Large Scale Probabilistic and Multi-output Benthic Habitat Mapping

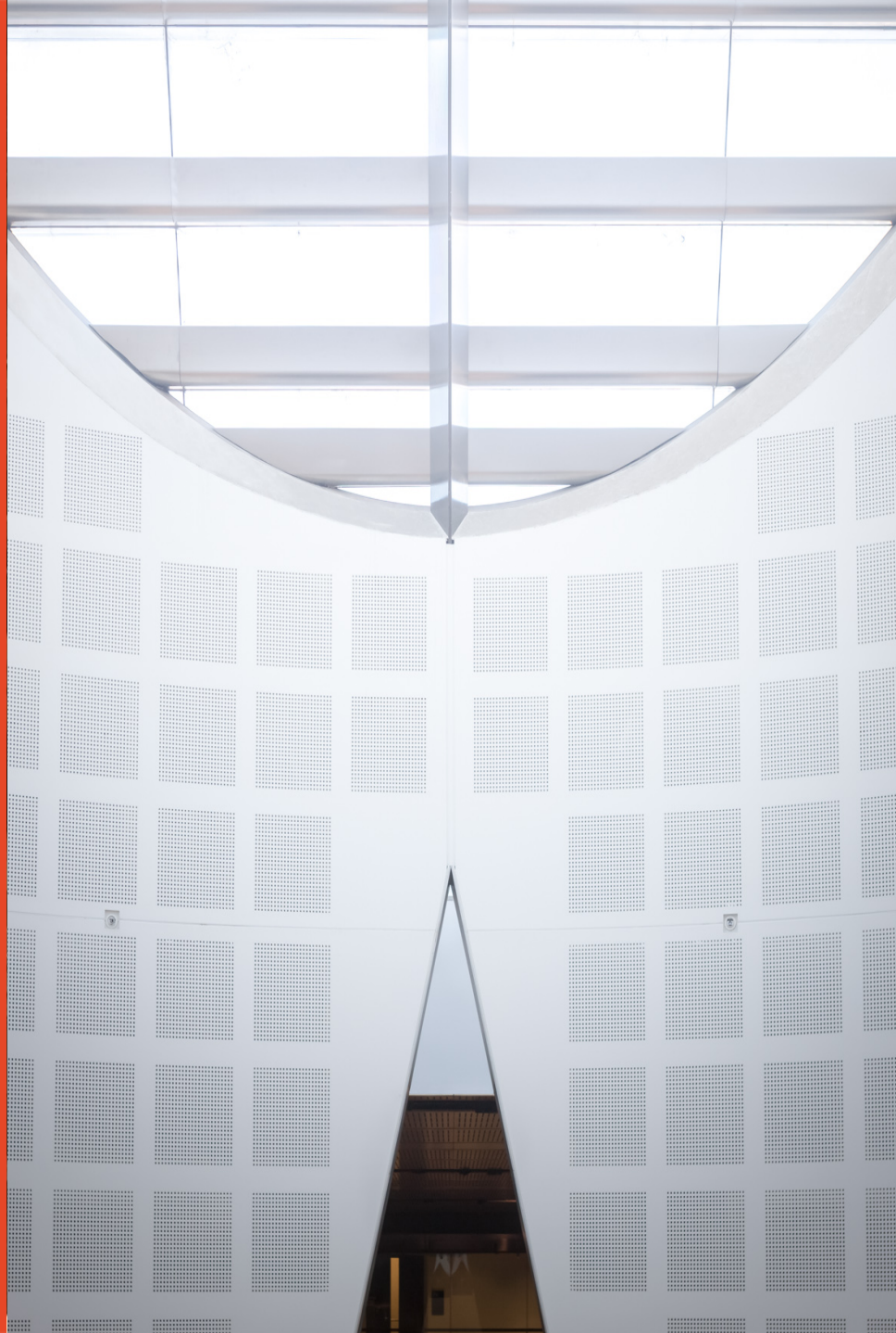
Honours Thesis

Presented by

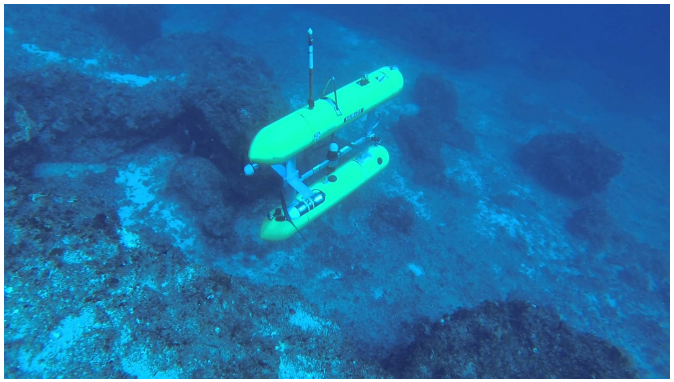
Justin Ting, School of Enginnering and IT

