QUEEN'S UNIVERSITY BELFAST
SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Data Analysis & Visualisation

CSC3062

BEng (CS & SE), MEng (CS & SE), BIT & CIT

Dr Reza Rafiee

Semester 1 – 2019/2020

Missing data & multiple imputation modelling

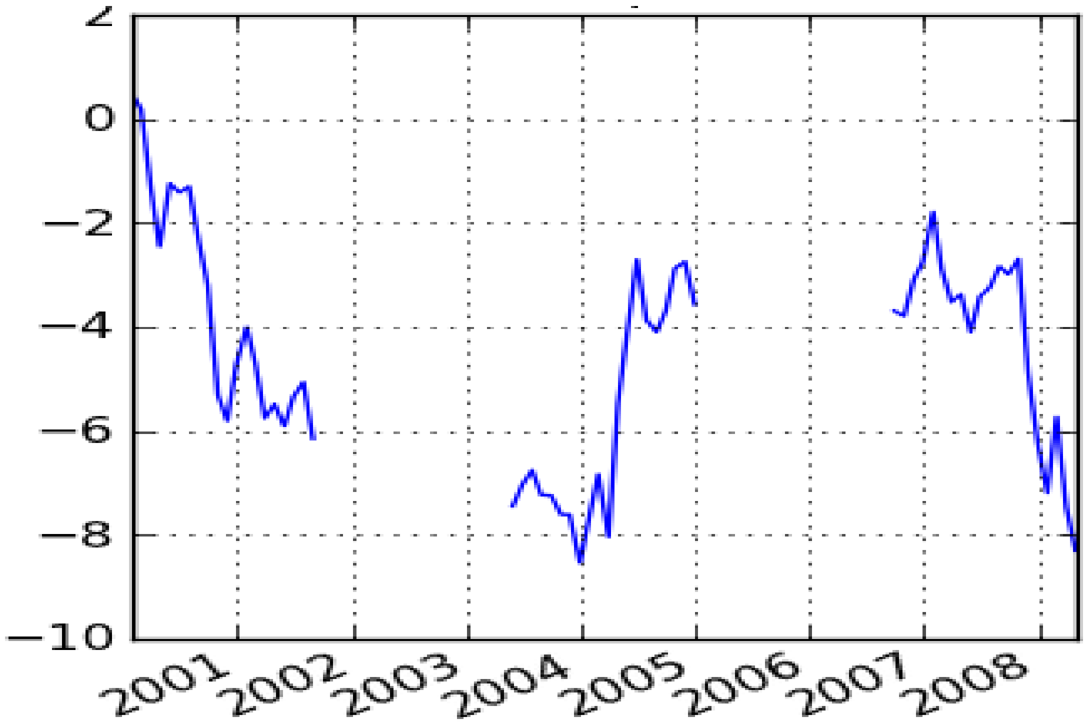
QUEEN'S UNIVERSITY BELFAST

SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Missing data is everywhere

In almost any research you perform, or any data analysis task, there is the potential for missing or incomplete data.

Continuous data (signal)

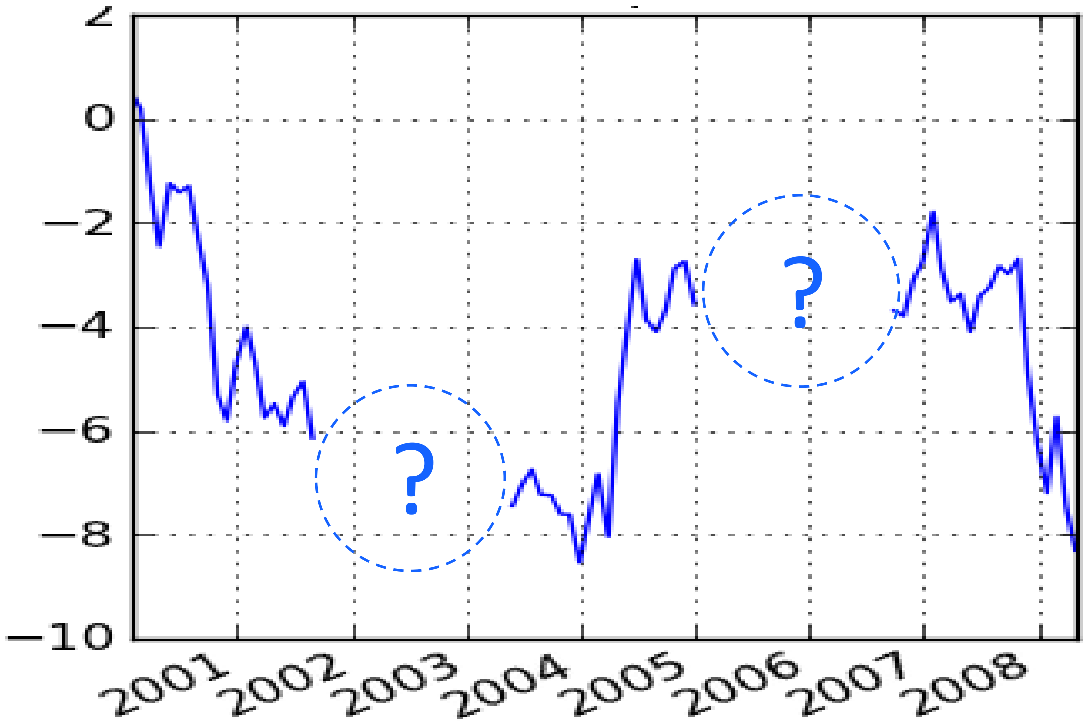


	Sample 1	Sample 2	…
cg00583535	NA	0.317394283	…
cg18788664	1	0.192024985	…
cg08123444	0.532659205	0.867010408	…
cg17185060	0.774338632	0.70392815	…
cg04541368	0.079894678	0.659468157	…
cg25923609	0.109138594	0.600225461	…
cg06795768	0.04605561	0.870753578	…
cg19336198	0.713845623	0.707326444	…
cg05851505	NA	0.981375746	…
cg20912770	0.039837473	0.0646352	…
cg09190051	1	0.336904134	…
cg01986767	NA	NA	…
cg01561259	0.133410152	0.113869472	…
cg12373208	NA	0.04628476	…
cg24280645	0.163157983	0.088281769	…
cg00388871	0.239179168	0.308942014	…
cg09923107	0.091227524	0.121433558	…

Missing data is everywhere

In almost any research you perform, or any data analysis task, there is the potential for missing or incomplete data.

Continuous data (signal)



	Sample 1	Sample 2	…
cg00583535	NA	0.317394283	…
cg18788664	1	0.192024985	…
cg08123444	0.532659205	0.867010408	…
cg17185060	0.774338632	0.70392815	…
cg04541368	0.079894678	0.659468157	…
cg25923609	0.109138594	0.600225461	…
cg06795768	0.04605561	0.870753578	…
cg19336198	0.713845623	0.707326444	…
cg05851505	NA	0.981375746	…
cg20912770	0.039837473	0.0646352	…
cg09190051	1	0.336904134	…
cg01986767	NA	NA	…
cg01561259	0.133410152	0.113869472	…
cg12373208	NA	0.04628476	…
cg24280645	0.163157983	0.088281769	…
cg00388871	0.239179168	0.308942014	…
cg09923107	0.091227524	0.121433558	…

How to address missingness?

