

Machine learning for ecosystem assessment in the Amazon, using environmental DNA data

In Collaboration with NatureMetrics

by

Adamos Spanashis

Supervised by

Dr Ben Calderhead and Dr Oliver Ratmann

Contents

1. Motivation

Why we applied machine learning

2. Data

Our data set; features and class labels

3. Train-Validation-Test

How we evaluated the classifiers

4. Results

Results of the project

Motivation

Why monitor the environment

- ▶ Rising human populations.
- ▶ Increase in unmanaged waste, especially running waters.
- ▶ Assessment of environmental health developed as a response.

Why monitor the environment

- ▶ Rising human populations.
- ▶ Increase in unmanaged waste, especially running waters.
- ▶ Assessment of environmental health developed as a response.

Why monitor the environment

- ▶ Rising human populations.
- ▶ Increase in unmanaged waste, especially running waters.
- ▶ Assessment of environmental health developed as a response.

Traditional environmental monitoring

- Selecting indicator species associated with specific kind of pollution.
- Morpho-taxonomic identification of organisms.
- Filling up biotic indices to quantify pollution.

Traditional environmental monitoring

- Selecting indicator species associated with specific kind of pollution.
- Morpho-taxonomic identification of organisms.
- Filling up biotic indices to quantify pollution.

Traditional environmental monitoring

- Selecting indicator species associated with specific kind of pollution.
- Morpho-taxonomic identification of organisms.
- Filling up biotic indices to quantify pollution.

Traditional monitoring disadvantages

- Taxonomic identification requires experts, is time consuming and expensive.
- Selection of indicator species is arbitrary.
- Taxonomic resolution is low.

Traditional monitoring disadvantages

- Taxonomic identification requires experts, is time consuming and expensive.
- Selection of indicator species is arbitrary.
- Taxonomic resolution is low.

Traditional monitoring disadvantages

- Taxonomic identification requires experts, is time consuming and expensive.
- Selection of indicator species is arbitrary.
- Taxonomic resolution is low.

Traditional monitoring disadvantages



Chironomus Plumosus



Chironomus Zealandicus

Genomics; A new hope?

- Sanger Sequencing 1977
- Barcoding and taxonomic reference libraries
- Next Generation Sequencing
- Costs go down by orders of magnitude

Genomics; A new hope?

- Sanger Sequencing 1977
- Barcoding and taxonomic reference libraries
- Next Generation Sequencing
- Costs go down by orders of magnitude

Genomics; A new hope?

- Sanger Sequencing 1977
- Barcoding and taxonomic reference libraries
- Next Generation Sequencing
- Costs go down by orders of magnitude

Genomics; A new hope?

- Sanger Sequencing 1977
- Barcoding and taxonomic reference libraries
- Next Generation Sequencing
- Costs go down by orders of magnitude

Genomics; A new hope?



Sampling from
Environment



Selecting
appropriate
primers



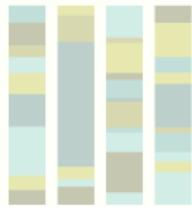
2

3

tttgagtatacaact
ttcgagcatacgact
aacgtccaaaggagt
ttggagcatacgact
aagggtccaaaggagt
ttcgagcatacgact
atcgltccaatggagt
aagggtccaaacgagt
aacgtccaaaggagt
tttgagtatacaact

Sequencing DNA
metabarcoding
libraries

4



OTU clustering
and taxonomic
identification

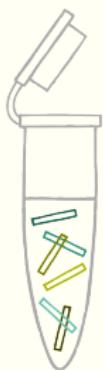
Genomics; A new hope?

1



Sampling from
Environment

2



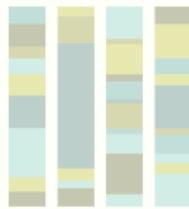
Selecting
appropriate
primers

3

tttgagtatacaact
ttcgagcatacgact
aacgtccaaaggagt
ttggagcatacgact
aagggtccaaaggagt
ttcgagcatacgact
atcgltccaatggagt
aagggtccaaacgagt
aacgtccaaaggagt
tttgagtatacaact

Sequencing DNA
metabarcoding
libraries

4



OTU clustering
and taxonomic
identification

Genomics; A new hope?



Sampling from
Environment



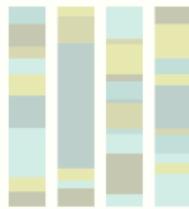
Selecting
appropriate
primers



2

tttgagtatacaact
ttcgagcatacact
aacgtccaaaggagt
ttggagcatacact
aagggtccaaaggagt
ttcgagcatacact
atcgtccaatggagt
aagggtccaaacgagt
aacgtccaaaggagt
tttgagtatacaact

3



4

Sequencing DNA
metabarcoding
libraries

OTU clustering
and taxonomic
identification

Genomics; A new hope?