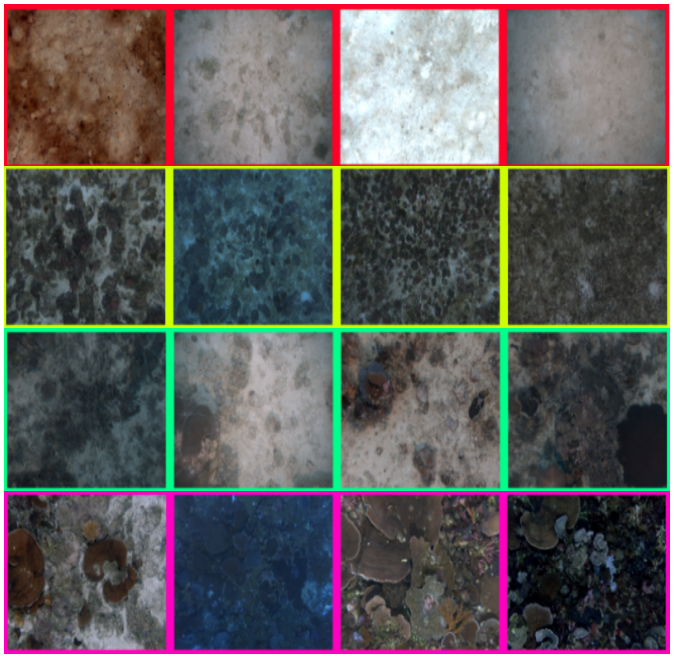
Supervisor: Simon O'Callaghan

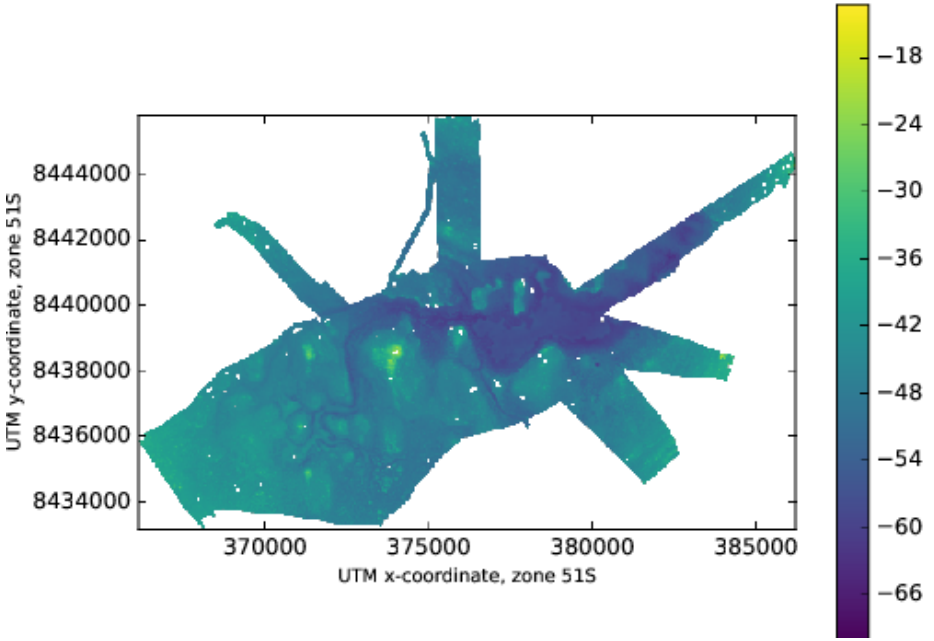
THE UNIVERSITY OF SYDNEY

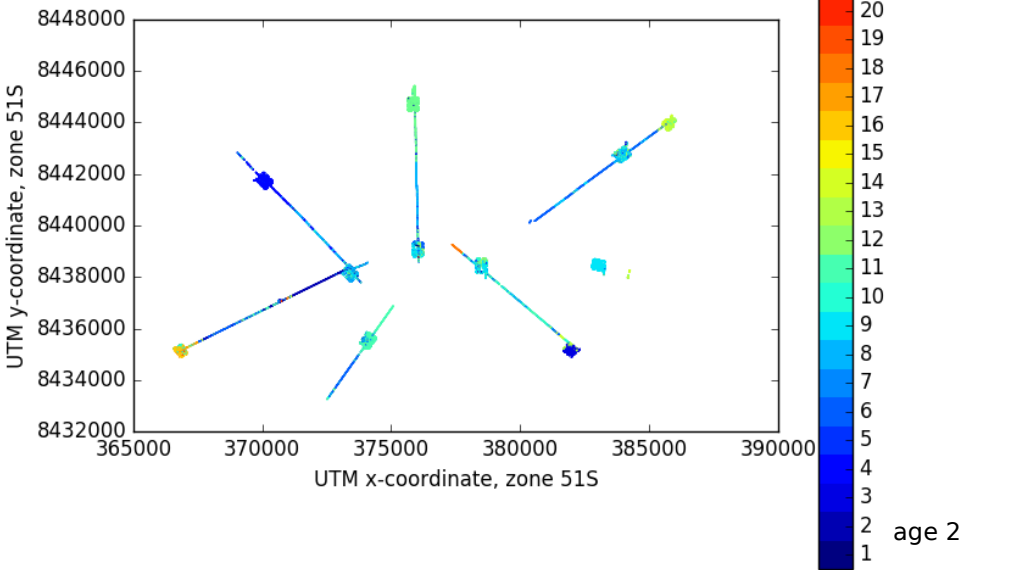


Motivation

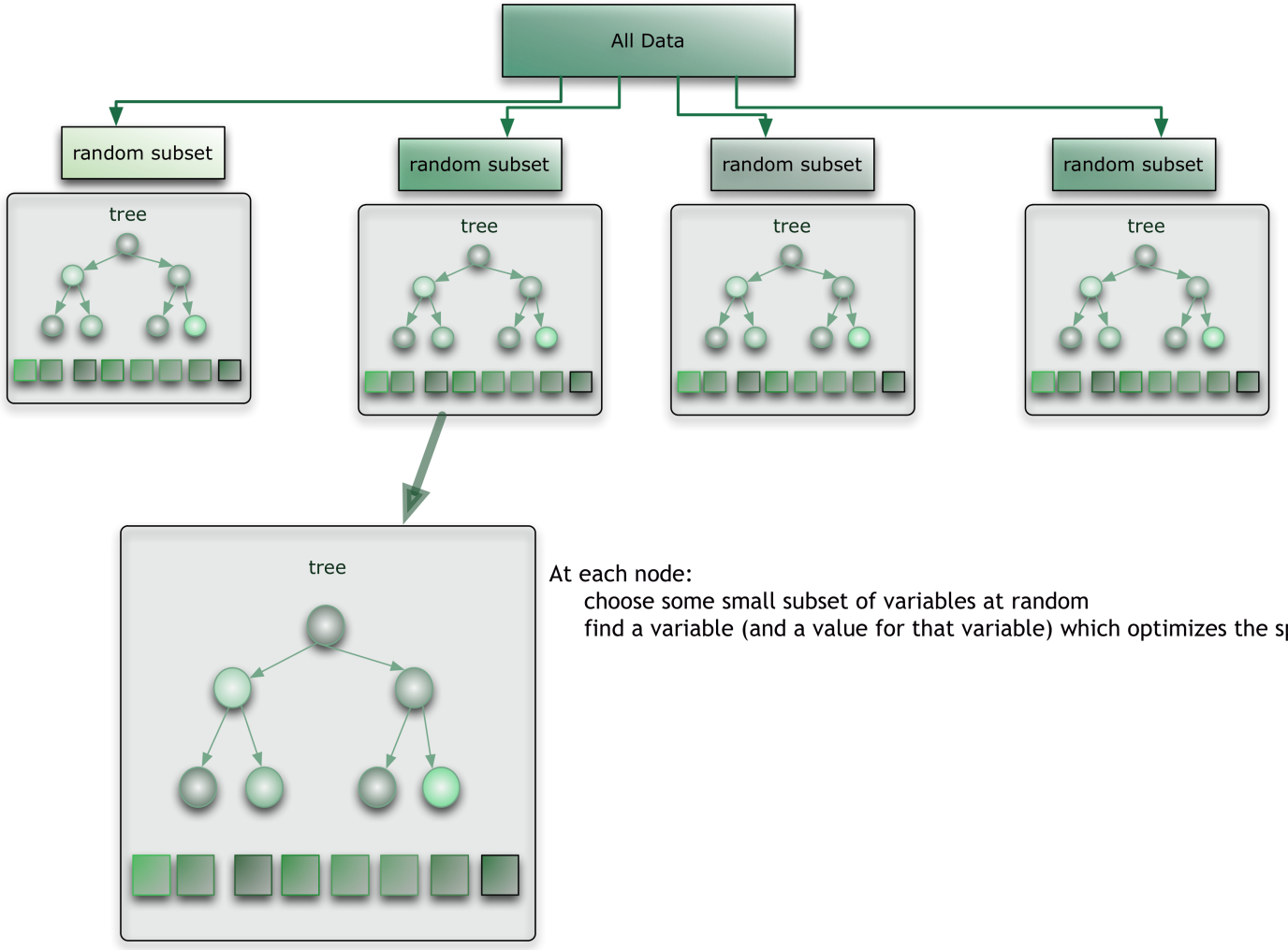





8444000
8442000
8440000
8438000
8436000
8434000
UTM y-coordinate, zone 51S
370000
375000
380000
385000
UTM x-coordinate, zone 51S
−18
−24
−30
−36
−42
−48
−54
−60
−66

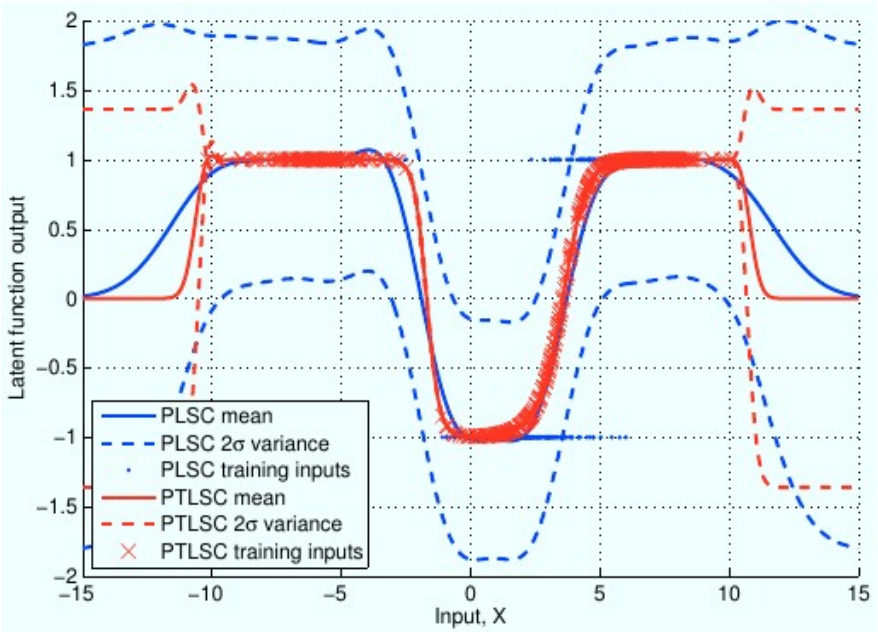

8448000
8446000
8444000
8442000
8440000
8438000
8436000
8434000
8432000
UTM y-coordinate, zone 51S
365000
370000
375000
380000
385000
390000
UTM x-coordinate, zone 51S
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1

