The Land Cover Classification by the Crowdsourced Mapping

116010282 詹铸成
116010093 黄志伟
116010067 何吉米
116010299 张熹哲

CONTENTS

Data Denoising

- Chi-square test and replaced by mean
- Chi-square test and replaced by median
- Boxplot method

Identify the outliers in the predictor

	Predictors		
obvs	...		
within a class	...	X_{ij}	...
	...		

p-value of X_{ij} within the column $< p_0$

Identify the outliers in the response

	Predictors		
obvs	...		
within a class	...	NA NA NA	...
	...		

Delete

num of NA > K

	Predictors		
obvs	...		
within a class	...	NA	...
	...		

Deal with the outliers

	Predictors		
obvs	...		
within a class	...	mean of col	...
	...		

	Predictors		
obvs	...		
within a class	...	median of col	...
	...		

Limitation

P-value is a sensitive measurement.

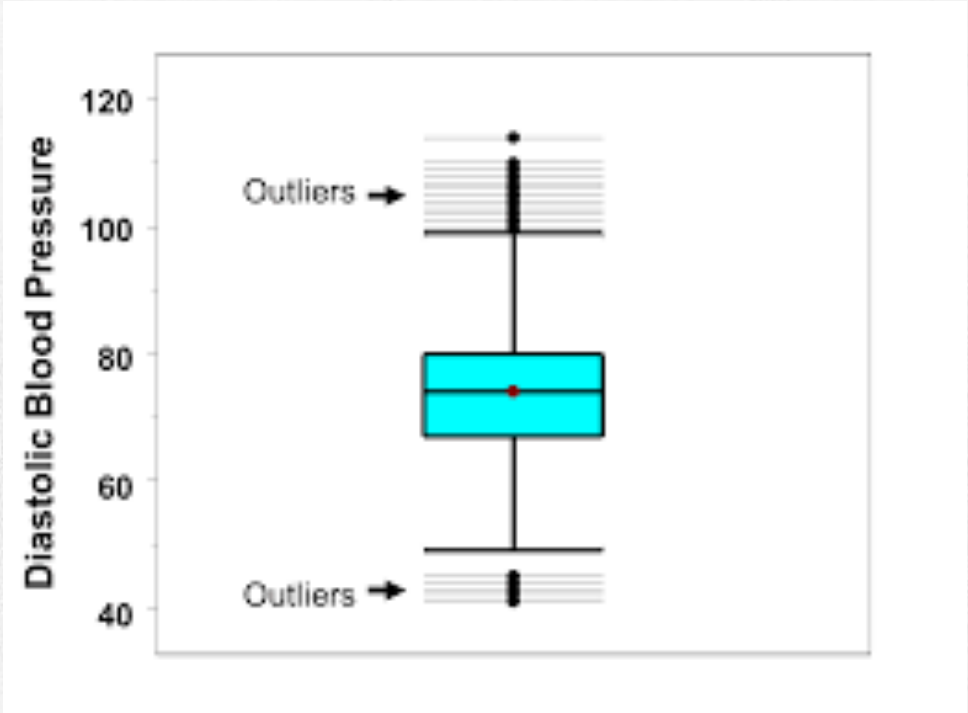
water.data	205 obs. of 29 variables
water.data.prune	31 obs. of 29 variables
impervious.data	969 obs. of 29 variables
impervious.data.prune	34 obs. of 29 variables

Hard to find a uniform p-value and K

Too many possible combinations make the optimization very hard

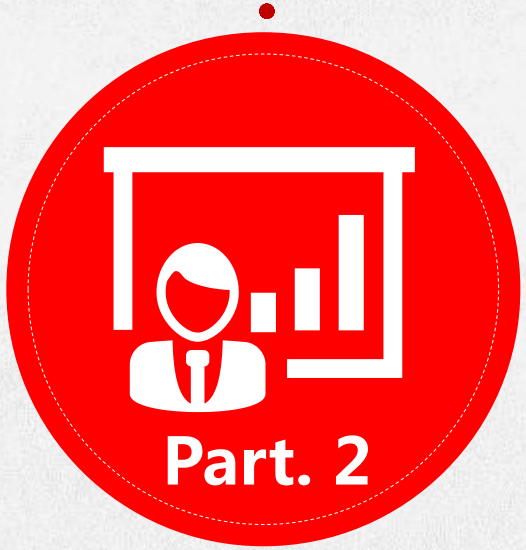
The interpretability of the model is not enough

