Evaluating CLEMI

CLustering to Evaluate Multiple Imputation

Anthony Chapman

Dr Steve Turner Dr Wei Pang Dr Lorna Aucott

Dept. of Applied Medical Sciences, University of Aberdeen
Dept. of Computing Science, University of Aberdeen
e-mail: r01ac14@abdn.ac.uk

Outline

1. Testing Dataset 1
2. Mini-ABDN Summary
3. Discussion

Content

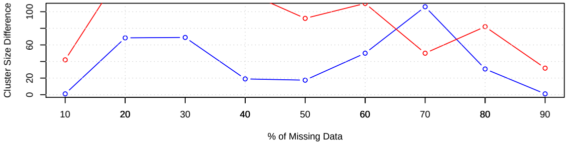
1. Testing Dataset 1
2. Mini-ABDN Summary
3. Discussion

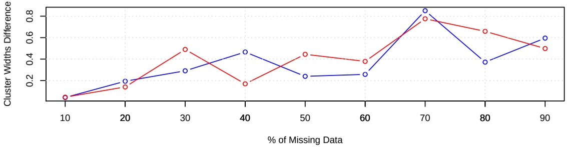
Forest Fires Summary

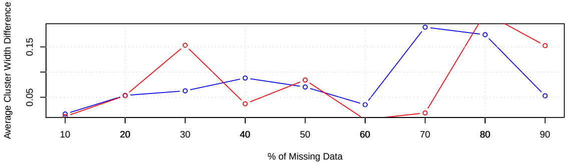
Forest Fires:

- 517 records, 13 variables
- Mixed numerical and categorical variables
- Will select $x\%$ of records randomly
- Will remove a random amount of variables from the selected records
- Chosen increments of 10% (10%-90%)

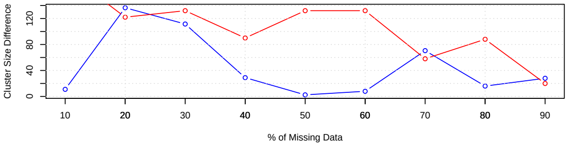
Evaluation ctn.

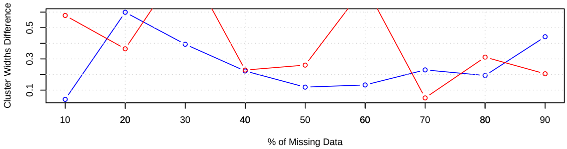


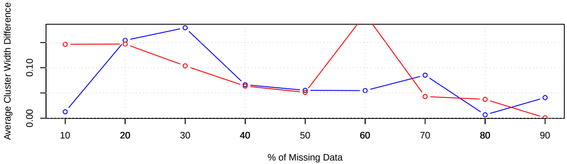




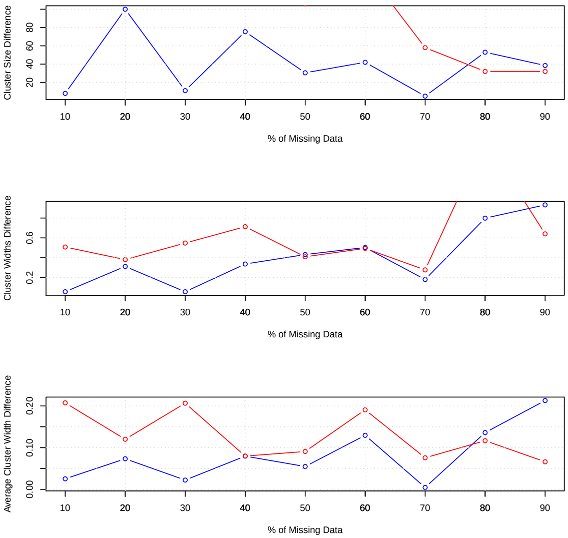
Evaluation ctn.



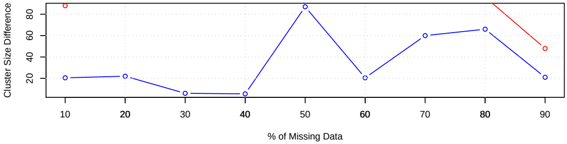


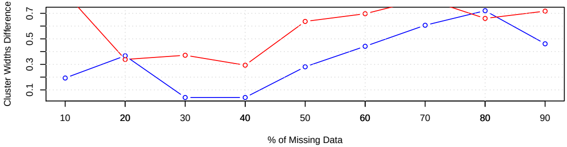


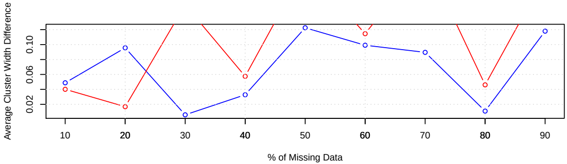
Evaluation ctn.