Related Literature – Deterministic Machine Learning Methods



At each node:
choose some small subset of variables at random
find a variable (and a value for that variable) which optimizes the split

Related Literature – Probabilistic Machine Learning Methods

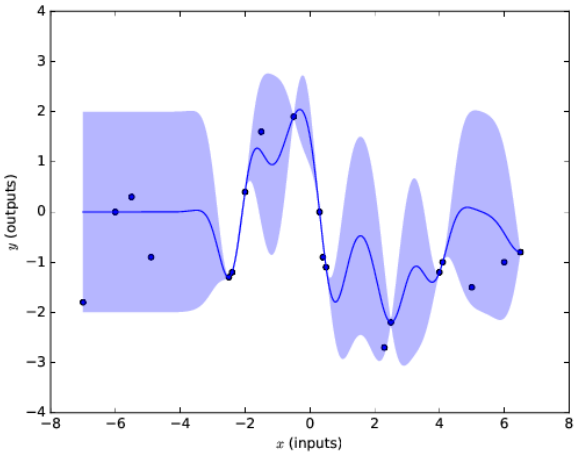


$$\mathbf{K}_{\mathbf{x}_p,\mathbf{x}_q} = \sigma_f^2 e^{\left(-\frac{1}{2}(\mathbf{x}_p-\mathbf{x}_q)^T L^{-1}(\mathbf{x}_p-\mathbf{x}_q)\right)} + \sigma_n^2 \mathbf{I}$$

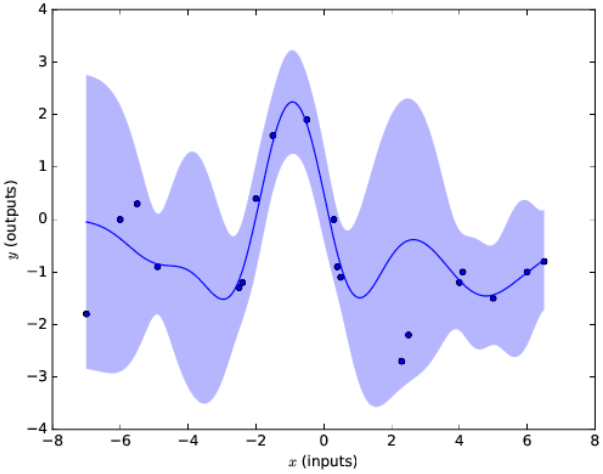
$$\mu_i = \mathbf{y}_i - \frac{\left[\mathbf{K}^{-1}\mathbf{y}\right]_{-i}}{\left[\mathbf{K}^{-1}\right]_{-ii}} \qquad \text{and} \qquad \sigma_i^2 = \frac{1}{\left[\mathbf{K}^{-1}\right]_{-ii}}$$