Sampling from Environment

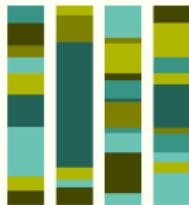


Selecting appropriate primers



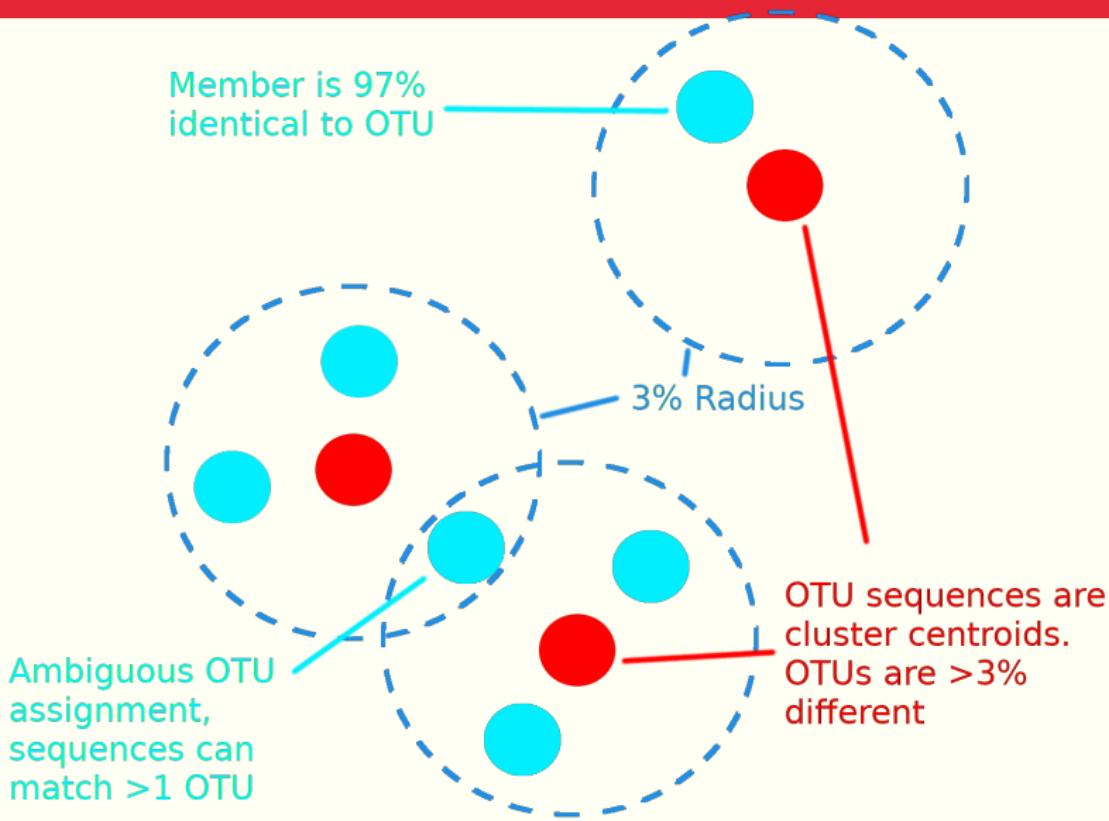
tttgagtatacaact
ttcgagcatacgact
aacgtccaaaggagt
ttggagcatacgact
aagggtccaaaggagt
ttcgagcatacgact
atcgtccaatggagt
aagggtccaaacgagt
aacgtccaaaggagt
tttgagtatacaact

