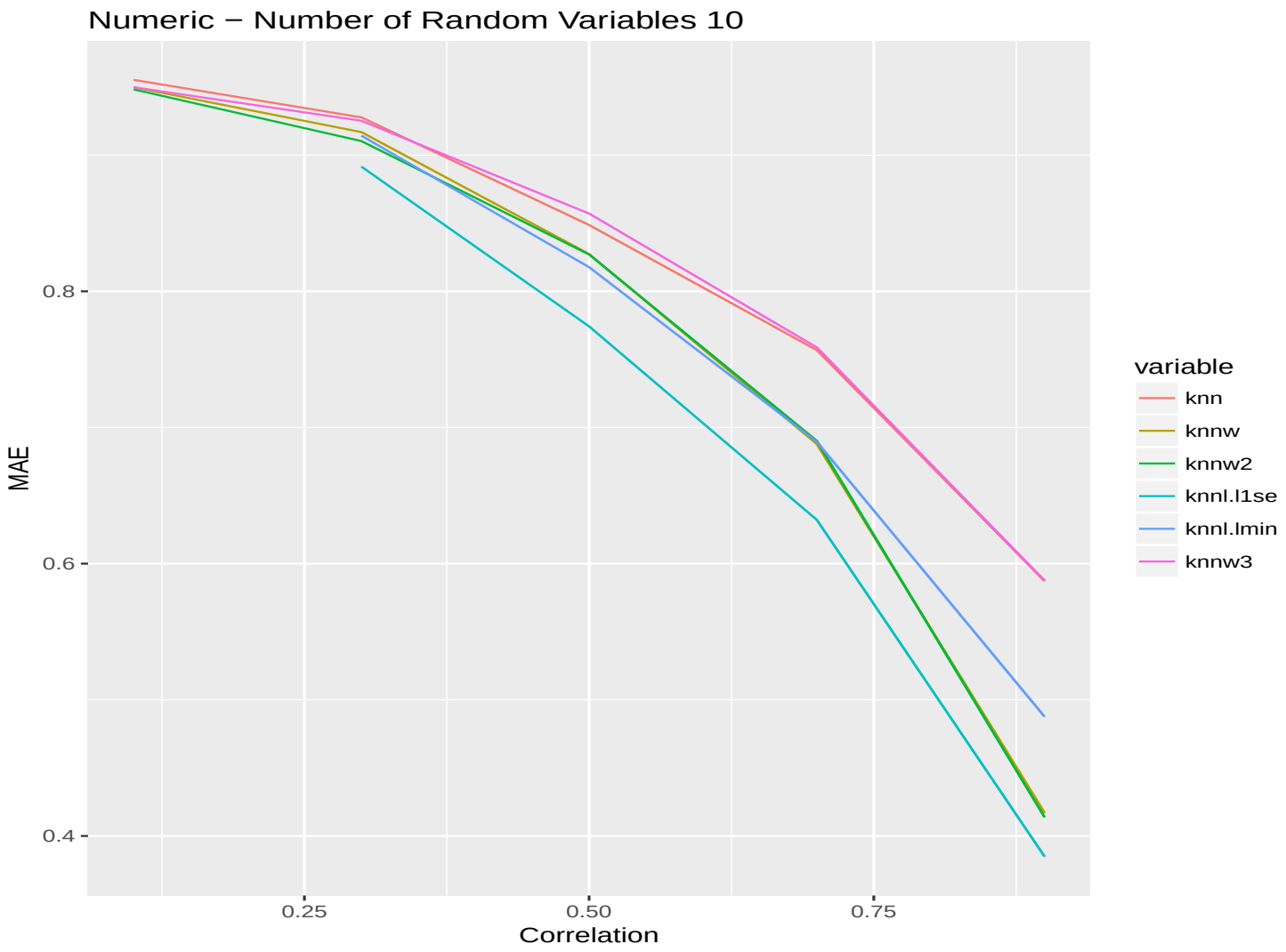
# Results - numeric target variable