Approach – Gaussian Process Approximation Methods

PoGPE



GPoGPE



$$O(n_1{}^3)+O(n_2{}^3)+\ldots+O(n_{k-1}{}^3)+O(n_k{}^3) \ll O(N^3)$$

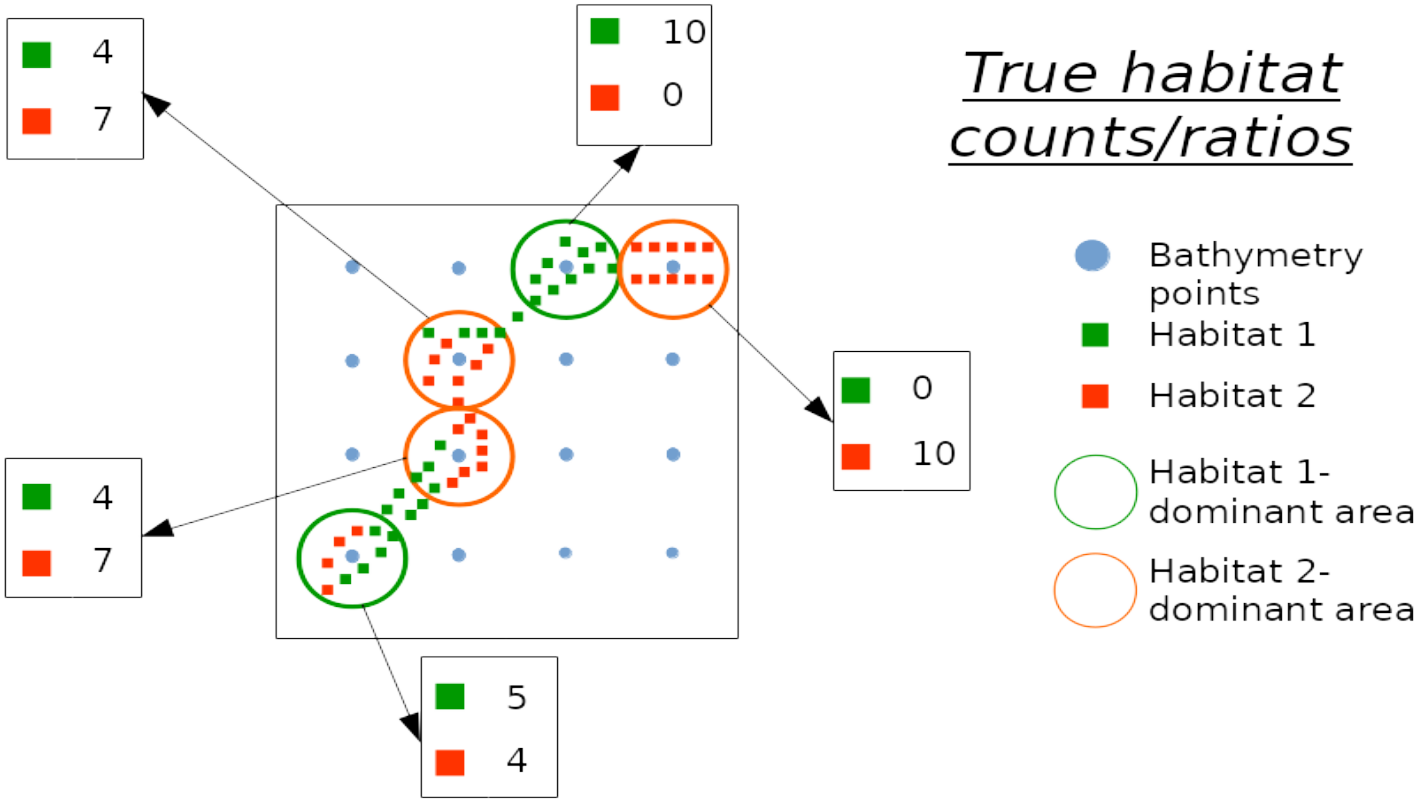
$$\text{where } n_1 + n_2 + \ldots + n_{k-1} + n_k = N$$

$$\text{if } n_k = n_k = \ldots = n_{k-1} = n_k, \text{ and } k = 100, N = 20000$$

$$100 \times O(200^3) \ll O(20000^3) \rightarrow 8 \times 10^8 \ll 8 \times 10^{12}$$

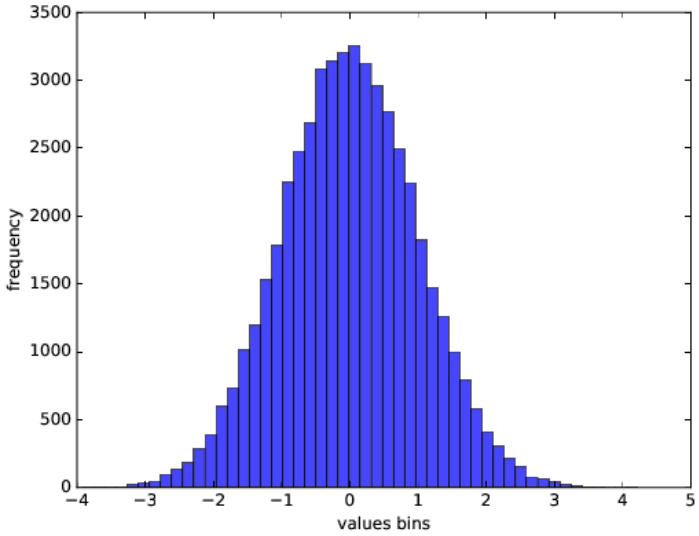
Approach – Intro to Multi-output Methods