Sequencing DNA metabarcoding libraries



OTU clustering and taxonomic identification

OTUs

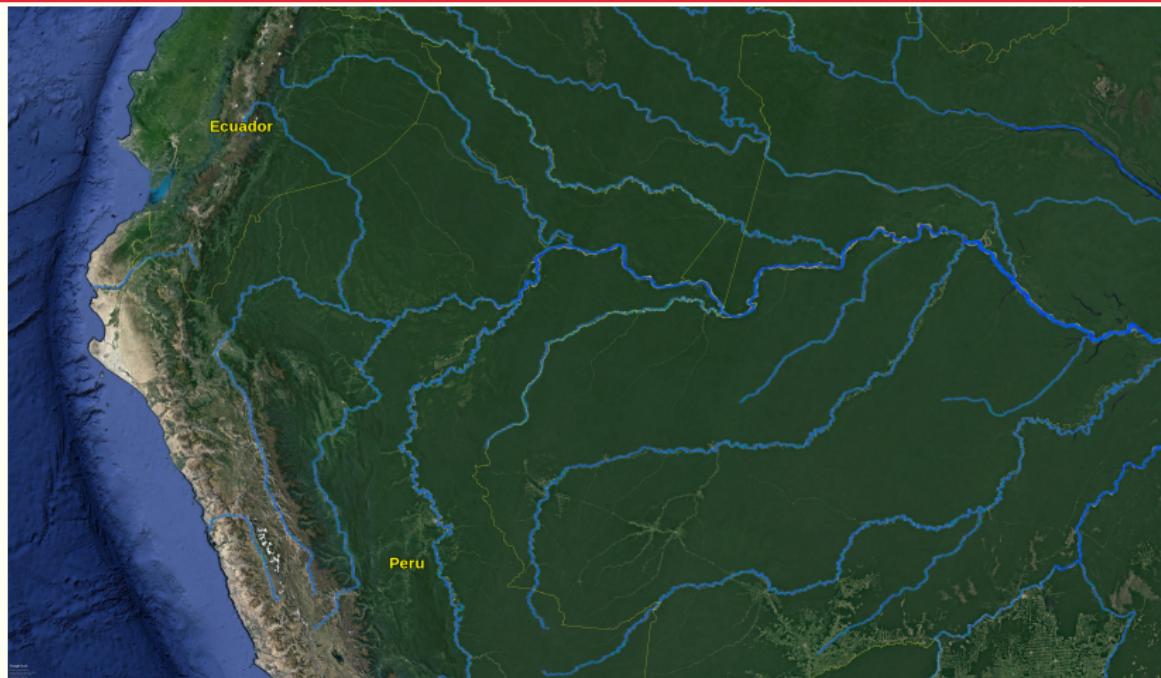


Data

Sampling

- Collected 164 samples from Northern Peruvian rivers
- Sequenced using Next generation sequencing
- 675 OTUs matched with taxonomies from reference libraries
- Discarded the rest
- Recorded colour of water and location

Sampling



Sampling



Sampling

- Collected 164 samples from Northern Peruvian rivers
- Sequenced using Next generation sequencing
- 675 OTUs matched with taxonomies from reference libraries
- Discarded the rest
- Recorded colour of water and location

Sampling

- Collected 164 samples from Northern Peruvian rivers
- Sequenced using Next generation sequencing
- 675 OTUs matched with taxonomies from reference libraries
- Discarded the rest
- Recorded colour of water and location

Sampling

- Collected 164 samples from Northern Peruvian rivers
- Sequenced using Next generation sequencing
- 675 OTUs matched with taxonomies from reference libraries
- Discarded the rest
- Recorded colour of water and location

Sampling

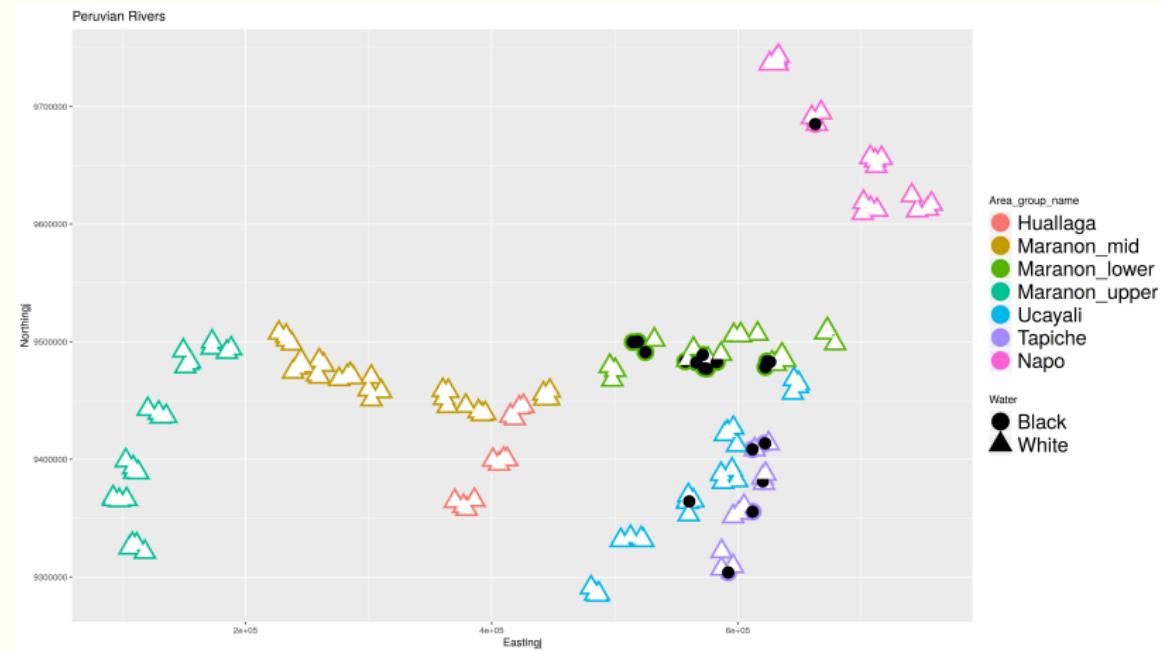
- Collected 164 samples from Northern Peruvian rivers
- Sequenced using Next generation sequencing
- 675 OTUs matched with taxonomies from reference libraries
- Discarded the rest
- Recorded colour of water and location