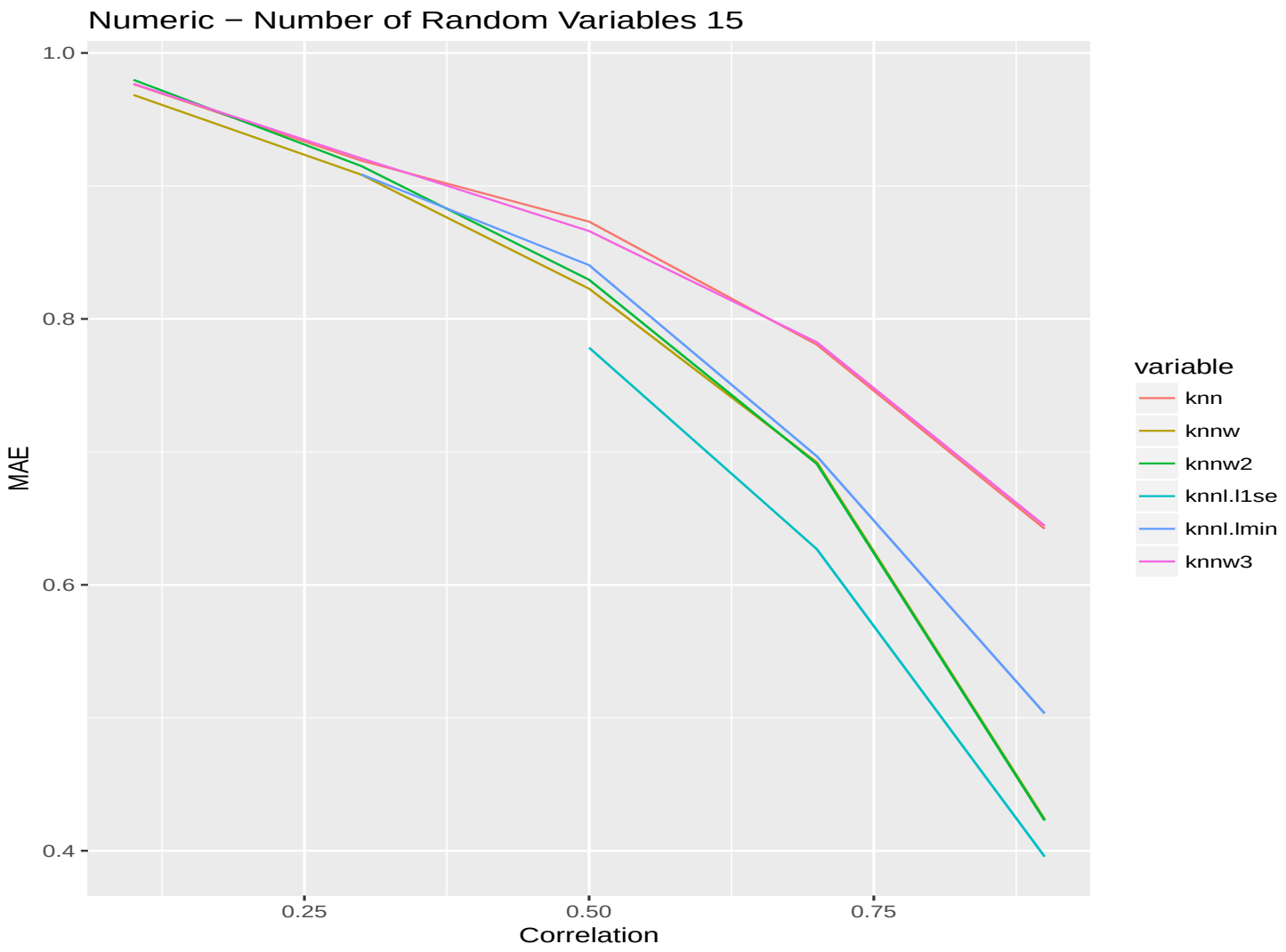
# Results - numeric target variable



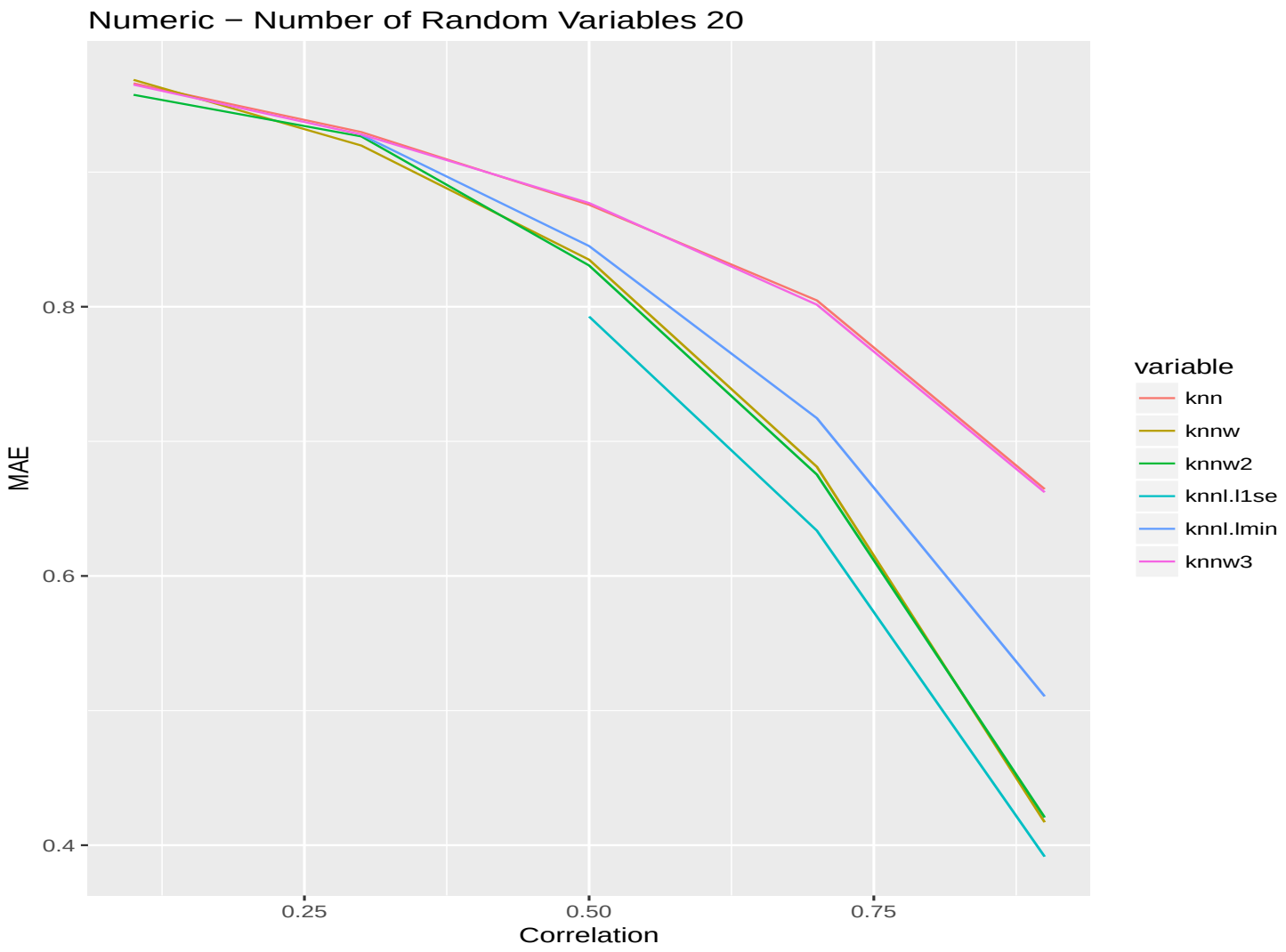