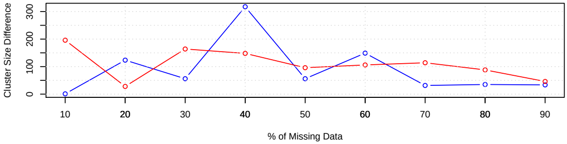
Evaluation ctn.

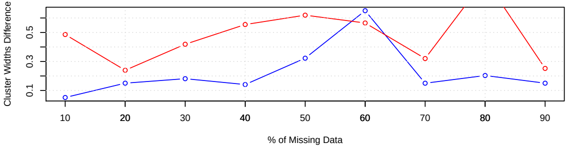


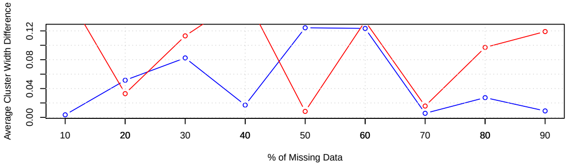




Evaluation ctn.







Forest Fires Outcomes

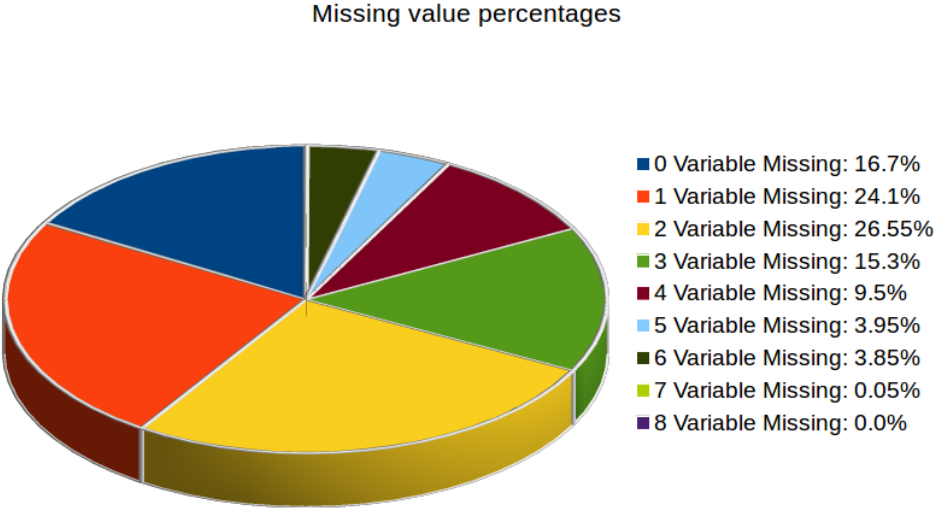
Outcomes:

- MICE works best when 20% to 80% of records contain missing values
- If less than 20% of the records have missing values, MICE might not be so effective. Could be as MICE needs lots of records with missing values to impute, more missing records means better prediction.
- If a dataset has more than 80% missing records MICE might not be so efficient, too much missingness could give false relationships and thus wrong imputation

Content

1. Testing Dataset 1
2. **Mini-ABDN Summary**
3. Discussion

Missing Values in Mini-ABDN

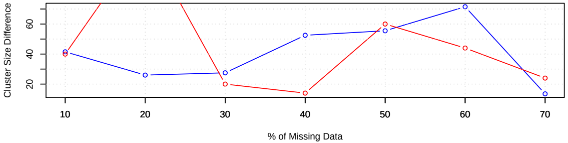


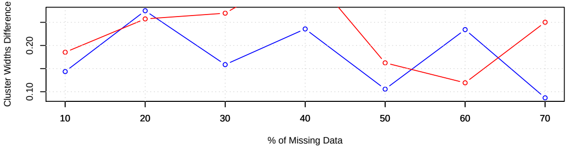
Mini-ABDN

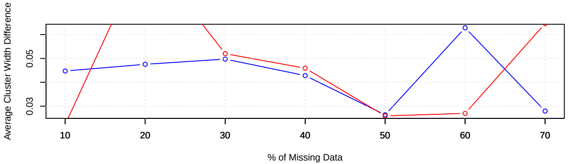
Mini-ABDN:

- 2000 records, 9 variables
- Mixed numerical and categorical variables
- Can be used to find the amount of missing allowed for imputation
- Will allow increments of 10% missing per records.
- Chosen increments 20% to 90%

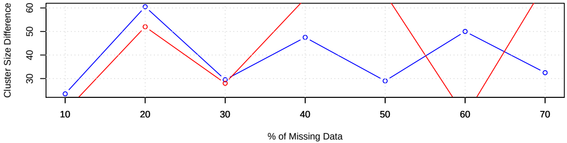
Evaluate ABDN ctn.