Meta Data



White vs Black

Meta Data



Features

- Sparse data matrix
- Dimensionality reduction
 - Principle Components Analysis
 - Principle Coordinates Analysis¹
 - Non-metric Multidimensional Scaling²
 - High spearman correlation features removed
- Cumulative sum scaling normalisation and log transformation³

¹Warren S. Torgerson, "Multidimensional scaling: I. Theory and method", In: *Psychometrika* 17.4 (Dec. 1, 1952).

²J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis", In: *Psychometrika* 25.1 (Mar. 1958), pp. 115–129.

³Joseph N. Paulson et al. "Robust methods for differential abundance analysis in marker gene surveys". In: *Nature methods* 10.12 (), pp. 1200–1202.

Features

- Sparse data matrix
- Dimensionality reduction
 - Principle Components Analysis
 - Principle Coordinates Analysis¹
 - Non-metric Multidimensional Scaling²
 - High spearman correlation features removed
- Cumulative sum scaling normalisation and log transformation³

¹Warren S. Torgerson. "Multidimensional scaling: I. Theory and method". In: *Psychometrika* 17.4 (Dec. 1, 1952).

²J. B. Kruskal. "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". In: *Psychometrika* 29.1 (Mar. 1964), pp. 1–27.

³Joseph N. Paulson et al. "Robust methods for differential abundance analysis in marker gene surveys". In: *Nature methods* 10.12 (), pp. 1200–1202.

Features

- Sparse data matrix
- Dimensionality reduction
 - Principle Components Analysis
 - Principle Coordinates Analysis¹
 - Non-metric Multidimensional Scaling²
 - High spearman correlation features removed
- Cumulative sum scaling normalisation and log transformation³

¹Warren S. Torgerson. "Multidimensional scaling: I. Theory and method". In: *Psychometrika* 17.4 (Dec. 1, 1952).

²J. B. Kruskal. "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". In: *Psychometrika* 29.1 (Mar. 1964), pp. 1–27.

³Joseph N. Paulson et al. "Robust methods for differential abundance analysis in marker gene surveys". In: *Nature methods* 10.12 (), pp. 1200–1202.

Features

- ▶ Sparse data matrix
- ▶ Dimensionality reduction
 - ▶ Principle Components Analysis
 - ▶ Principle Coordinates Analysis¹
 - ▶ Non-metric Multidimensional Scaling²
 - ▶ High spearman correlation features removed
- ▶ Cumulative sum scaling normalisation and log transformation³

¹Warren S. Torgerson. "Multidimensional scaling: I. Theory and method". In: *Psychometrika* 17.4 (Dec. 1, 1952).

²J. B. Kruskal. "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". In: *Psychometrika* 29.1 (Mar. 1964), pp. 1–27.

³Joseph N. Paulson et al. "Robust methods for differential abundance analysis in marker gene surveys". In: *Nature methods* 10.12 (), pp. 1200–1202.

Features

- ▶ Sparse data matrix
- ▶ Dimensionality reduction
 - ▶ Principle Components Analysis
 - ▶ Principle Coordinates Analysis¹
 - ▶ Non-metric Multidimensional Scaling²
 - ▶ High spearman correlation features removed
- ▶ Cumulative sum scaling normalisation and log transformation³

¹Warren S. Torgerson. "Multidimensional scaling: I. Theory and method". In: *Psychometrika* 17.4 (Dec. 1, 1952).

²J. B. Kruskal. "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". In: *Psychometrika* 29.1 (Mar. 1964), pp. 1–27.

³Joseph N. Paulson et al. "Robust methods for differential abundance analysis in marker gene surveys". In: *Nature methods* 10.12 (), pp. 1200–1202.

Features

- Sparse data matrix
- Dimensionality reduction
 - Principle Components Analysis
 - Principle Coordinates Analysis¹
 - Non-metric Multidimensional Scaling²
 - High spearman correlation features removed
- Cumulative sum scaling normalisation and log transformation³

¹Warren S. Torgerson. "Multidimensional scaling: I. Theory and method". In: *Psychometrika* 17.4 (Dec. 1, 1952).