- Addressing missing data is one the most common challenges in data analysis and machine learning when analysing real-world data.
- Many data analysis and machine learning algorithms (or techniques) rely on a complete dataset.
  - Most visualisation functions in various data analytics programming
  - Most classification and clustering methods, etc.

Example of missingness in prediction

	CSC3062_108_2	CSC3062_109_4	CSC3062_110_4	CSC3062_112_2	CSC3062_113_2	CSC3062_125_4	CSC3062_127_3
feature_1	0.290874776	0.89080331	0.81032173	0.094939587	0.150149242	0.894331320	0.275124510
feature_2	0.810257812	0.08627098	0.24416510	0.821881924	0.709218768	0.103017283	0.909203863
feature_3	0.865808069	0.92201287	0.89654937	0.956021386	0.735552896	0.952562500	0.889457035
feature_4	0.862365076	0.06557660	0.08144940	0.985600007	0.858727746	0.053551341	0.050593115
feature_5	0.966055005	0.05415225	0.08579509	0.997462814	0.887684164	0.028919888	0.059173152
feature_6	0.983397001	0.06252419	0.10805568	0.998506562	0.950558010	0.046704008	0.309769218
feature_7	0.859219550	0.32792332	0.10845017	0.914788949	0.868725210	0.236932825	0.081102726
feature_8	0.771524676	0.70493114	0.80318701	0.236309397	0.217483490	0.749603835	0.806856586
feature_9	0.993792219	0.98066608	0.98448790	0.998409165	0.966818311	0.990438742	0.989619716
feature_10	0.442210237	0.05116329	0.06296946	0.701879320	0.040175667	0.053461366	0.117850085
feature_11	0.899340588	0.06391362	0.13571643	0.453133041	0.892199087	0.058884539	0.686559491
feature_12	0.993988972	0.99149929	0.98880779	0.997791094	0.984008240	0.988925295	0.992761839
feature_13	0.085722787	0.05524618	0.04647980	0.105029451	0.061793271	0.036679333	0.934434637
feature_14	0.044409295	0.02614883	0.02810520	0.030050428	0.028856994	0.037977469	0.872106236
feature_15	0.009247367	0.01671519	0.02064240	0.004755696	0.024785343	0.016187983	0.013172200
feature_16	0.728546654	0.08666581	0.05717338	0.871429414	0.740817463	0.062748032	0.082223218
feature_17	0.009349993	0.00775338	NA	0.011523751	0.009470701	0.009221186	0.006522704

Example of missingness in prediction

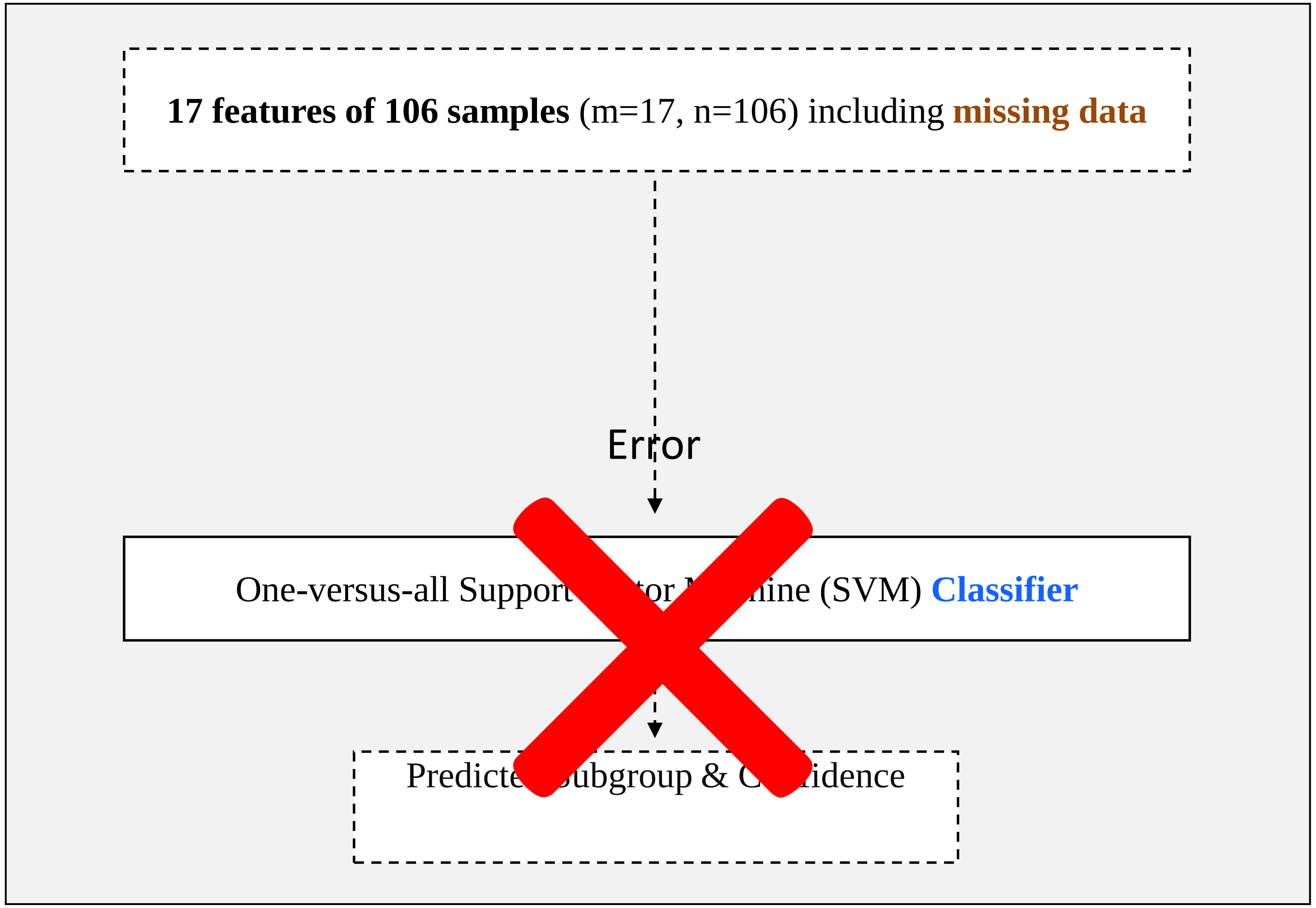
17 features of 106 samples (m=17, n=106) including **missing data**