- However, the data collection process does not often align exactly with this

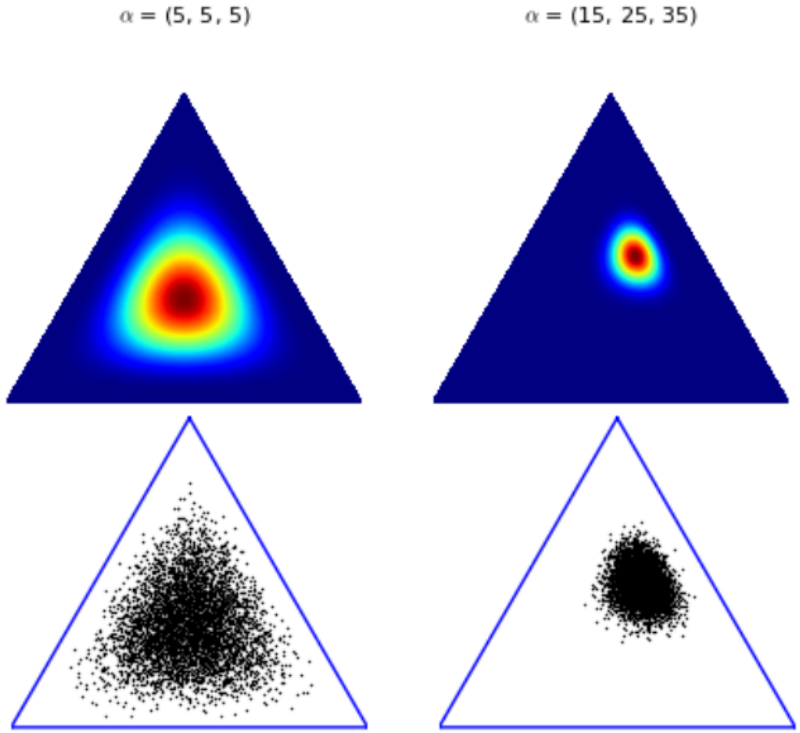


Approach – Dirichlet Multinomial Regression

Multinomial Distribution



Dirichlet Distribution



<todo – equation/s?>

Results – GP vs DM vs Deterministic Methods

		SVM	Random Forest	Logistic Regression	K-Nearest Neighbour	Gaussian Process Classification	Generalised Product of Gaussian Process Experts	Dirichlet Multinomial
Aggregated (4) labels	Accuracy	0.75	0.79	0.77	0.78	**0.83**	0.75	0.77
	F-score	0.21	0.47	0.34	0.47	**0.53**	0.33	0.38
Full (24) labels	Accuracy	0.29	0.35	0.31	0.33	**0.39**	0.21	0.28
	F-score	0.10	0.22	0.13	0.26	**0.32**	0.19	0.11

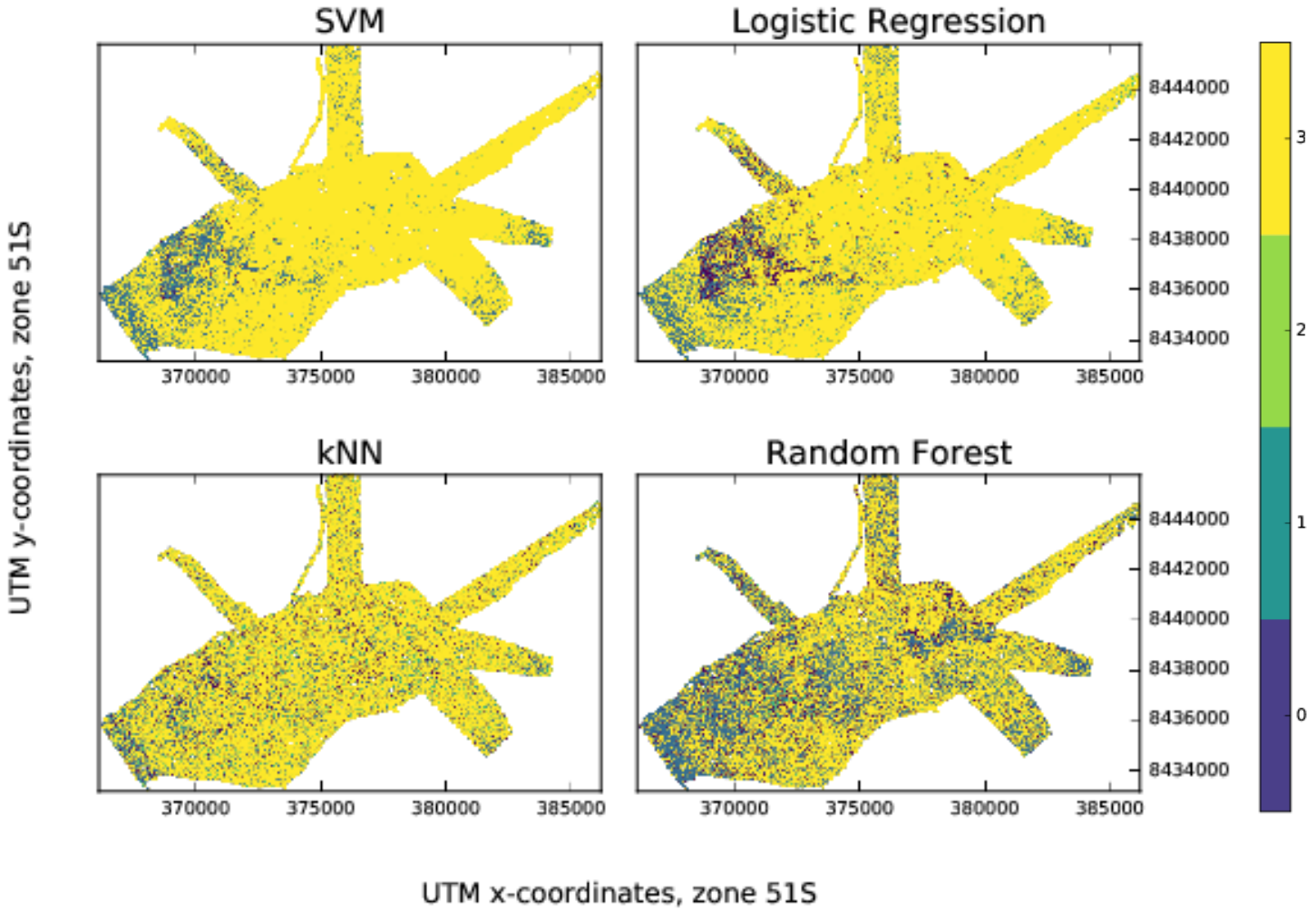
Results

Time of GP vs approx GP vs DM

Time taken to perform model fitting and predictions on 4-habitats, with 4700 training points, and 500,000 test points