²J. B. Kruskal. "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". In: *Psychometrika* 29.1 (Mar. 1964), pp. 1–27.

³Joseph N. Paulson et al. "Robust methods for differential abundance analysis in marker gene surveys". In: *Nature methods* 10.12 (), pp. 1200–1202.

Features

- Sparse data matrix
- Dimensionality reduction
 - Principle Components Analysis
 - Principle Coordinates Analysis¹
 - Non-metric Multidimensional Scaling²
 - High spearman correlation features removed
- Cumulative sum scaling normalisation and log transformation³

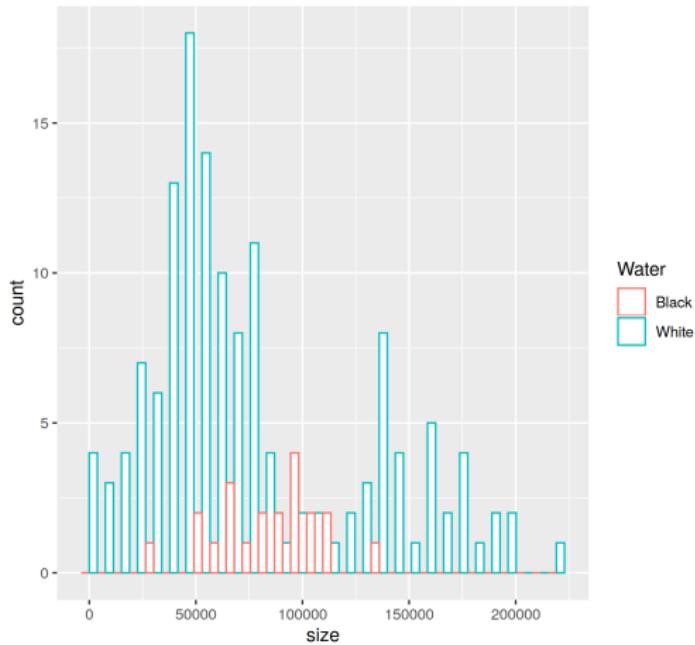
¹Warren S. Torgerson. "Multidimensional scaling: I. Theory and method". In: *Psychometrika* 17.4 (Dec. 1, 1952).

²J. B. Kruskal. "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". In: *Psychometrika* 29.1 (Mar. 1964), pp. 1–27.

³Joseph N. Paulson et al. "Robust methods for differential abundance analysis in marker gene surveys". In: *Nature methods* 10.12 (), pp. 1200–1202.