Boxplot method



Find outliers

Delete outliers



Datasets and Model Selection

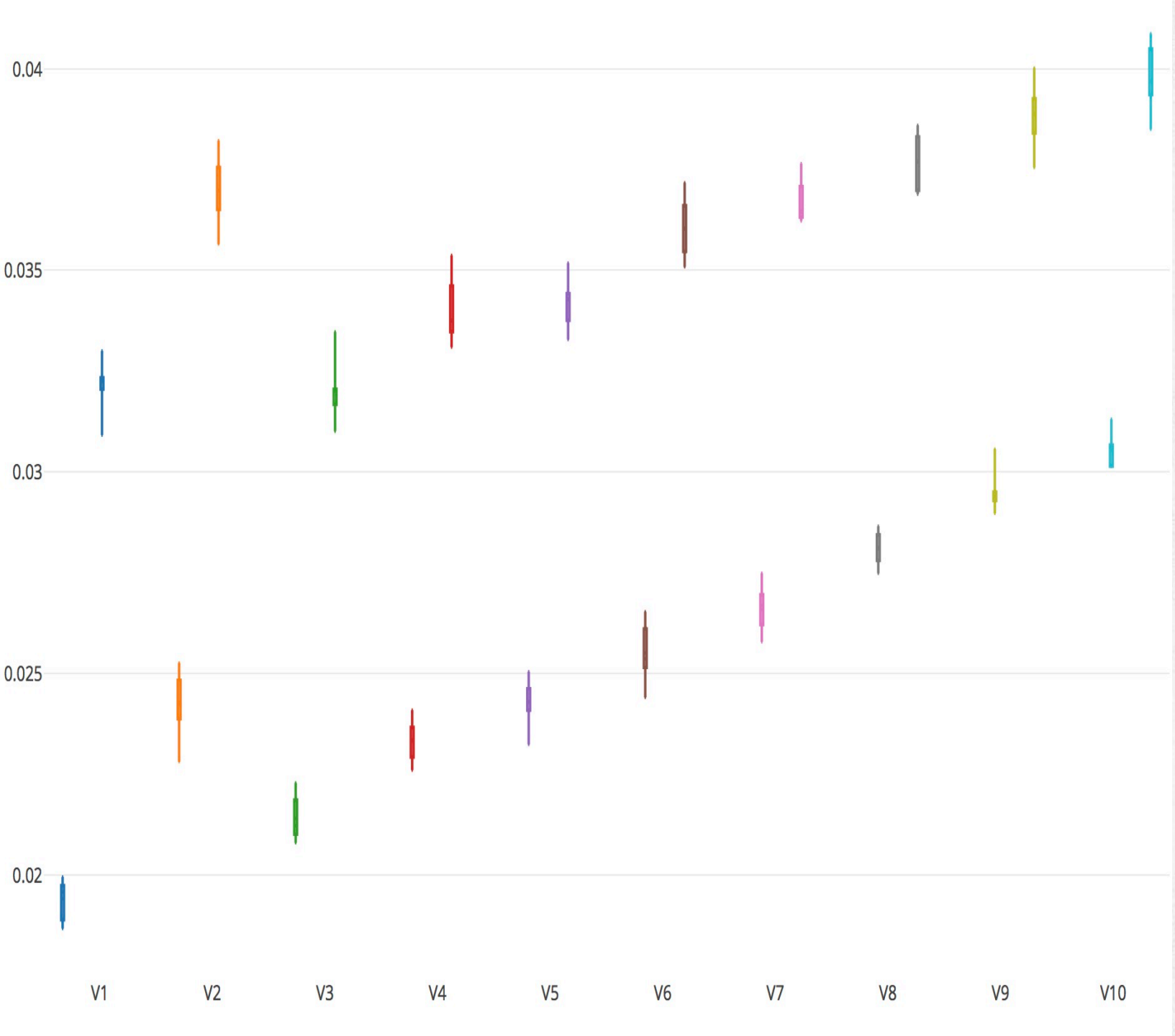
- Model selection
- Dataset selection

Method for Selection