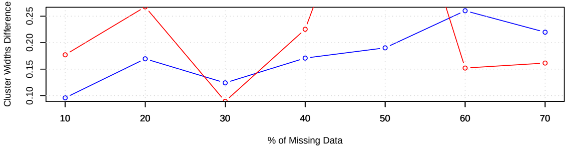


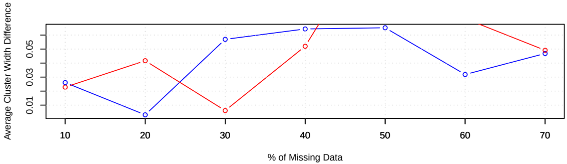




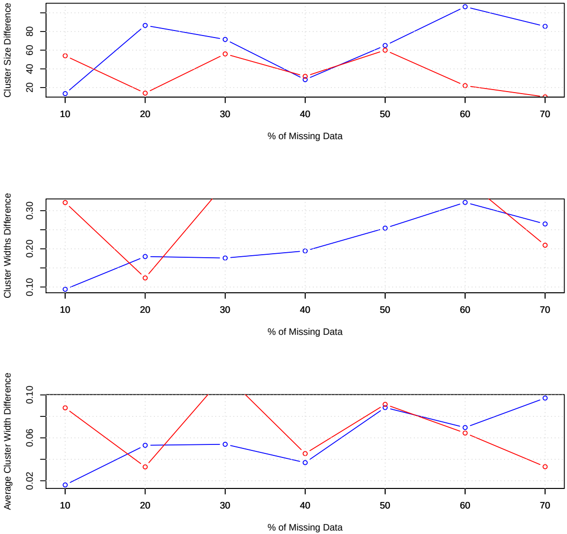
Evaluate ABDN ctn.



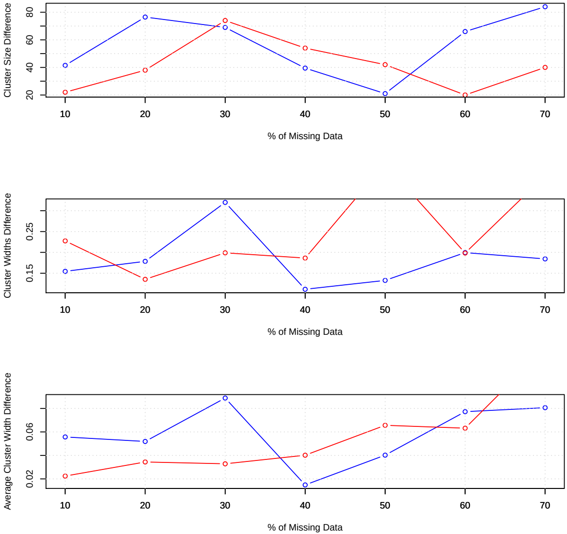




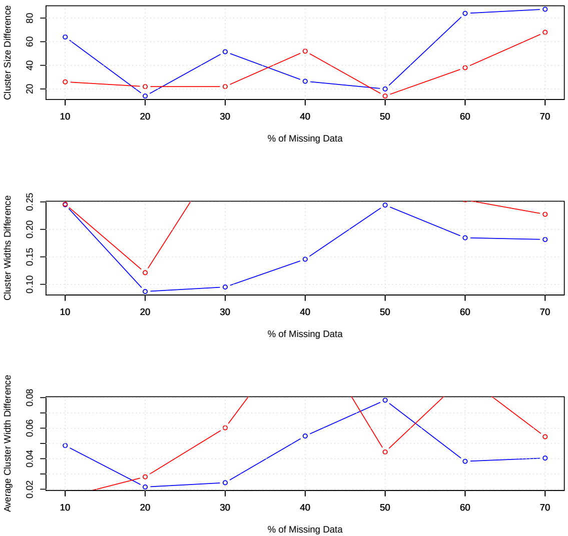
Evaluate ABDN ctn.



Evaluate ABDN ctn.



Evaluate ABDN ctn.



Mini-ABDN Outcomes

Outcomes:

- MICE works best when there is 40% to 80% missingness
- If a dataset has less than 40% missing variables on each record, might not be best to use MICE
- If a dataset has more than 80% missing variabes on each records, only keep the records with less than 80%

Content

1. Testing Dataset 1
2. Mini-ABDN Summary
3. Discussion

Discussion & Conclusion

Limitations

- Output is subjective
- Some may over-interpret the results
- What if the complete subset is too small

Outcomes

- Optimised number of ignored records
- Compare different imputation methods
- Optimize imputation features

To Consider

- Use modelling to verify the outcome
- Use more imputation method

Thanks & Questions

Thanks for your attention!
Question & Comments