# Results - numeric target variable



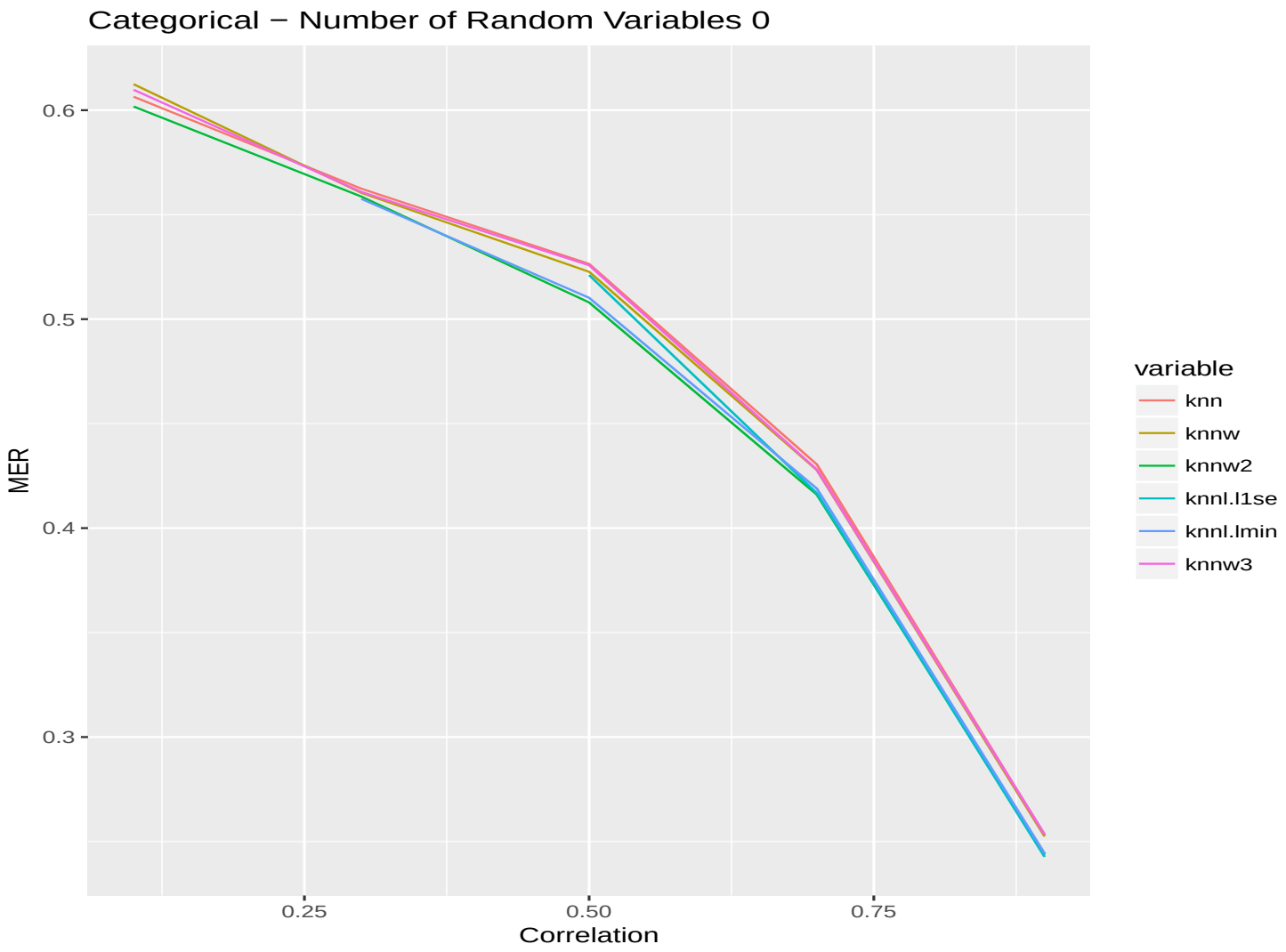
# Results - numeric