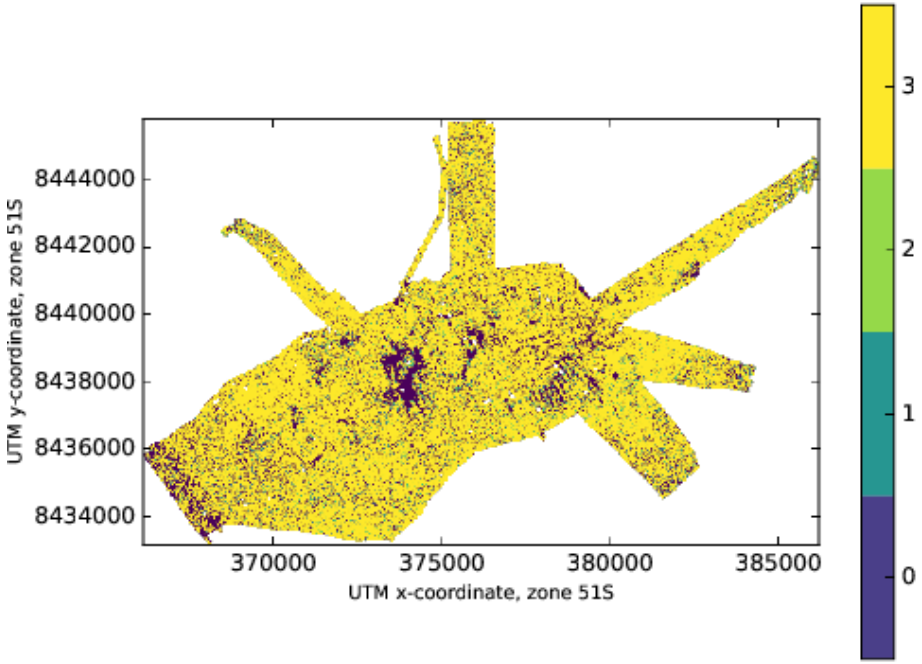
Heading 1	Gaussian Process Classification	Generalised Product of Gaussian Process Experts	Dirichlet Multinomial Regression
Model fitting	47:41	00:23	00:01
Predictions	20:41	03:35	00:05