The model candidates: KNN, Tree, Logistics, LDA and QDA

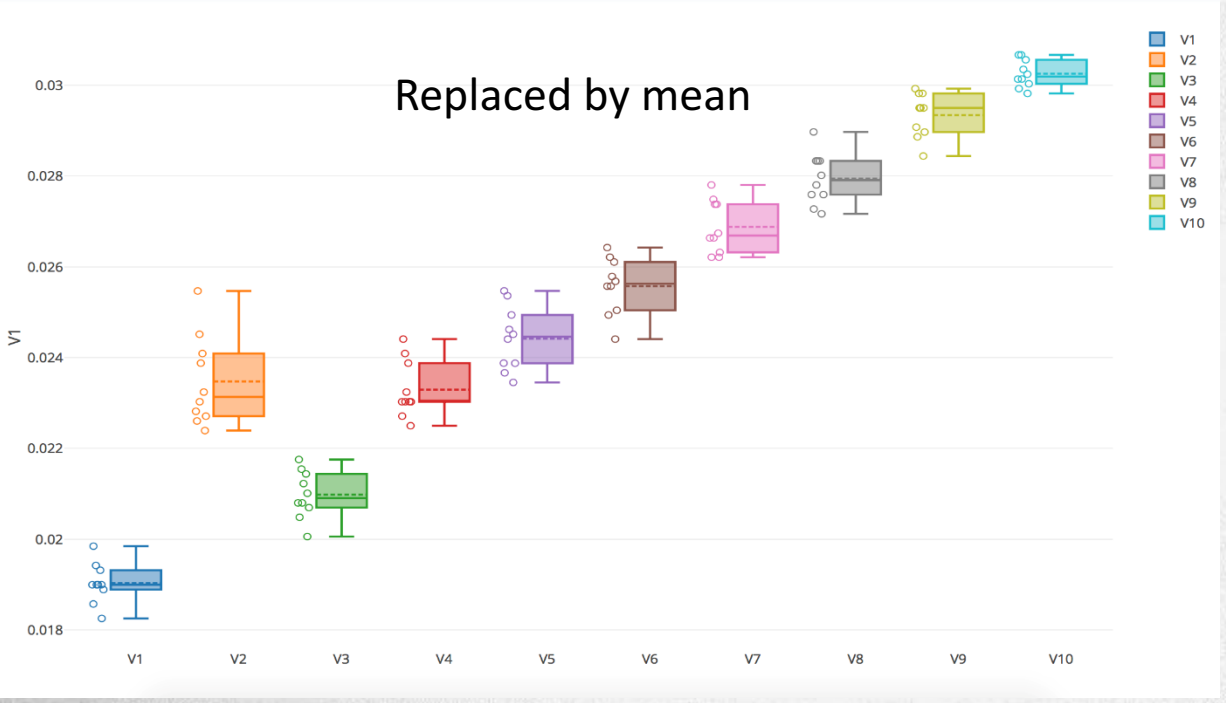
Based on the bias-variance tradeoff theory,
we chose the 10-fold cross validation.