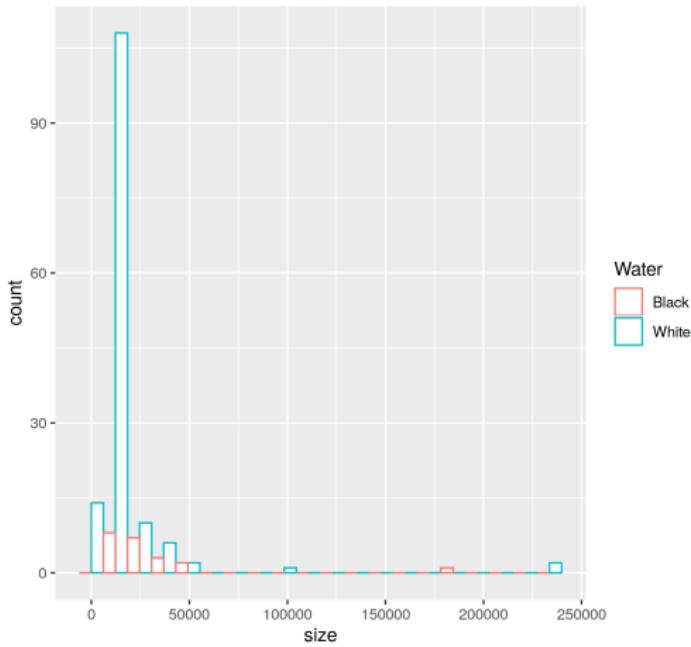
Features

Histogram of total sample counts



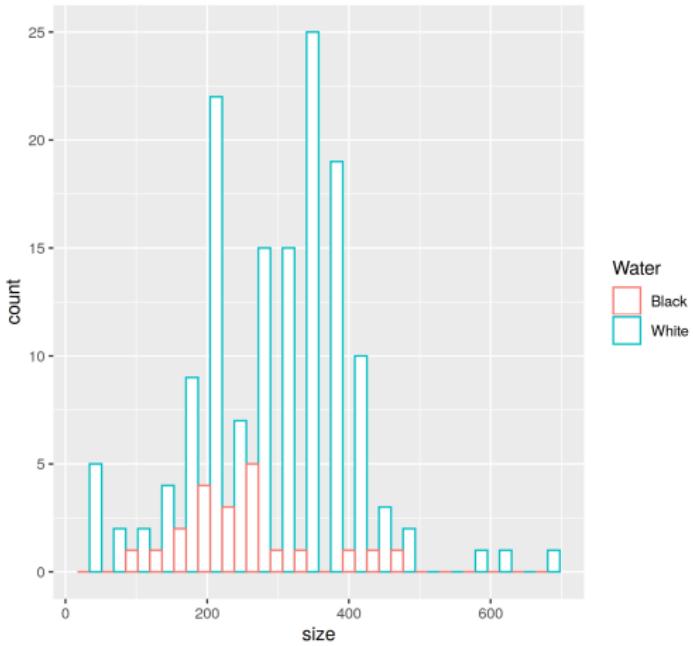
Features

Histogram of total sample counts CSS



Features

Histogram of total sample counts CSS Log



Train-Validation-Test

Location and prediction

- Location of train and test sets will affect performance of classifiers.
- Splitting data simulates different learning conditions .
- Maximally similar
- Maximally dissimilar
- Random with equal distributions of black water in all sets

Location and prediction

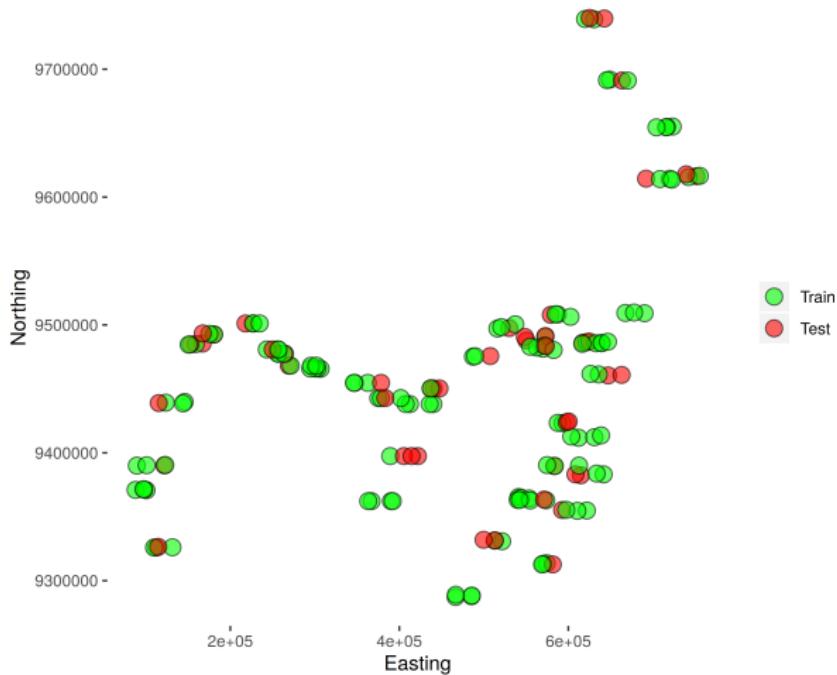
- Location of train and test sets will affect performance of classifiers.
- Splitting data simulates different learning conditions .
- Maximally similar
- Maximally dissimilar
- Random with equal distributions of black water in all sets

Location and prediction

- Location of train and test sets will affect performance of classifiers.
- Splitting data simulates different learning conditions .
- Maximally similar
- Maximally dissimilar
- Random with equal distributions of black water in all sets