target variable



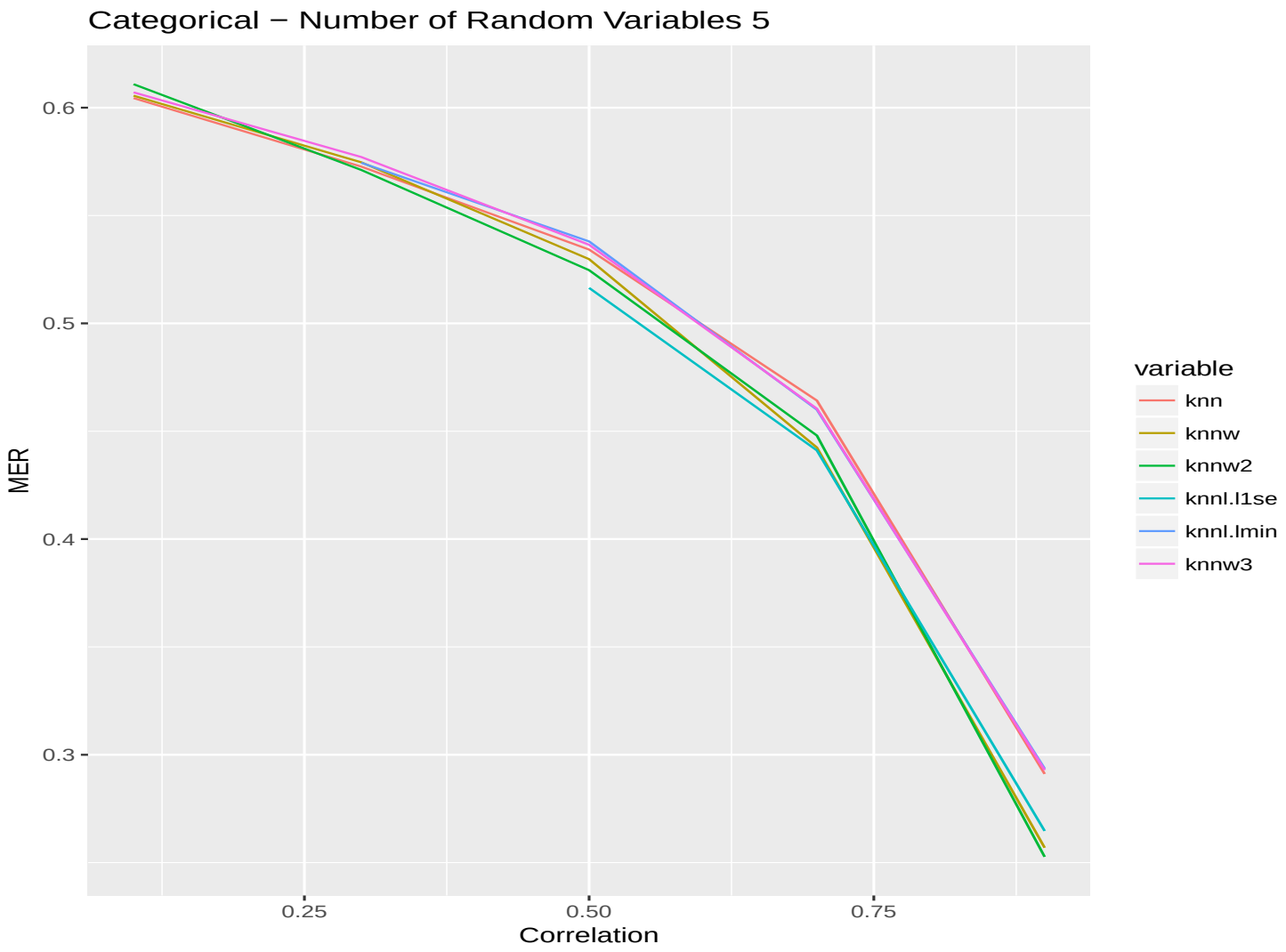
# Results - numeric target variable