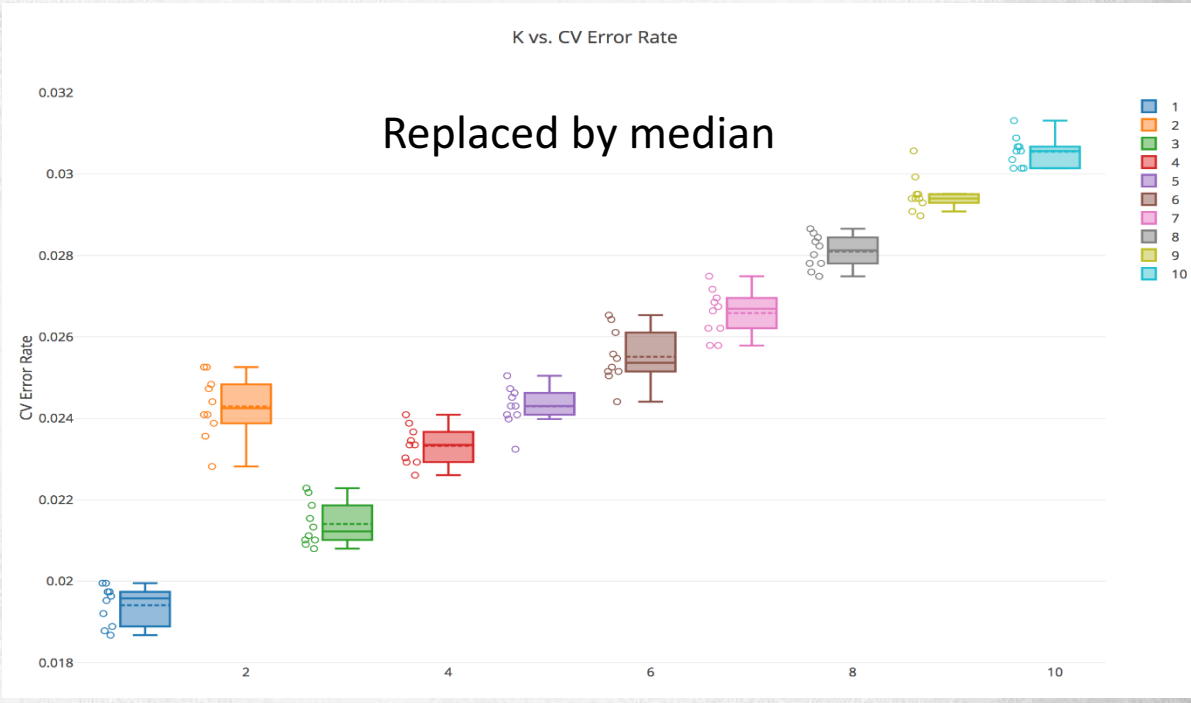
To reduce the randomness, we ran each model ten times by 10-fold cross validation, resulting in an error set.