Location and prediction

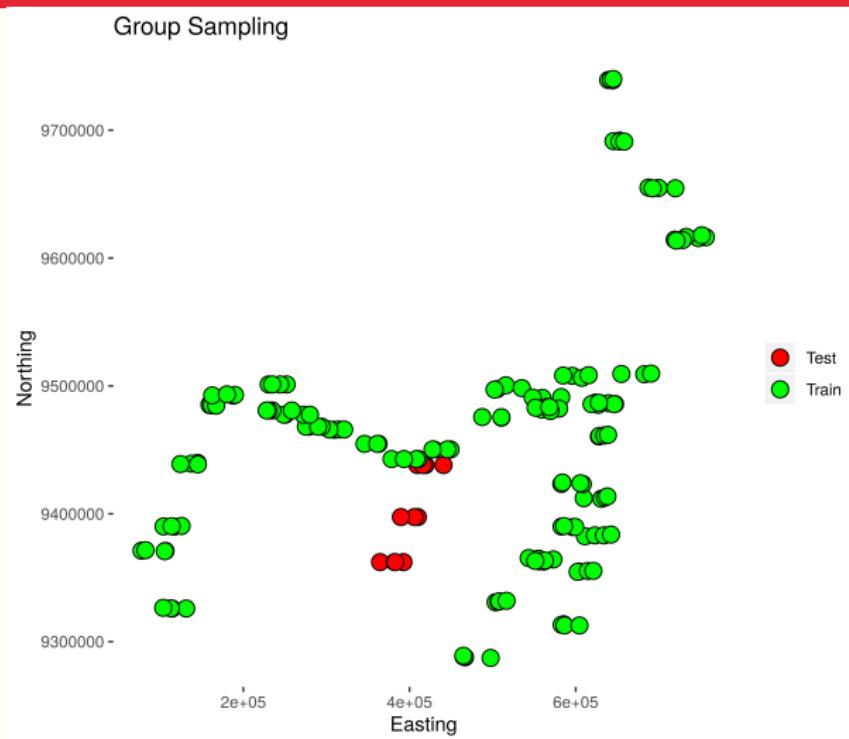
Stratified Sampling



Location and prediction

- Location of train and test sets will affect performance of classifiers.
- Splitting data simulates different learning conditions .
- Maximally similar
- Maximally dissimilar
- Random with equal distributions of black water in all sets

Location and prediction



Location and prediction

- Location of train and test sets will affect performance of classifiers.
- Splitting data simulates different learning conditions .
- Maximally similar
- Maximally dissimilar
- Random with equal distributions of black water in all sets

Model evaluation

- ➊ Split data into train and test sets.
- ➋ Cross-validate model on validation sets created from train sets using the same principle as in train-test split; pick best hyperparameters based on accuracy.
- ➌ Test on the remaining samples in the test set.
- ➍ Repeat until all samples are found in the test set.

Model evaluation

- ➊ Split data into train and test sets.
- ➋ Cross-validate model on validation sets created from train sets using the same principle as in train-test split; pick best hyperparameters based on accuracy.
- ➌ Test on the remaining samples in the test set.
- ➍ Repeat until all samples are found in the test set.

Model evaluation

- ➊ Split data into train and test sets.
- ➋ Cross-validate model on validation sets created from train sets using the same principle as in train-test split; pick best hyperparameters based on accuracy.
- ➌ Test on the remaining samples in the test set.
- ➍ Repeat until all samples are found in the test set.

Model evaluation

- ➊ Split data into train and test sets.
- ➋ Cross-validate model on validation sets created from train sets using the same principle as in train-test split; pick best hyperparameters based on accuracy.
- ➌ Test on the remaining samples in the test set.
- ➍ Repeat until all samples are found in the test set.

Classifiers

- ▶ Random Forest
- ▶ Logistic regression with L_1 and L_2 penalty
- ▶ Bayesian Logistic Regression with Laplacian and Gaussian priors

Classifiers

- ▶ Random Forest
- ▶ Logistic regression with L_1 and L_2 penalty
- ▶ Bayesian Logistic Regression with Laplacian and Gaussian priors

Classifiers

- ▶ Random Forest
- ▶ Logistic regression with L_1 and L_2 penalty
- ▶ Bayesian Logistic Regression with Laplacian and Gaussian priors

Results

Evaluating performance

- ▢ Can not use F score or ROC based on black water samples
- ▢ Some validation sets do not contain any black waters so most measures are undefined. Accuracy picks better models than F score based on white water.

Results presented in confusion matrix and accuracy:

Black samples Predicted correctly	Black samples Predicted falsely
White samples Predicted falsely	White samples Predicted correctly

(1)

Evaluating performance

- Can not use F score or ROC based on black water samples
- Some validation sets do not contain any black waters so most measures are undefined. Accuracy picks better models than F score based on white water.

Results presented in confusion matrix and accuracy:

Black samples Predicted correctly	Black samples Predicted falsely
White samples Predicted falsely	White samples Predicted correctly (1)

Evaluating performance

- Can not use F score or ROC based on black water samples
- Some validation sets do not contain any black waters so most measures are undefined. Accuracy picks better models than F score based on white water.

Results presented in confusion matrix and accuracy:

Black samples Predicted correctly	Black samples Predicted falsely
White samples Predicted falsely	White samples Predicted correctly

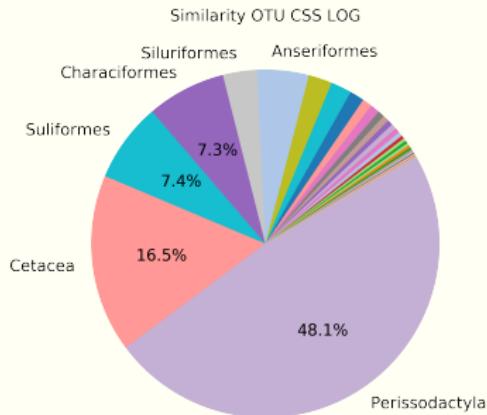
(1)

