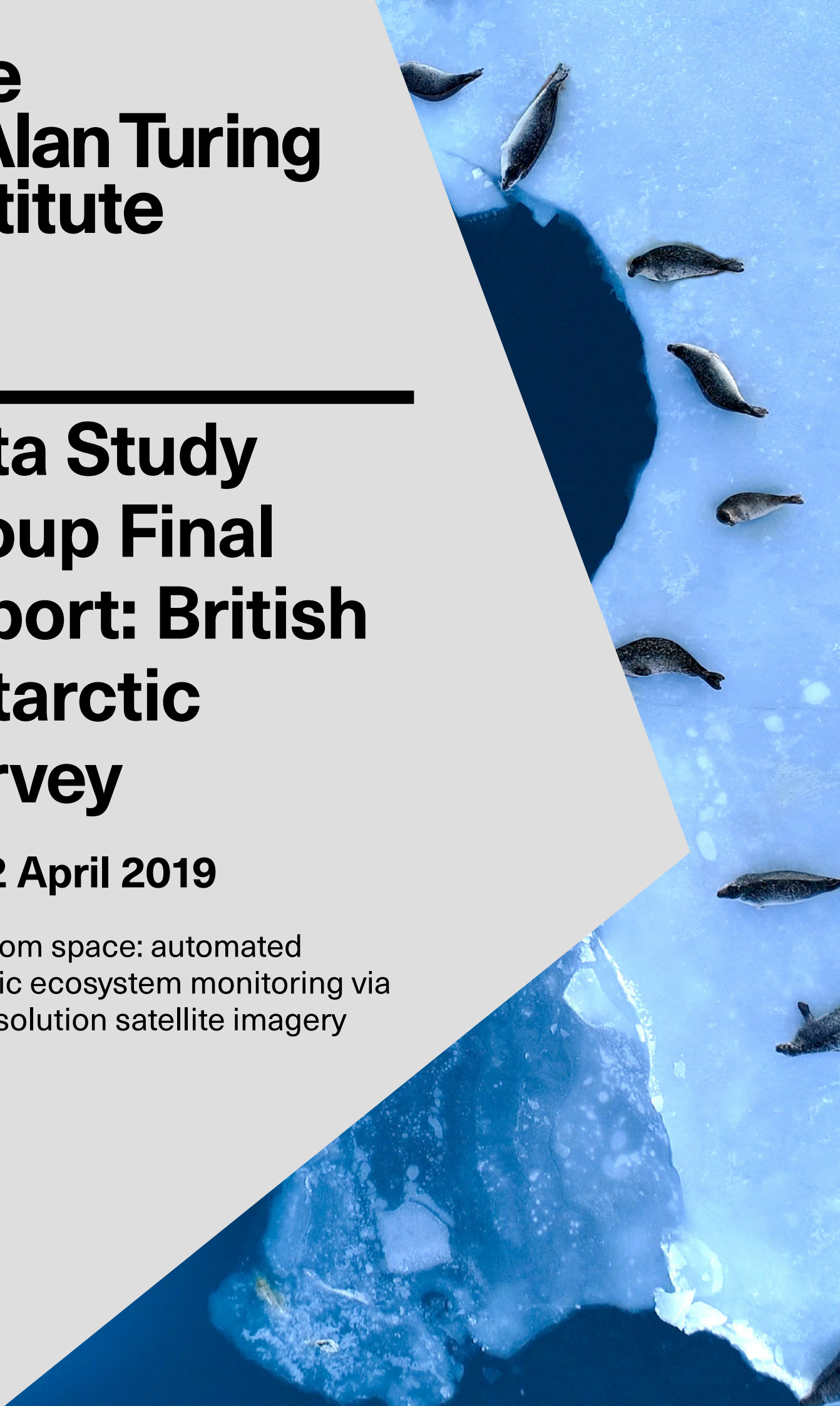# The Alan Turing Institute

---

Data Study Group Final Report: British Antarctic Survey

8 – 12 April 2019

Seals from space: automated Antarctic ecosystem monitoring via high-resolution satellite imagery