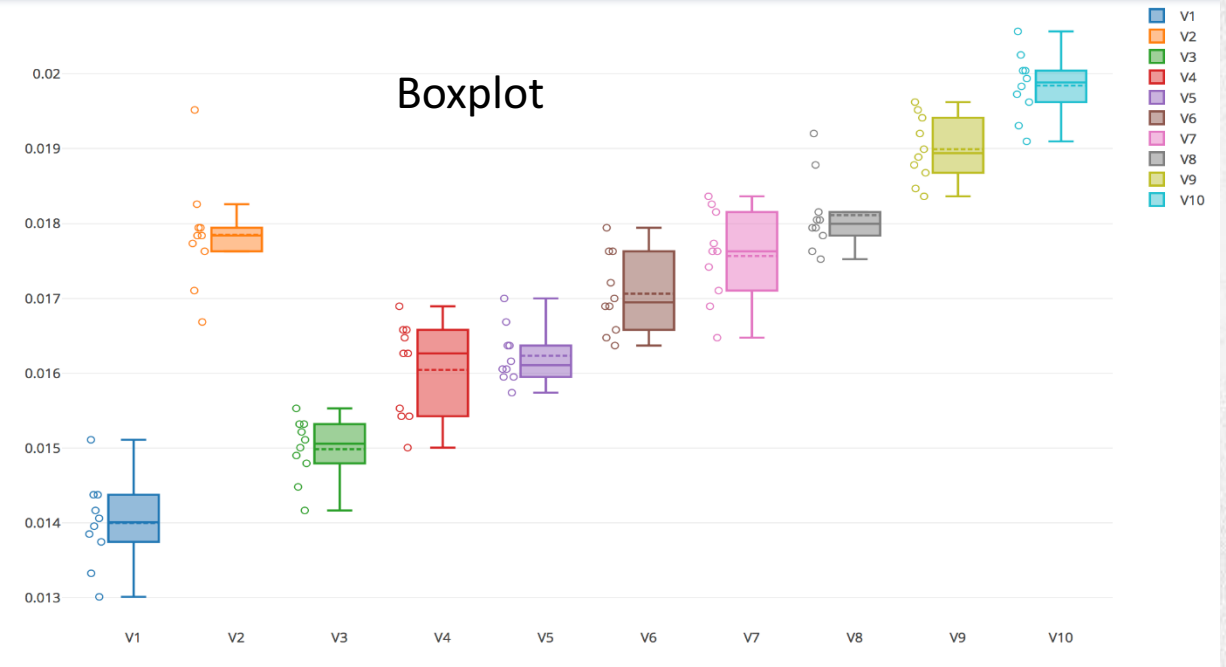
The upper boxes is the 10-Fold Cross-Validation of KNN method based on raw data set.

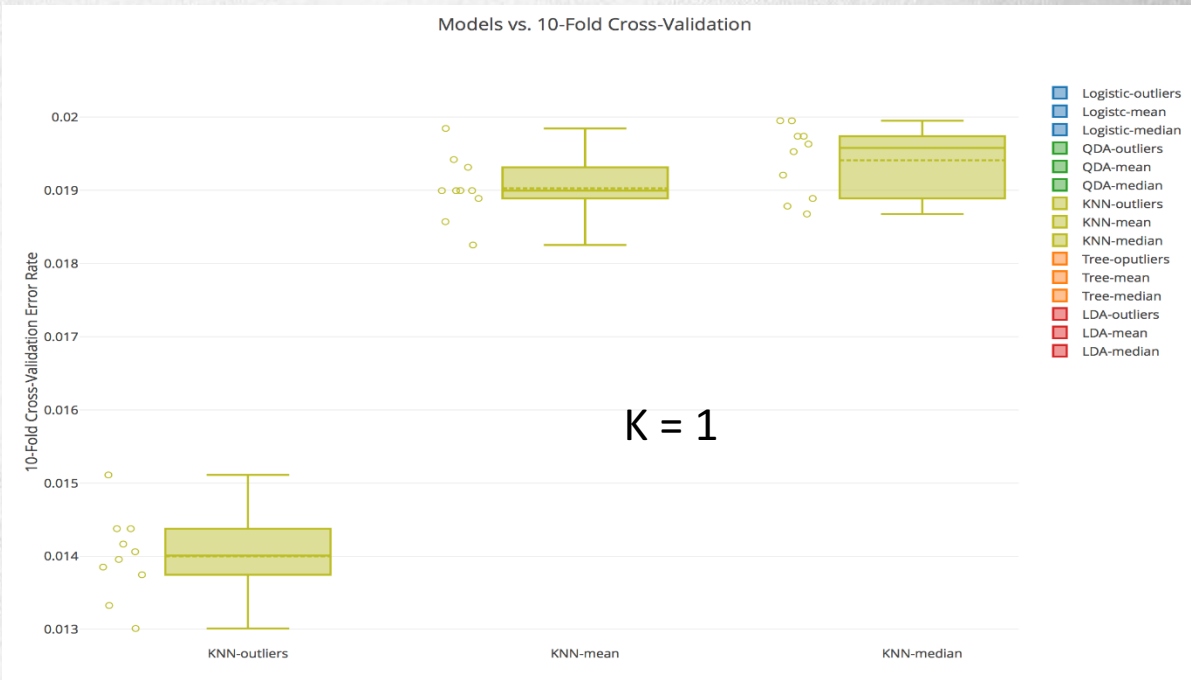
The lower boxes is the 10-Fold Cross-Validation of KNN method based on data set after denoising by replacing outliers by means of its columns





One can tell that k = 1 is the optimized k for 3 data set.