Best of Maximally similar

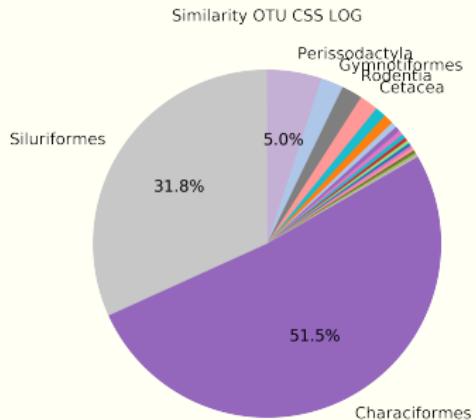
Features used	Confusion Matrix		Accuracy
	Predicted Black	Predicted White	
LR L1 OTU	19 1	2 142	98.17%
LR L1 OTU LOW	19 0	2 143	98.78%
LR L1 OTU CSS	18 0	3 143	98.17%
LR L1 OTU CSS LOG	19 1	2 142	98.17%
LR L2 OTU CSS LOG	20 1	1 142	98.78%
RF OTU LOW	18 3	3 140	96.34%
RF OTU CSS	18 3	3 140	96.34%
RF OTU CSS LOG	19 3	2 140	96.95%

Most important taxonomic orders

CSS LOG.bb



CSS LOG.bb



Aggregating importance by
averaging over Order

Aggregating importance by
summing over Order

Most important taxonomic orders



South American Tapir¹



Amazon river dolphin²

¹ Sharp Charles. *Tapir*. In: *Wikipedia*. Page Version ID: 913557340. Sept. 1, 2019.
URL: <https://en.wikipedia.org/w/index.php?title=Tapir&oldid=913557340> (visited on 09/13/2019)

² *5 amazing facts about the Amazon pink river dolphin*. Aqua Expeditions. Apr. 13, 2019. URL:
<https://www.aquaexpeditions.com/blog/amazon/facts-amazon-pink-river-dolphin/> (visited on 09/13/2019)

Best of Maximally Dissimilar

Features used	Confusion Matrix		Accuracy
	Predicted Black	Predicted White	
L1 PCoA CSS 99%	8	13	89.02%
	5	138	
L1 PCoA 90%	7	14	88.41%
	5	138	
L2 OTU CSS LOG	12	9	93.90%
	1	142	
RF OTU CSS LOG	7	14	90.24%
	2	141	
RF OTU CSS	7	14	90.24%
	2	141	

Maximally Similar LR

Features used	Confusion Matrix		Accuracy
	Predicted Black	Predicted White	
OTU	19	2	
	1	142	98.17%
OTU LOW	19	2	
	0	143	98.78%
OTU CSS	18	3	
	0	143	98.17%
OTU Min CSS	18	3	
	0	136	98.09%
OTU CSS LOG	19	2	
	1	142	98.17%
PCoA Bray-Curtis	16	5	
	3	140	95.12%
PCoA Bray-Curtis CSS	16	5	
	1	142	96.34%

Maximally Similar RF

Features used	Confusion Matrix		Accuracy
	Predicted Black	Predicted White	
OTU	17	4	95.73%
	3	140	
OTU LOW	18	3	96.34%
	3	140	
OTU CSS	18	3	96.34%
	3	140	
OTU Min CSS	18	3	96.18%
	3	133	
OTU CSS LOG	19	2	96.95%
	3	140	
PCoA Bray-Curtis	6	15	88.41%
	4	139	
PCoA Bray-Curtis CSS	4	17	89.63%
	0	143	

Maximally Dissimilar LR

Features used	Confusion Matrix		Accuracy
	Predicted Black	Predicted White	
OTU	5 18	16 125	79.27%
OTU LOW	5 17	16 126	79.88%
OTU CSS	7 14	14 129	82.93%
OTU Min CSS	3 14	18 122	79.61%
OTU CSS LOG	5 4	16 139	87.80%
PCoA Bray-Curtis	7 8	14 135	86.59%
PCoA Bray-Curtis CSS	9 8	12 135	87.80%

Maximally Dissimilar RF

Features used	Confusion Matrix		Accuracy
	Predicted Black	Predicted White	
OTU	4	17	
	10	133	83.50%
OTU LOW	2	19	
	10	133	82.32%
OTU CSS	7	14	
	2	141	90.24%
OTU Min CSS	7	14	
	1	135	90.45%
OTU CSS LOG	7	14	
	2	141	90.24%
PCoA Bray-Curtis	0	21	
	1	142	86.59%
PCoA Bray-Curtis CSS	0	21	
	5	138	84.15%