One-versus-all Support Vector Machine (SVM) **Classifier**

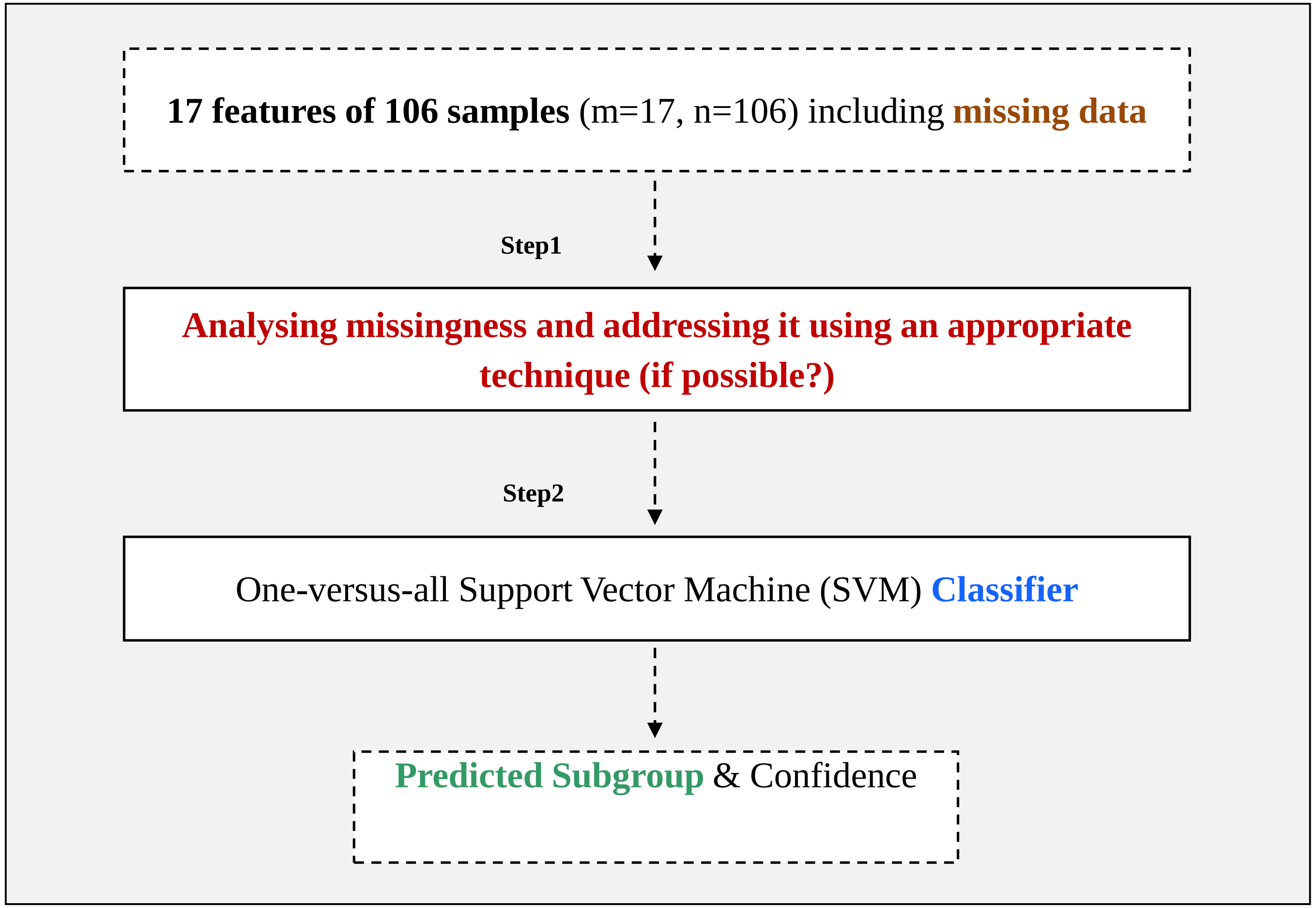
Predicted Subgroup & Confidence

A package called **e1071** in R

Example of missingness in prediction



Example of missingness in prediction



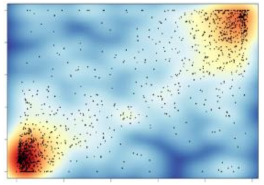
What if we couldn't address missingness using an imputation technique?

QUEEN'S UNIVERSITY BELFAST

SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Example of missingness in prediction

MS-MIMIC (certified assay)



NewGene Next Generation Diagnostics

106 samples, 17 features (dataset)

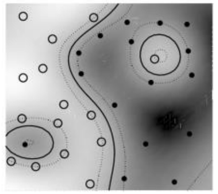
Missing data

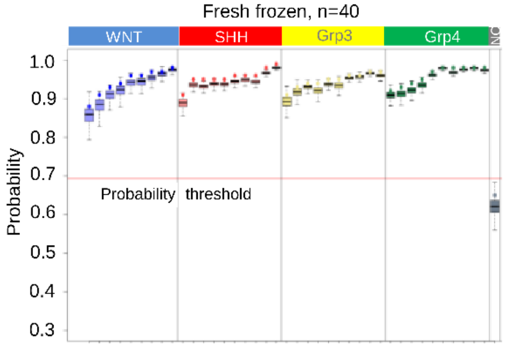
Handling missing data

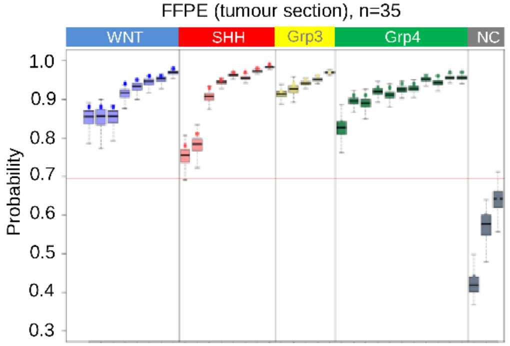
Complete data

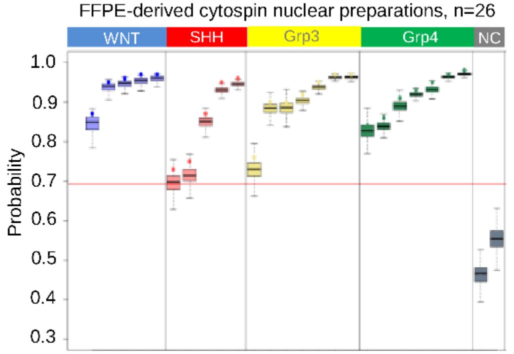
Already designed classifiers

Multiclass, non-linear SVM classifier









+ Schwalbe E.C.*, Hicks D.*, **Rafiee G.***, et al., "Minimal methylation classifier (MIMIC): A novel method for derivation and rapid diagnostic detection of disease-associated DNA methylation signatures", ***Nature Scientific Reports***, *(* denotes joint first-authorship)*, October 2017.

Categories of missingness

- Failure in:
  - Responding to a question (in surveys)
  - Equipment (sensors), recording mechanisms
  - Data entry
  - …

Missing at Random (MAR)

Missing Completely at Random (MCAR)

Missing Not at Random (MNAR)

The probability that a value is missing depends only on observed values.

$\underbrace{\hspace{10em}}$

the missingness cannot be predicted from any other variables or sets of variables

What is imputation?

- This statistical technique (algorithm) takes the incomplete dataset (i.e., including missing values) and returns a final imputed (filled in) dataset with no missing values.

The aim is to impute (fill in) the values of the missing data that resemble the underlying complete data as closely as possible[+].