Non-parametric model selection

Data Denoising

Datasets and Model Selection

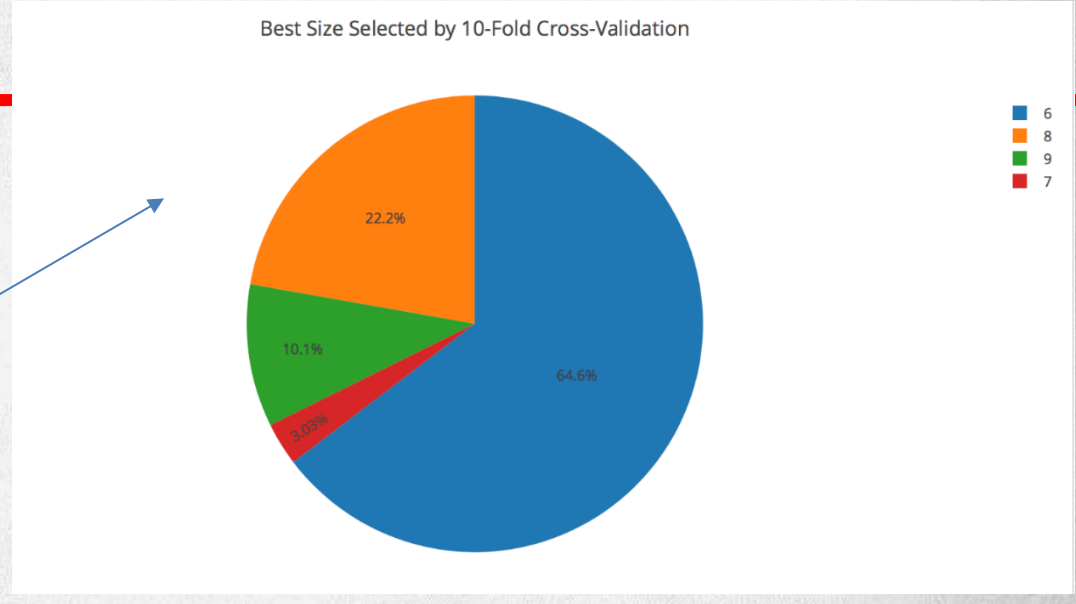
Conclusion

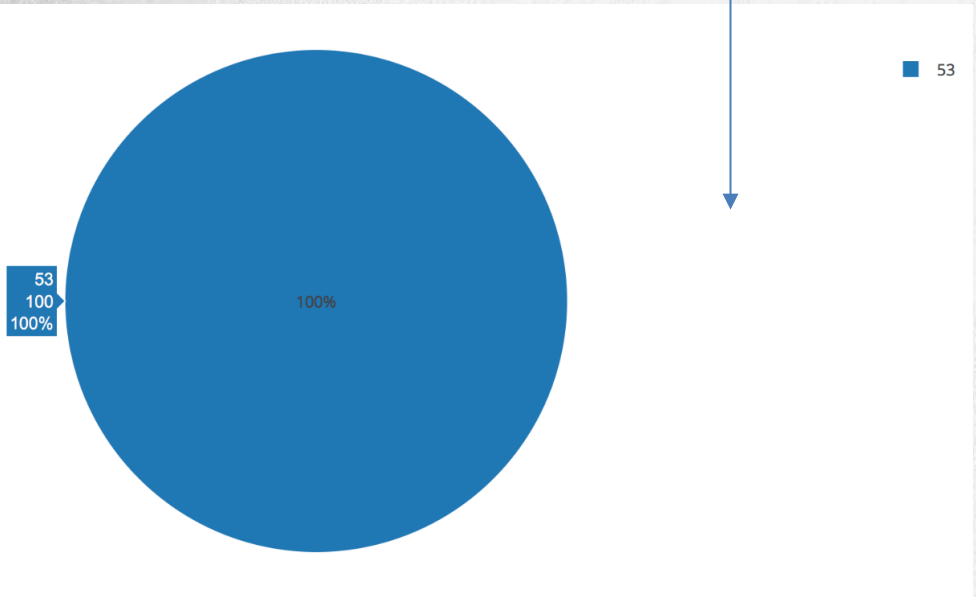
Based on 3 data sets, we selected 3 best sizes of trees.