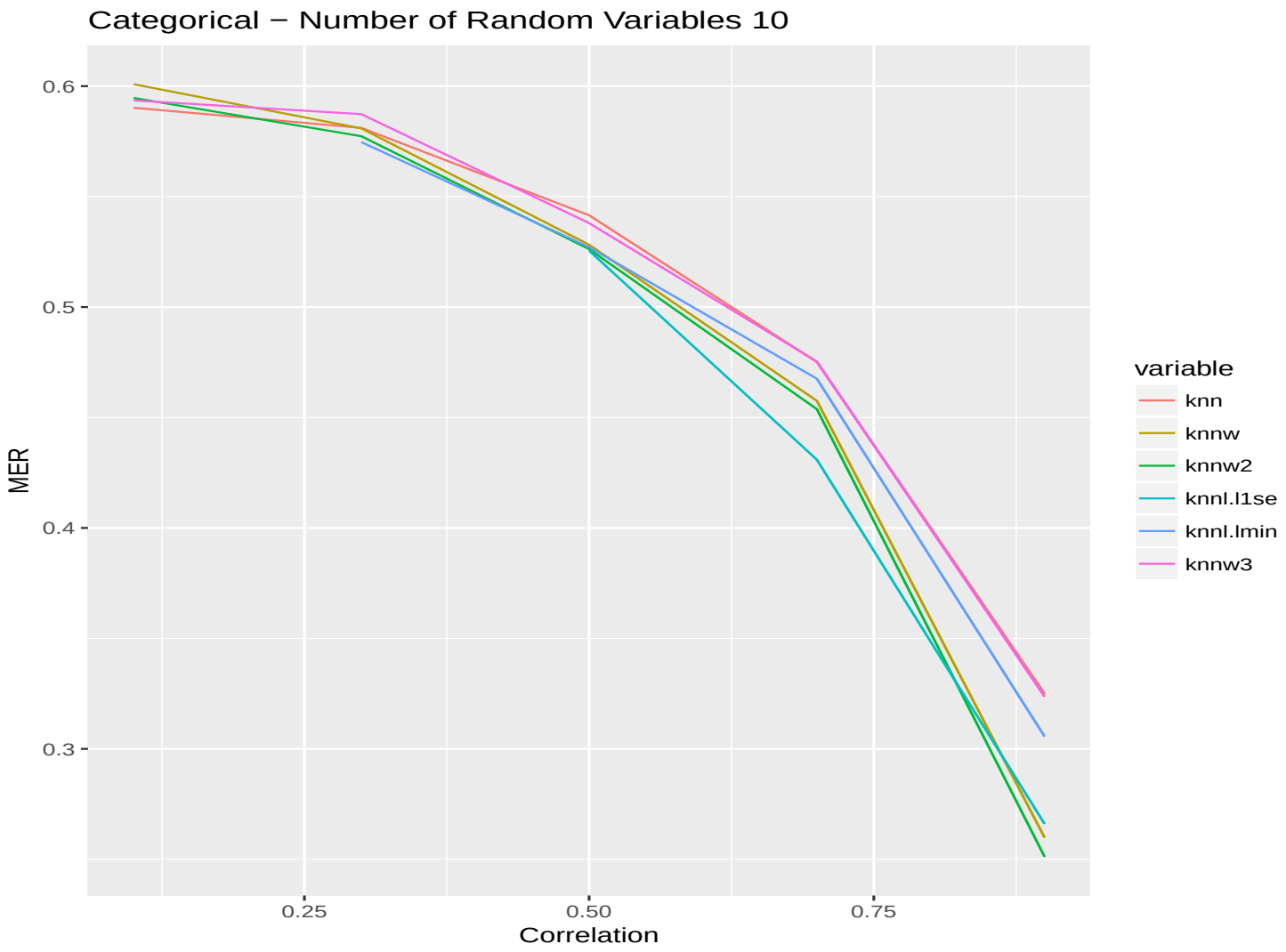
# Results - categorical target variable