[+]*Dimitris Bertsimas, Colin Pawlowski, Ying Daisy Zhuo*; 18(196):1−39, 2018, *Journal of Machine Learning Research 18 (2018) 1-39*, **"From Predictive Methods to Missing Data Imputation: An Optimization Approach"**

Current methods for imputation

Bertsimas, Pawlowski, and Zhuo

Method Name	Category	Software	Reference
Mean impute (`mean`)	Mean		Little and Rubin (1987)
Expectation-Maximization (`EM`)	EM		Dempster et al. (1977)
EM with Mixture of Gaussians and Multinomials	EM		Ghahramani and Jordan (1994)
EM with Bootstrapping	EM	`Amelia II`	Honaker et al. (2011)
K-Nearest Neighbors (`knn`)	K-NN	`impute`	Troyanskaya et al. (2001)
Sequential K-Nearest Neighbors	K-NN		Kim et al. (2004)
Iterative K-Nearest Neighbors	K-NN		Caruana (2001); Brás and Menezes (2007)
Support Vector Regression	SVR		Wang et al. (2006)
Predictive-Mean Matching (`pmm`)	LS	`MICE`	Buuren and Groothuis-Oudshoorn (2011)
Least Squares	LS		Bø et al. (2004)
Sequential Regression Multivariate Imputation	LS		Raghunathan et al. (2001)
Local-Least Squares	LS		Kim et al. (2005)
Sequential Local-Least Squares	LS		Zhang et al. (2008)
Iterative Local-Least Squares	LS		Cai et al. (2006)
Sequential Regression Trees	Tree	`MICE`	Burgette and Reiter (2010)
Sequential Random Forest	Tree	`missForest`	Stekhoven and Bühlmann (2012)
Singular Value Decomposition	SVD		Troyanskaya et al. (2001)
Bayesian Principal Component Analysis	SVD	`pcaMethods`	Oba et al. (2003); Mohamed et al. (2009)
Factor Analysis Model for Mixed Data	FA		Khan et al. (2010)

Table 1: List of Imputation Methods

Journal of Machine Learning Research 18 (2018) 1-39, **“From Predictive Methods to Missing Data Imputation: An Optimization Approach”**

Current methods for imputation

BERTSIMAS, PAWLOWSKI, AND ZHUO

Method Name	Category	Software	Reference
Mean impute (`mean`)	Mean		Little and Rubin (1987)
Expectation-Maximization (`EM`)	EM		Dempster et al. (1977)
EM with Mixture of Gaussians and Multinomials	EM		Ghahramani and Jordan (1994)
EM with Bootstrapping	EM	`Amelia II`	Honaker et al. (2011)
K-Nearest Neighbors (`knn`)	K-NN	`impute`	Troyanskaya et al. (2001)
Sequential K-Nearest Neighbors	K-NN		Kim et al. (2004)
Iterative K-Nearest Neighbors	K-NN		Caruana (2001); Brás and Menezes (2007)
Support Vector Regression	SVR		Wang et al. (2006)
Predictive-Mean Matching (`pmm`)	LS	`MICE`	Buuren and Groothuis-Oudshoorn (2011)
Least Squares	LS		Bø et al. (2004)
Sequential Regression Multivariate Imputation	LS		Raghunathan et al. (2001)
Local-Least Squares	LS		Kim et al. (2005)
Sequential Local-Least Squares	LS		Zhang et al. (2008)
Iterative Local-Least Squares	LS		Cai et al. (2006)
Sequential Regression Trees	Tree	`MICE`	Burgette and Reiter (2010)
Sequential Random Forest	Tree	`missForest`	Stekhoven and Bühlmann (2012)
Singular Value Decomposition	SVD		Troyanskaya et al. (2001)
Bayesian Principal Component Analysis	SVD	`pcaMethods`	Oba et al. (2003); Mohamed et al. (2009)
Factor Analysis Model for Mixed Data	FA		Khan et al. (2010)

Table 1: List of Imputation Methods

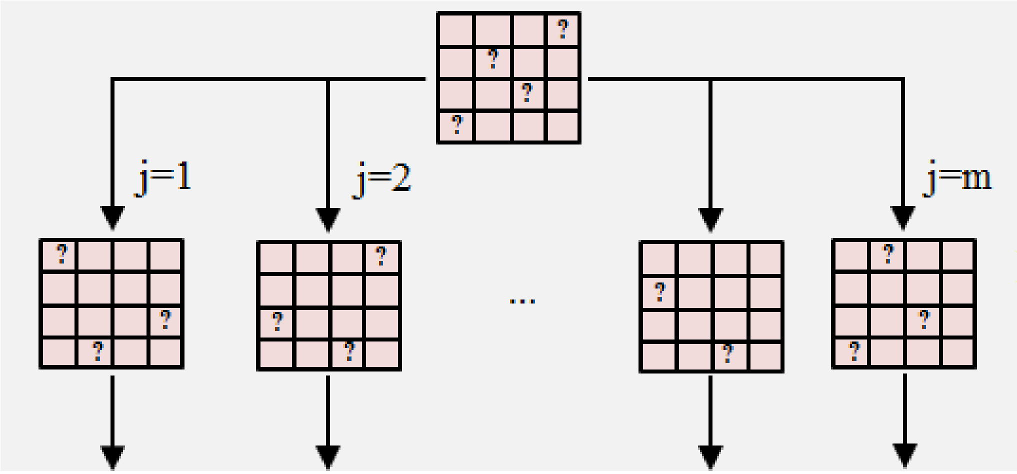
Journal of Machine Learning Research 18 (2018) 1-39, **"From Predictive Methods to Missing Data Imputation: An Optimization Approach"**

QUEEN'S UNIVERSITY BELFAST | SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

What is multiple imputation?

- This statistical technique (algorithm) takes the incomplete dataset (i.e., including missing data) and returns m imputed datasets with no missing values.

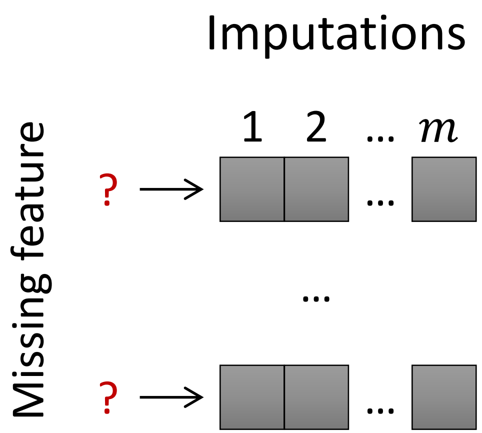
m is a user-selected parameter



Multiple imputation

- Each missing feature is imputed (filled in) with a set of $m > 1$ plausible values which reflect the uncertainty about the missing feature.

$$
\begin{array}{l c}
 & \begin{array}{ccc} 1 & \ldots & 106 \end{array} \\
\begin{array}{l} \text{Feature 1} \\ \ldots \\ \text{Feature 17} \end{array} & \begin{bmatrix} 0.9 & \cdots & ? \\ \vdots & \ddots & \vdots \\ ? & \cdots & 0.1 \end{bmatrix}
\end{array}
$$