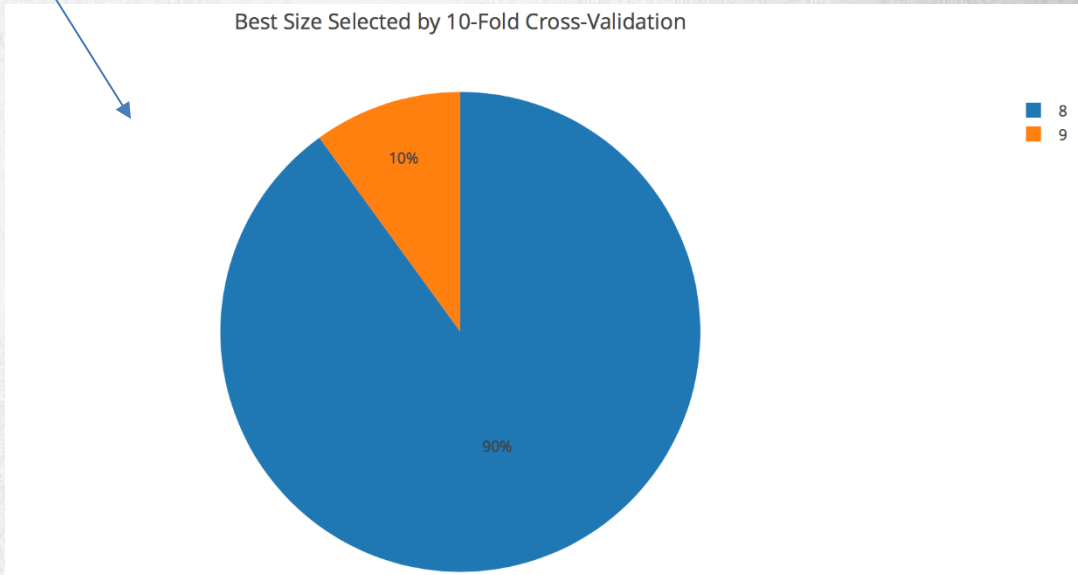
For data set denoising by replacing outliers by median, we selects size = 6.

For data set denoising by replacing outliers by mean, we selects size = 8.

For data set denoising by boxplot, we selects size = 53.