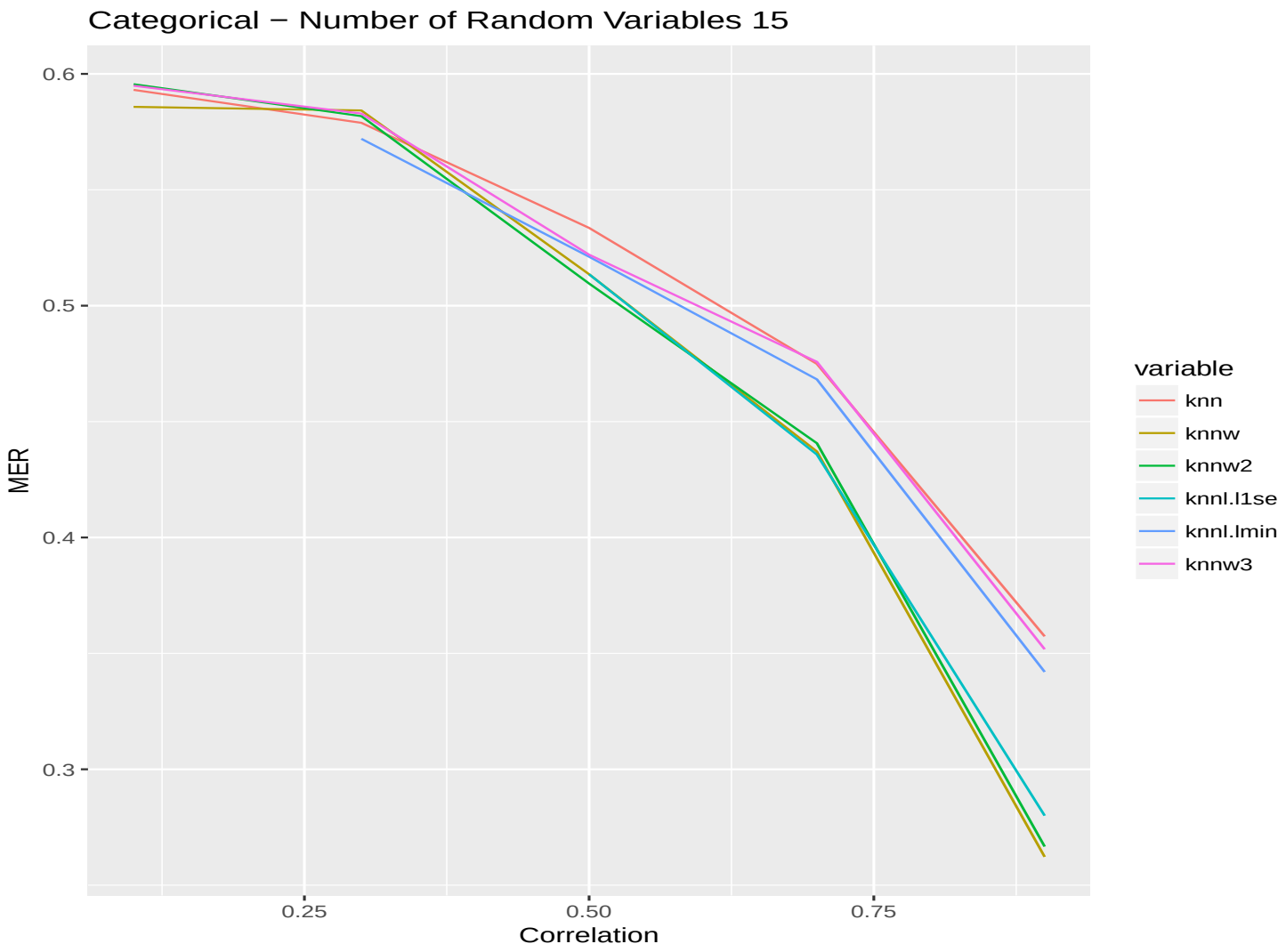
# Results - categorical target variable



# Results - categorical target variable



# Results - categorical target variable



# Results - categorical target variable