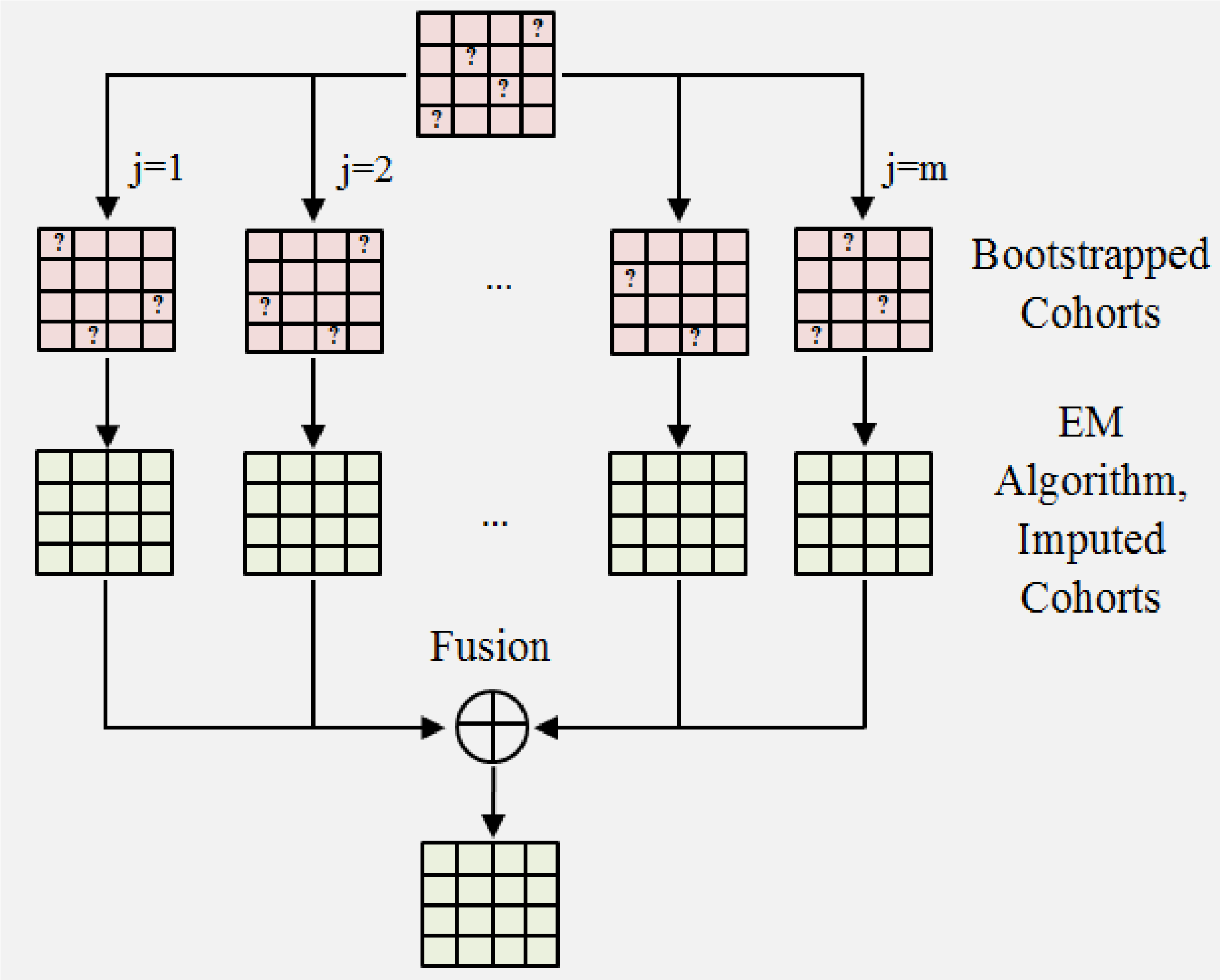
Imputation techniques and packages in R

- **M**ultivariate **I**mputation by **C**hained **E**quations (MICE)
  - https://cran.r-project.org/web/packages/mice/index.html
- **B**ootstrapped **E**xpectation-**M**aximisation (BEM)
  - https://cran.r-project.org/web/packages/Amelia/index.html
- **M**ultiple **I**mputation using an approximate Bayesian framework (MI)
  - https://cran.r-project.org/web/packages/mi/mi.pdf
- **V**isualisation and **I**mputation of **M**issing Values (VIM)
  - https://cran.r-project.org/web/packages/VIM/index.html

QUEEN'S UNIVERSITY BELFAST | SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Multiple imputation modelling techniques

- Bootstrapped Expectation-Maximisation (BEM)



Question

Which technique or imputation method would be appropriate for a specific dataset?

(How to choose an imputation technique from a list of available packages)

Which technique or imputation method?

Discussion: dataset dependency?

Number of missing per sample

	A	B	C	D	E	F	G	H	I	J	K
1		CSC3062_108_2	CSC3062_109_4	CSC3062_110_4	CSC3062_112_2	CSC3062_113_2	CSC3062_125_4	CSC3062_127_3	CSC3062_130_4	CSC3062_132_4	CSC3062_134_4
2	feature_1	0.290874776	0.89080331	0.81032173	0.094939587	0.150149242	0.89433132	0.27512451	0.827973513	0.83451715	0.825536552
3	feature_2		0.08627098	0.2441651	0.821881924	0.709218768	0.103017283	0.909203863	0.145577715	0.20381042	0.242565516
4	feature_3	0.865808069	0.92201287	0.89654937	0.956021386	0.735552896	0.9525625	0.889457035	0.928676048	0.68739547	0.947010103
5	feature_4		0.0655766			0.858727746					0.049718481
6	feature_5	0.966055005	0.05415225	0.08579509	0.997462814		0.028919888	0.059173152	0.056658601	0.3076646	0.011375909
7	feature_6		0.06252419	0.10805568	0.998506562	0.95055801	0.046704008	0.309769218	0.041905677	0.2857966	0.096072952
8	feature_7		0.32792332	0.10845017	0.914788949	0.86872521	0.236932825	0.081102726	0.420160222	0.24202751	0.234431085
9	feature_8	0.771524676	0.70493114	0.80318701	0.236309397	0.21748349	0.749603835	0.806856586	0.733636133	0.7106085	0.859536891
10	feature_9		0.98066608	0.9844879	0.998409165	0.966818311	0.990438742	0.989619716	0.991279458	0.98848102	0.97757636
11	feature_10	0.442210237	0.05116329	0.06296946	0.70187932	0.040175667	0.053461366	0.117850085	0.051542927	0.04986252	0.045808883
12	feature_11		0.06391362	0.13571643	0.453133041	0.892199087	0.058884539	0.686559491	0.112967744	0.13221349	0.189032123
13	feature_12		0.99149929	0.98880779	0.997791094	0.98400824	0.988925295	0.992761839	0.990677697	0.99289298	0.980565199
14	feature_13	0.085722787	0.05524618	0.0464798	0.105029451	0.061793271	0.036679333	0.934434637	0.069101161	0.21767877	0.057077769
15	feature_14	0.044409295	0.02614883	0.0281052	0.030050428	0.028856994	0.037977469	0.872106236	0.026116449	0.09474476	0.03148797
16	feature_15		0.01671519	0.0206424	0.004755696	0.024785343	0.016187983	0.0131722	0.011121978	0.00782123	0.019781302
17	feature_16	0.728546654	0.08666581	0.05717338	0.871429414	0.740817463	0.062748032	0.082223218	0.102088481	0.1406173	0.064767425
18	feature_17		0.00775338		0.011523751	0.009470701	0.009221186	0.006522704	0.007768349	0.0244767	0.006747257

9 Missing

QUEEN'S UNIVERSITY BELFAST

SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Number of missing per feature