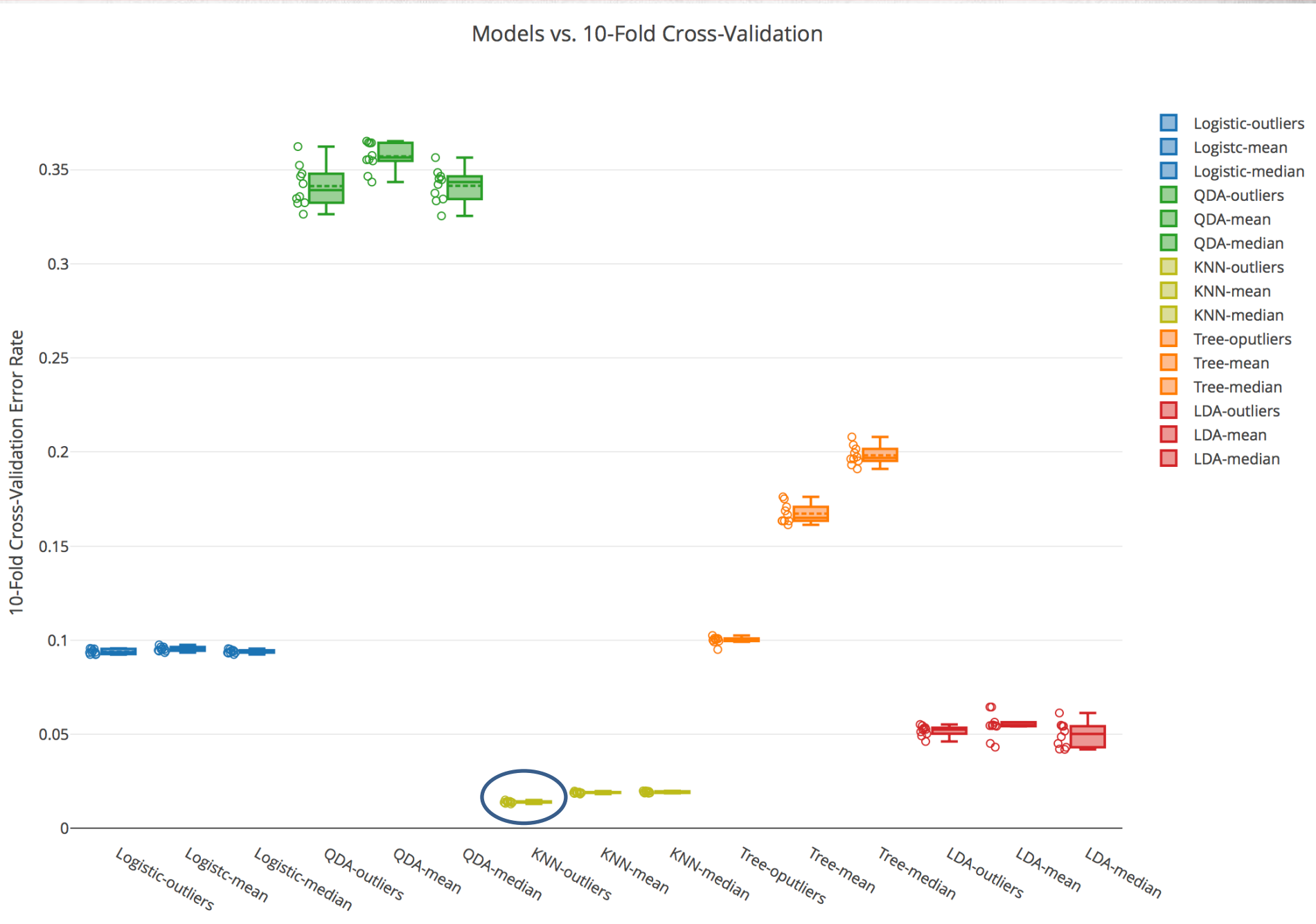






Overall model selection

- Data Denoising
- Datasets and Model Selection
- Conclusion



Conclusion

- Results
- Limitation

Test Error

Data Denoising

Dataset and Model Selection

Conclusion

61.333%

Meets With Our Expectation

KNN usually preforms well if the dataset is large it is hard to make some assumption for the predictors.

By intuition, the relationship between the predictors and the response is highly nonlinear and complicated, so K=1 which is the most flexible one is the best.

Since the chi-square test is based on the normal assumption, the boxplot method outperforming the chi-square method also indicates that the KNN method is better.

Reference: Ye H.2016. CBN Data：2009-2016 year double 11 data Yearbook. Retrieved from http://www.199it.com/archives/543531.html

Limitation

The error rate is not THAT satisfying

Although the boxplot method has greater interpretability, the parameter selection is still subjective.

Thank you for your listening!