Discussion – Deterministic Maps

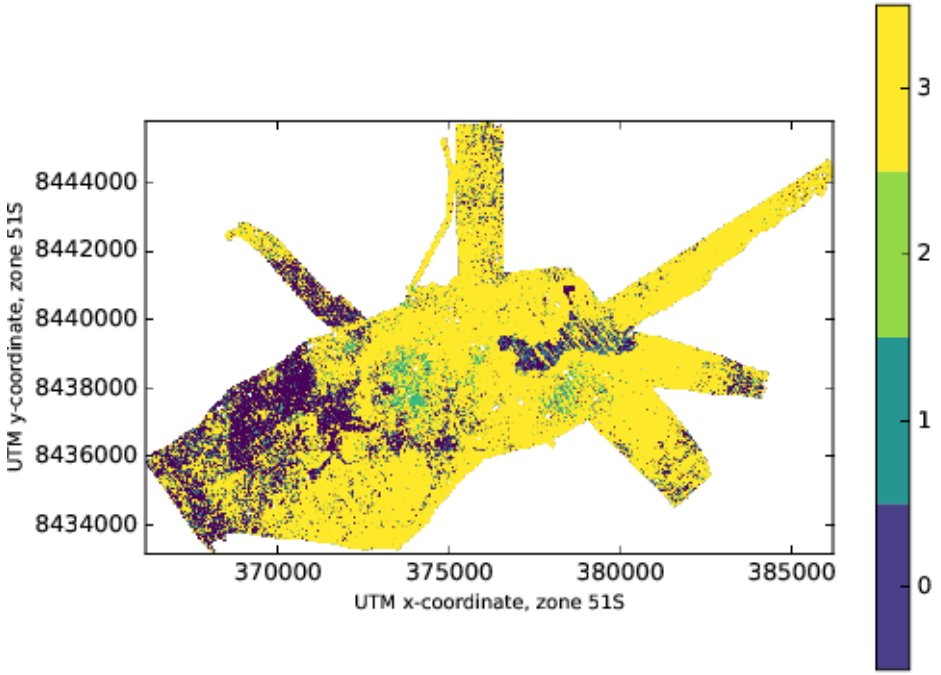


Discussion – Gaussian Process and Approximations

GP

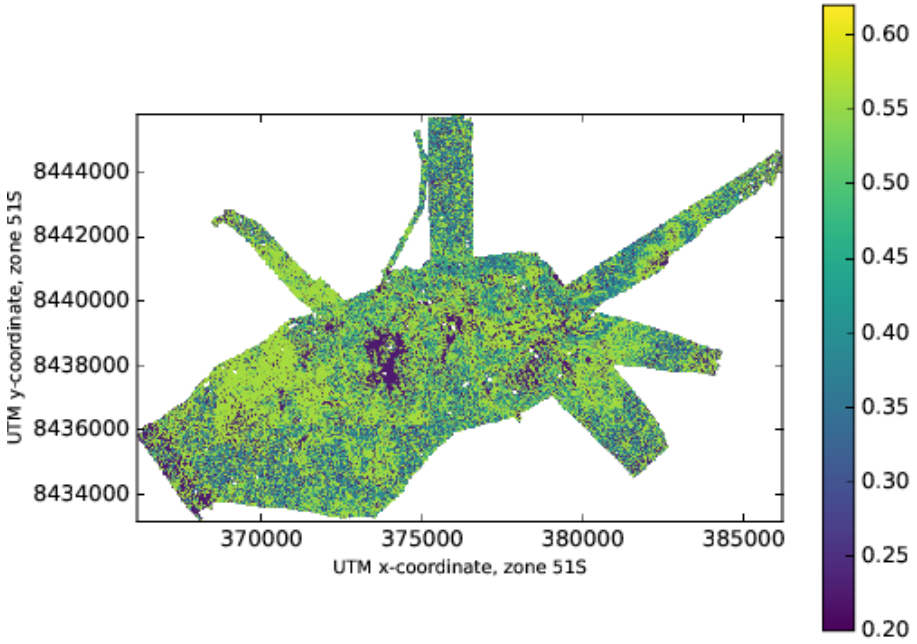


GPoGPE



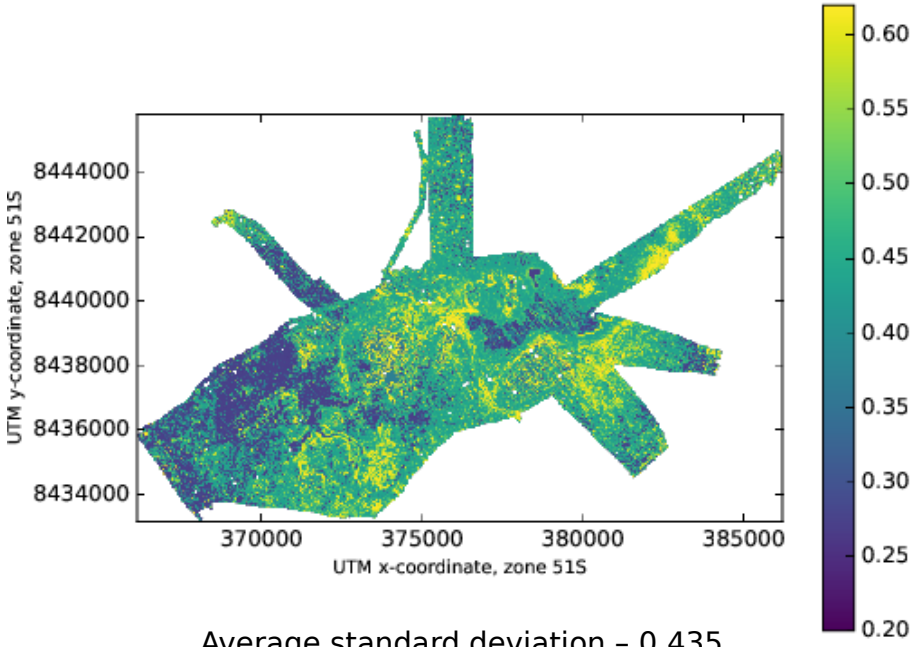
Discussion – Gaussian Process Variance

Gaussian Process Standard Deviation at most likely labels



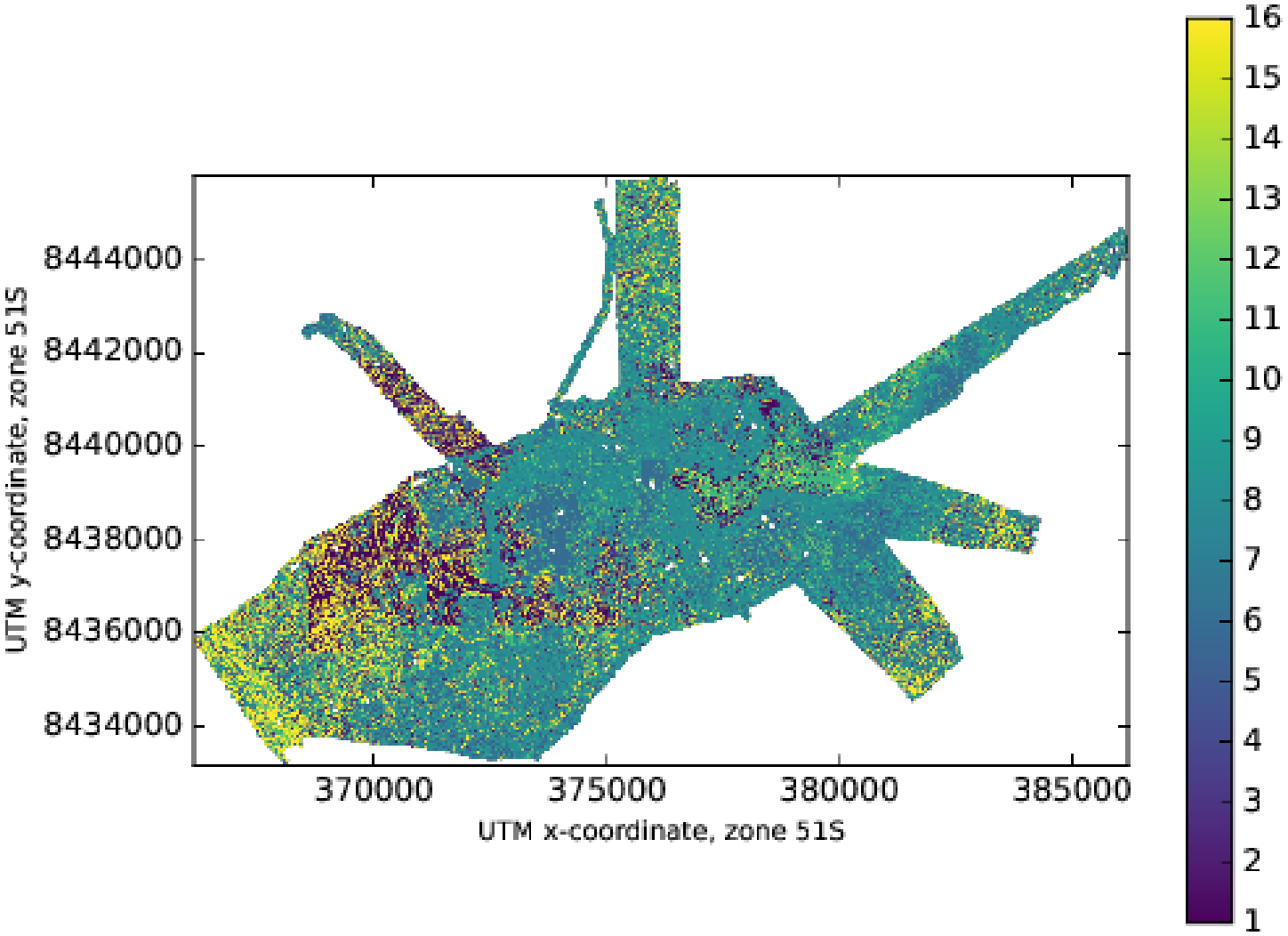
Average standard deviation – 0.449

Generalised Product of Gaussian Process Experts Standard Deviation at most likely labels