7 Missing

	A	B	C	D	E	F	G	H	I	J	K
1		CSC30[illegible]_108_2	CSC3062_109_4	CSC3062_110_4	CSC3062_112_2	CSC3062_113_2	CSC3062_125_4	CSC3062_127_3	CSC3062_130_4	CSC3062_132_4	CSC3062_134_4
2	feature_1	0.[illegible]90874776	0.89080331	0.81032173	0.094939587	0.150149242	0.89433132	0.27512451	0.827973513	0.83451715	0.825536552
3	feature_2		0.08627098	0.2441651	0.821881924	0.709218768	0.103017283	0.909203863	0.145577715	0.20381042	0.242565516
4	feature_3	0.865808069	0.92201287	0.89654937	0.956021386	0.735552896	0.9525625	0.889457035	0.928676048	0.68739547	0.947010103
5	feature_4		0.0655766			0.858727746					0.049718481
6	feature_5	0.966055005	0.05415225	0.08579509	0.997462814		0.028919888	0.059173152	0.056658601	0.3076646	0.011375909
7	feature_6		0.06252419	0.10805568	0.998506562	0.95055801	0.046704008	0.309769218	0.041905677	0.2857966	0.096072952
8	feature_7		0.32792332	0.10845017	0.914788949	0.86872521	0.236932825	0.081102726	0.420160222	0.24202751	0.234431085
9	feature_8	0.771524676	0.70493114	0.80318701	0.236309397	0.21748349	0.749603835	0.806856586	0.733636133	0.7106085	0.859536891
10	feature_9		0.98066608	0.9844879	0.998409165	0.966818311	0.990438742	0.989619716	0.991279458	0.98848102	0.97757636
11	feature_10	0.442210237	0.05116329	0.06296946	0.70187932	0.040175667	0.053461366	0.117850085	0.051542927	0.04986252	0.045808883
12	feature_11		0.06391362	0.13571643	0.453133041	0.892199087	0.058884539	0.686559491	0.112967744	0.13221349	0.189032123
13	feature_12		0.99149929	0.98880779	0.997791094	0.98400824	0.988925295	0.992761839	0.990677697	0.99289298	0.980565199
14	feature_13	0.085722787	0.05524618	0.0464798	0.105029451	0.061793271	0.036679333	0.934434637	0.069101161	0.21767877	0.057077769
15	feature_14	0.044409295	0.02614883	0.0281052	0.030050428	0.028856994	0.037977469	0.872106236	0.026116449	0.09474476	0.03148797
16	feature_15		0.01671519	0.0206424	0.004755696	0.024785343	0.016187983	0.0131722	0.011121978	0.00782123	0.019781302
17	feature_16	0.728546654	0.08666581	0.05717338	0.871429414	0.740817463	0.062748032	0.082223218	0.102088481	0.1406173	0.064767425
18	feature_17		0.00775338		0.011523751	0.009470701	0.009221186	0.006522704	0.007768349	0.0244767	0.006747257

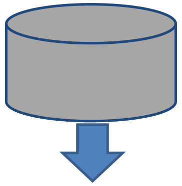
Fraction of missing in total

Dependent on your dataset statistics

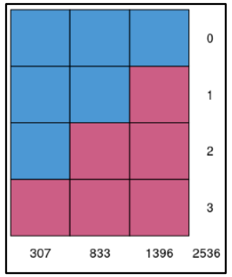
- Fraction of missing in total
    - The number of missing per sample
    - The number of missing per feature

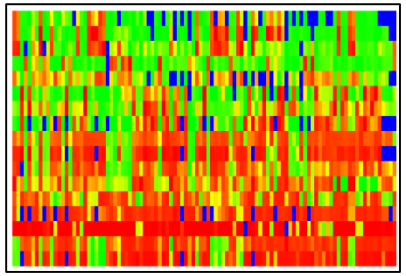
Impact of imputation on final result

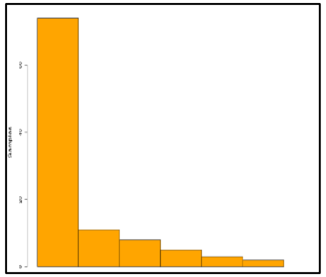
Incomplete dataset



Evaluation of missing





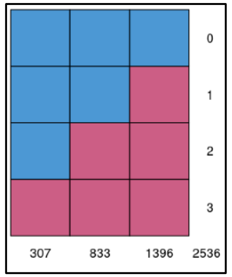


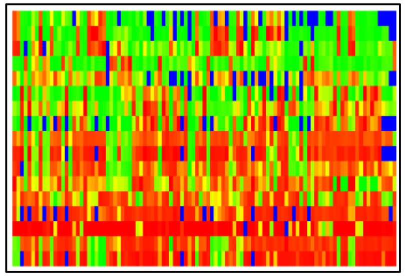
Impact of imputation on final result

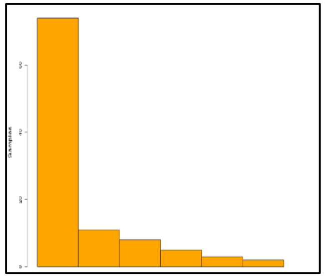
Incomplete dataset

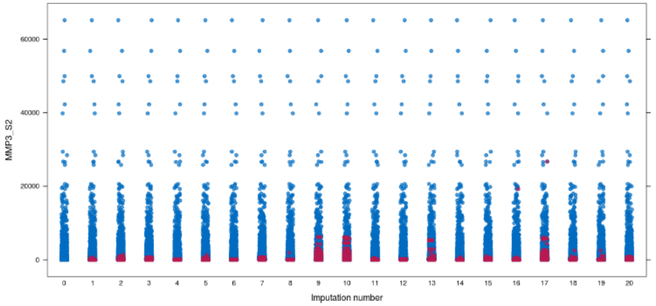
Evaluation of missing

Imputation method

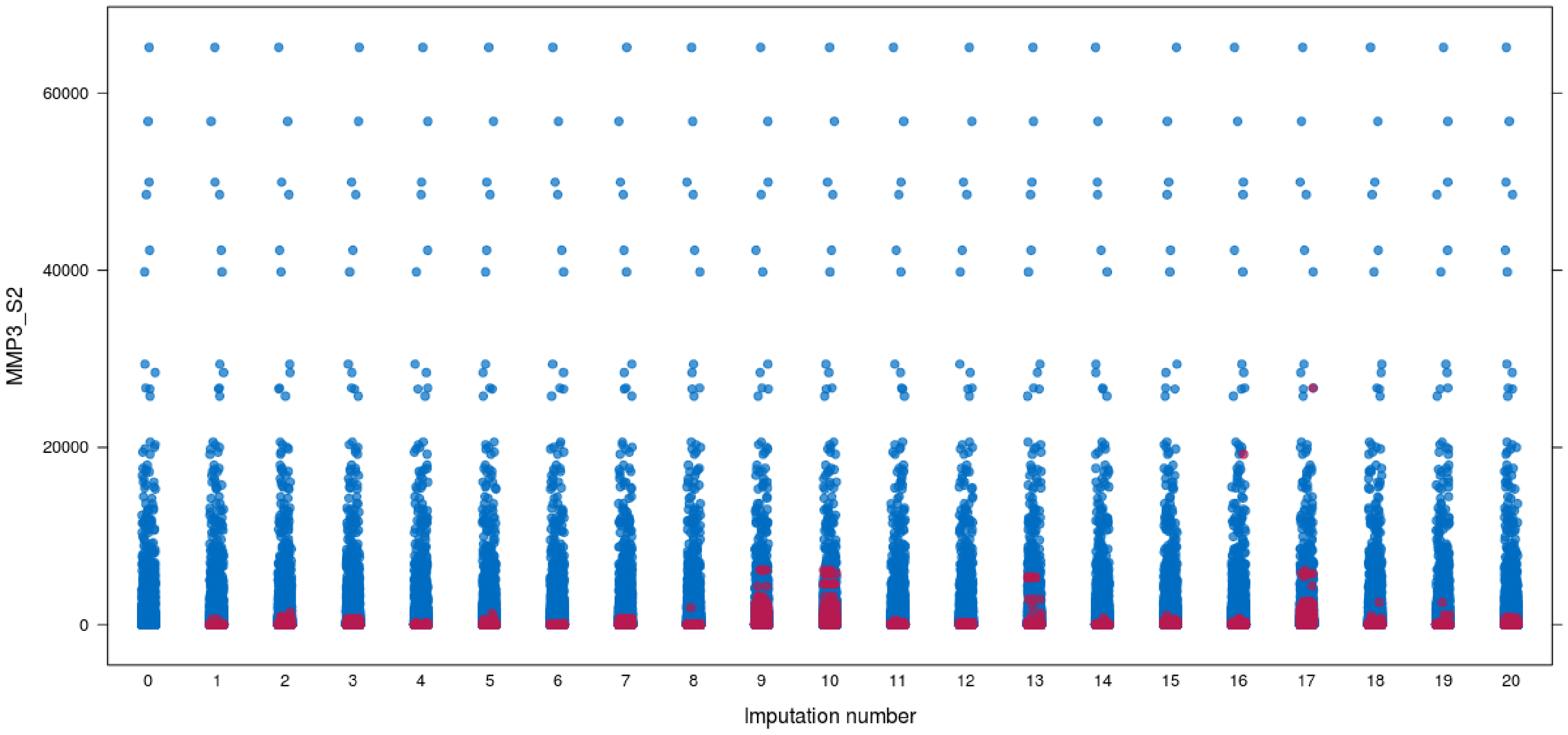








After applying a multiple imputation method



Observed

The figures show that the distribution of the imputed and the observed values are similar (observed data in blue, imputed in red).