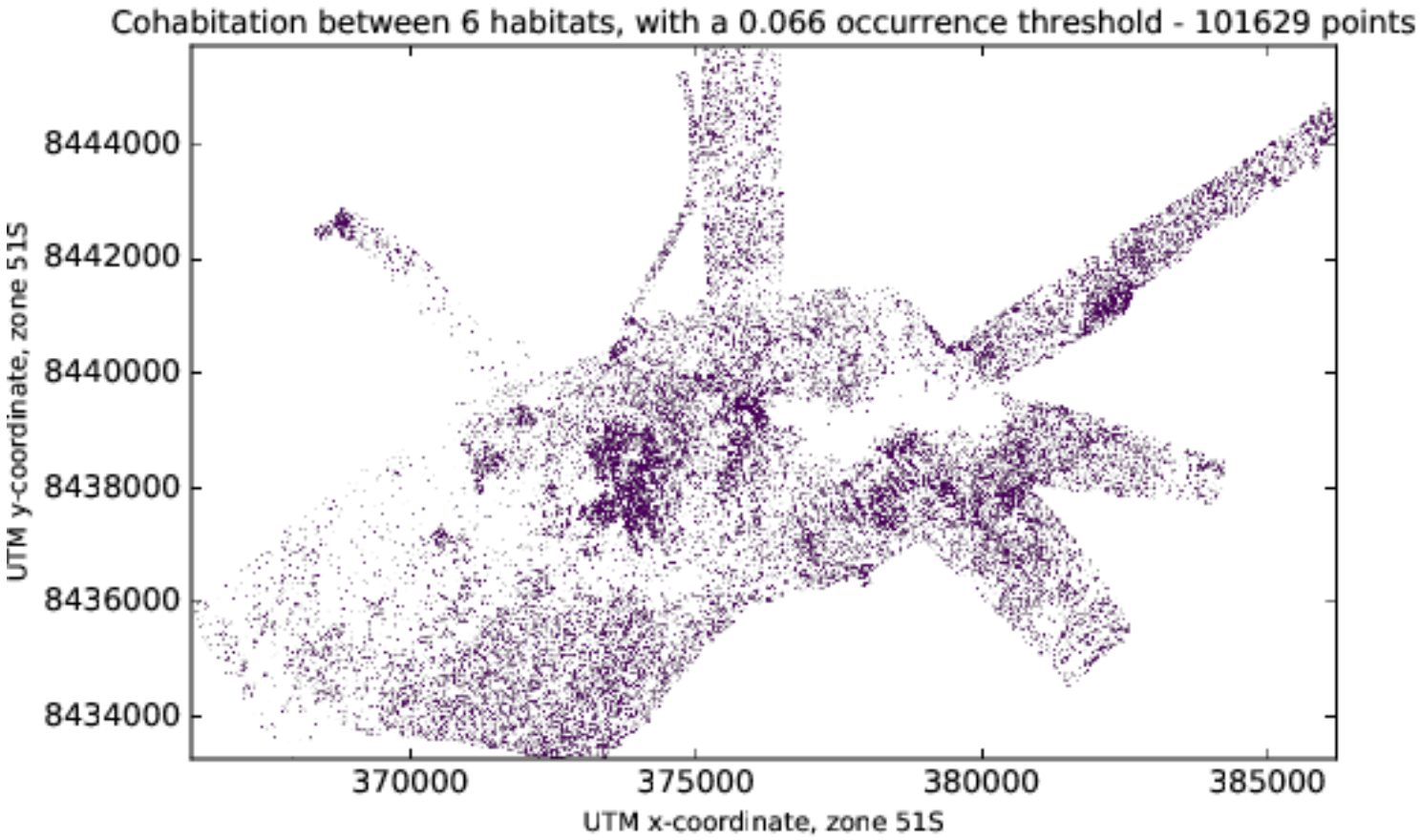


Average standard deviation – 0.435

Discussion – Dirichlet Multinomial Regression



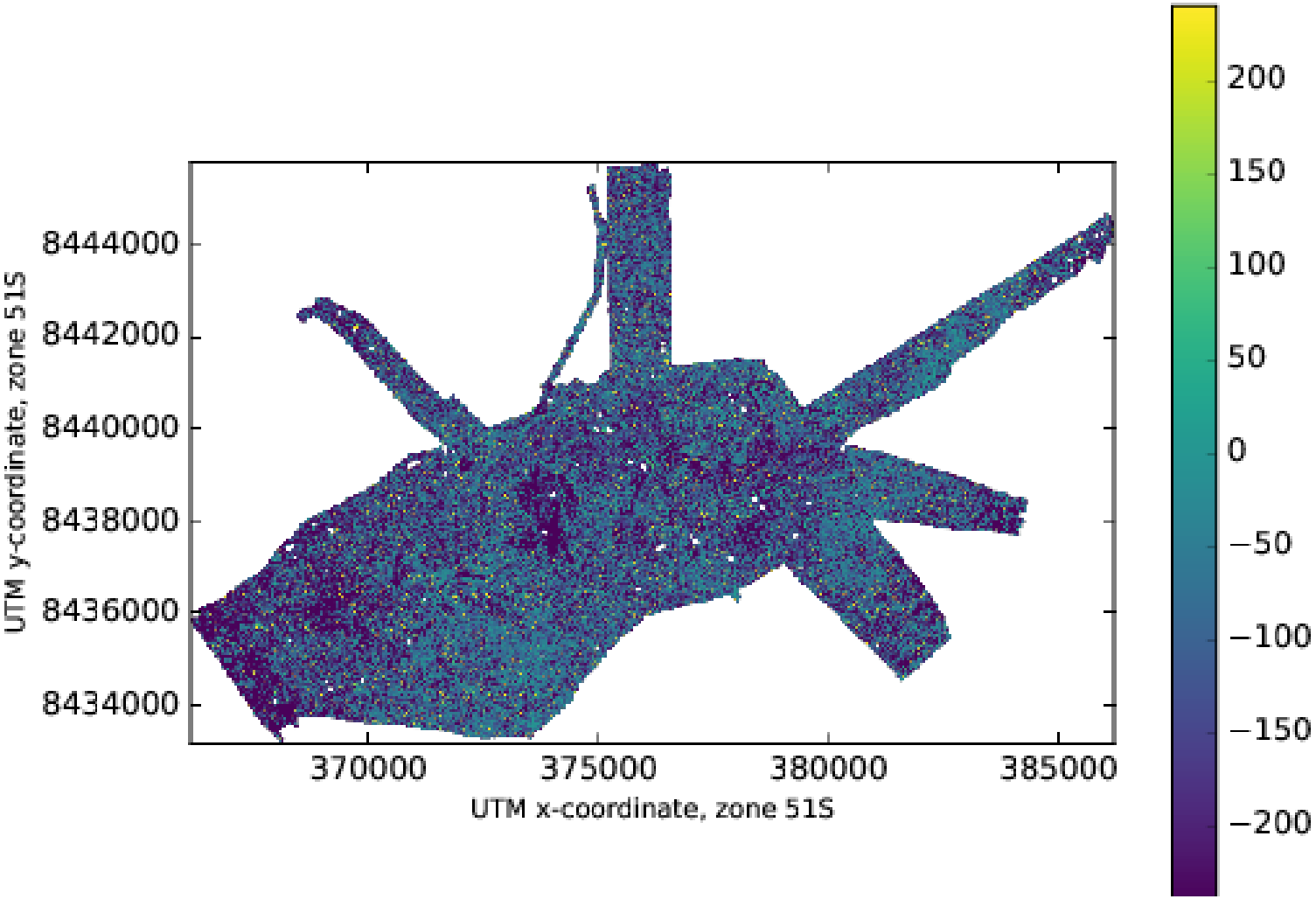
Discussion – Biodiversity



Discussion – Dirichlet Multinomial using parameters from full distribution

<todo – argmax gif>

Discussion – Dirichlet Multinomial Entropy



Limitations

- Marine biologist/expert verification of map quality
-
-

Conclusion

- Lorem Ipsum