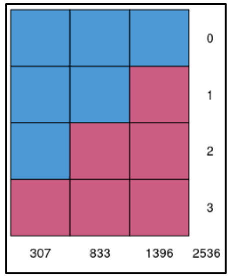
Impact of imputation on final result

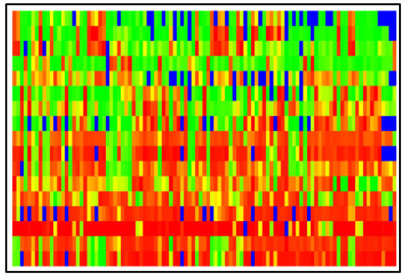
Incomplete dataset

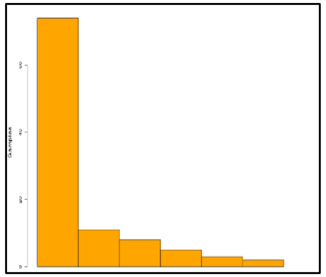
Evaluation of missing

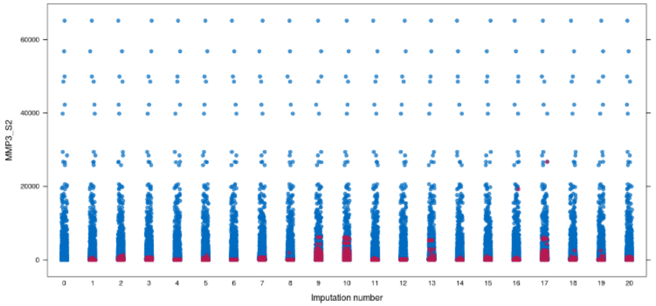
Imputation method

Complete dataset

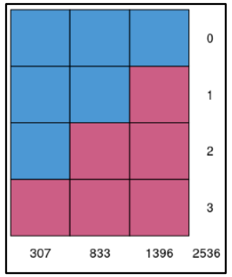


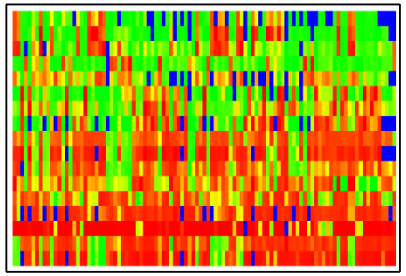


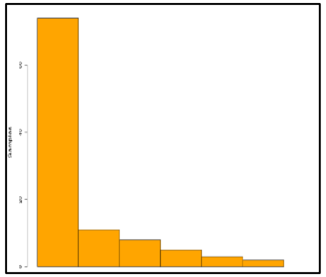




Impact of imputation on final result

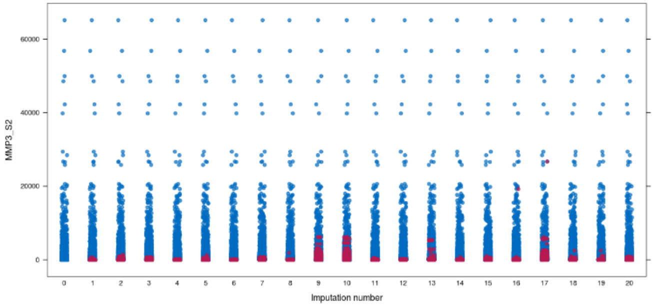






Incomplete dataset

Evaluation of missing



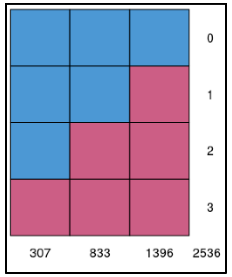
Imputation method

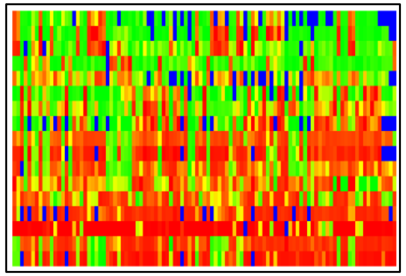
Complete dataset

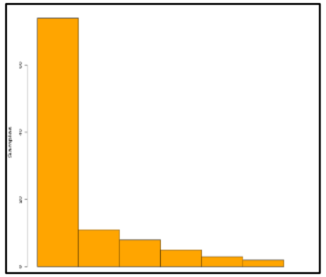
Main analysis

e.g., prediction

Impact of imputation on final result

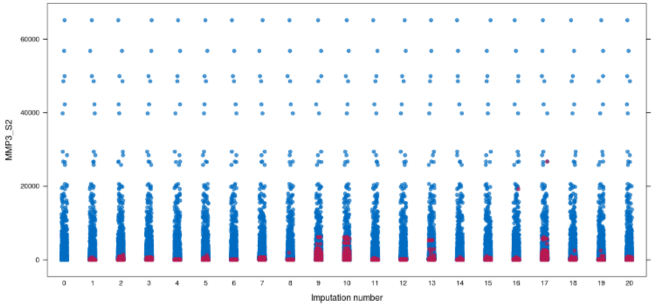






Incomplete dataset

Evaluation of missing



Imputation method

Complete dataset

Main analysis

e.g., prediction

Final result

Impact of imputation on final result

Incomplete dataset

Evaluation of missing

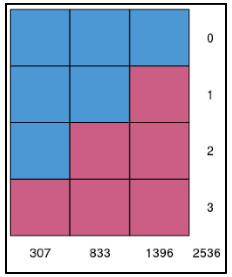
Imputation method

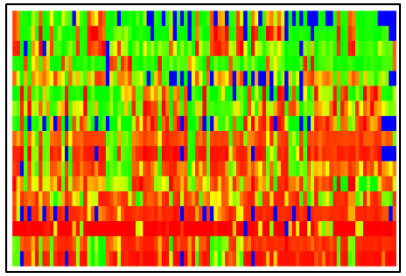
Complete dataset

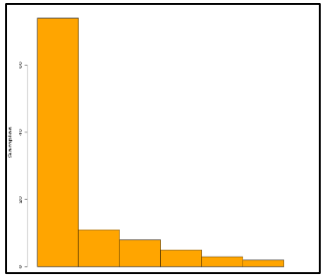
Main analysis

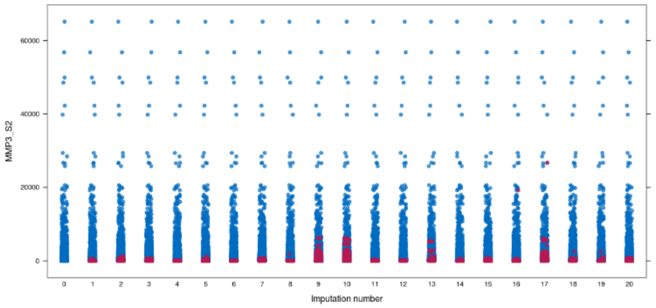
Final result

Inaccurate results



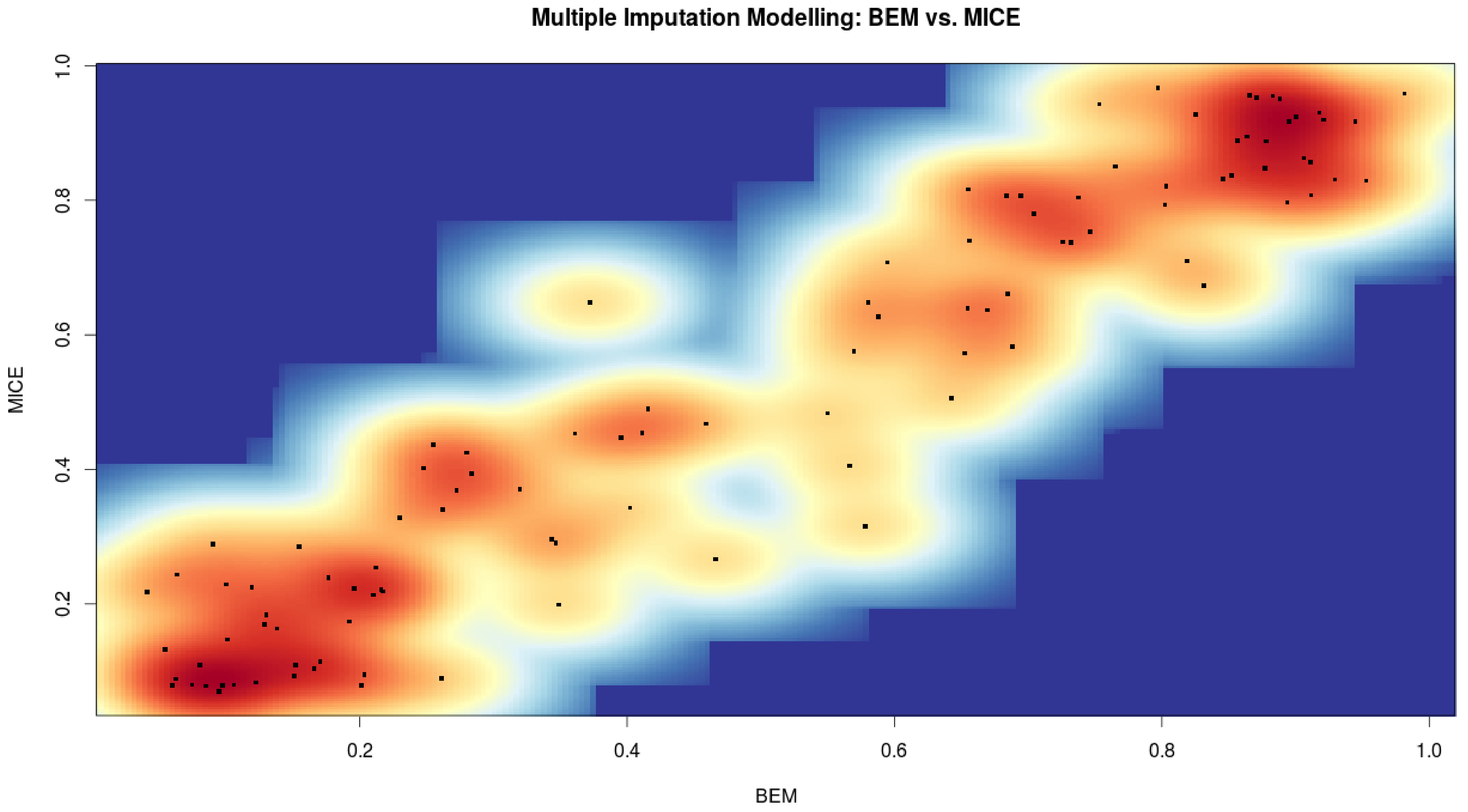






e.g., prediction

Visualising BEM vs. MICE using scatter plot



QUEEN'S UNIVERSITY BELFAST
SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Impact of a method (BEM or MICE)

Different multiple imputation methods may affect the final results (e.g., classification results)

BEM

		Reference subgroup			
		Grp1	**Grp2**	**Grp 3**	**Grp 4**
Predicted Subgroup	**Grp1**	**22**	0	0	0
	Grp2	0	**23**	0	0
	Grp3	0	0	**23**	0
	Grp4	0	0	0	**28**
	NC+	**2**	**4**	**1**	**0**
	Total	**24**	**27**	**24**	**28**

MICE

		Reference subgroup			
		Grp1	**Grp2**	**Grp3**	**Grp4**
Predicted Subgroup	**Grp1**	**22**	0	0	0
	Grp2	0	**22**	0	0
	Grp3	0	**1**	**23**	0
	Grp4	0	0	0	**28**
	NC	**2**	**5**	**0**	**0**
	Total	**24**	**28**	**23**	**28**

Summarising the performance of a classification algorithm using a "confusion matrix". A matrix (table) shows the discrepancy between predicted and reference subgroup.

+NC: Non-classifiable

Efficiency of Multiple Imputation

- Efficiency of an estimate based on m imputation is approximately:

$$(1 + \frac{\gamma}{m})^{-1}$$

Where γ is the fraction of missing information for the quality being estimated.

1) Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys . J. Wiley & Sons, New York.
2) Schafer, Joseph L. and Maren K. Olsen. 1998. \Multiple imputation for multivariate missing-data problems: A data analyst's perspective." Multivariate Behavioral Research 33(4):545-571.

QUEEN'S UNIVERSITY BELFAST

SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Efficiency of m imputations for 17 features

Feature #	missing fraction	Efficiency of m imputation per feature	Average of efficiency (12 features)
1	0.165048544	0.991815118	**0.995782732**
2	0.048543689	0.997578693	m=20
3	0.038834951	0.998062016	
4	0	-	
5	0.077669903	0.996131528	
6	0.009708738	0.999514799	
7	0.155339806	0.992292871	
8	0.009708738	0.999514799	
9	0.077669903	0.996131528	
10	0.038834951	0.998062016	
11	0	-	
12	0.242718447	0.988009592	
13	0	-	
14	0.019417476	0.999030068	
15	0	-	
16	0	-	
17	0.13592233	0.993249759	

Installation and Updates from R

To install the Amelia package on any platform, simply type the following at the R command prompt,

```
> install.packages("Amelia")
> update.packages()
```

Let’s look at an R script and doing multiple imputation

Any Questions?