Contents

1 Executive summary

1.1 Challenge overview

The Antarctic is constantly evolving as the ecosystem recovers from past exploitation (e.g. whaling, seal harvesting), adapts to climate change, and responds to current anthropogenic impacts including fishing (krill and Patagonian toothfish), shipping and tourism. Due to its vastness, relatively little is known about the ecology of the region and its species, and how best to mitigate and control anthropogenic impacts in this region. Traditional field methods are costly and limited in geographical extent due to areas of interest being difficult to access by ships. Remote sensing provides a low-cost, non-invasive method that can be used for ecological monitoring.

The overall aim of the challenge was to create an automated system for classifying sea ice and mapping seals which can then be used to transform the satellite raster images into images with vectorised features of ice and seals. Using these two outputs (seal counts and ice classification/environmental features) we can explore ecological questions such as what habitat features do seal prefer and how is the habitat changing over time. This report presents the outputs of a week-long collaboration between the Alan Turing Institute and the British Antarctic Survey (BAS), to scope an automated system to classify sea ice, count seals, and explore the environmental factors influencing seal density.

1.1.1 Why use remote sensing?

Fieldwork in the Antarctic is often very challenging due to the large size of the region, inhospitable terrain, limited accessibility, high costs associated with this work. Census data is often biased, as sampling locations are focused around easy to access, existing research stations. In contrast, remote sensing provides a low-cost, non-invasive alternative to field work that is able to cover an immense terrain. Remote sensing is able to

generate high-resolution images of large regions with no impact on vegetation or disruption to the animals themselves.

1.1.2 Why automate?

Manual inspection of satellite imagery is time consuming, subjective and requires trained experts. Automation will allow for faster processing of a larger number of data files/images and the opportunity to produce repeatable results with a standardised method. Studying broad areas, a continental-wide scale becomes possible.

1.2 Data overview

The data consisted of 53 `geoTIFF` files covering two focal areas, the Antarctic Penisula (Crystal Sound and Marguerite Bay) and Signy Island. The geoTIFF files for Signy Island are colour images with a higher resolution of $30\mathrm{cm}$ per pixel, while other `geoTIFF` files are greyscale images with a lower resolution of $50\mathrm{cm}$ per pixel. Location data on 2023 manually-counted seals accompanied these images. This sample is a subset of a much larger satellite dataset, covering millions of square kilometres.

1.3 Main objectives

The sea ice provides an important dynamic seasonal habitat to a range of species such as pinnipeds in the Antartic. Sea ice is likely to vary greatly in the future due to climate change. Understanding the habitat features of the sea ice important to pinnipeds is crucial to understanding how pinnipeds and other species will respond to habitat changes in the future. The main aims of the data study group were to:

1. Classify sea ice into three broad types (open-water, thin brash ice, thick stable ice)

2. Extract small-scale features in the ice of potential biological importance to the seals (ice chunk size (area and perimeter), stable ice edge, thin brash ice edge).

3. Test the applicability of the automatic seal detector convolutional neural network (CNN), developed for the $30\mathrm{cm}$ resolution images to the $50\mathrm{cm}$ resolution images.

4. Explore the environmental features influencing seal density using data on the seal locations and the sea ice features (outputs of steps 1-3) identified to construct a spatial point process model.

1.4   Approach

Classifying sea ice and extracting the location of seals from images may both be considered as image processing tasks. In the case of the sea ice, pixels in the image were classified into 3 general categories (open-water, thin brash ice, thick stable ice) using edge detection algorithms. To count the seals, a CNN model developed for $30\mathrm{cm}$ images was tested on $50\mathrm{cm}$ images. To identify the ecological features that might be important to determining seal locations we constructed a spatial point process model.

1.5   Main conclusions

We developed preliminary a pipeline for processing the satellite images, classifying ice type (stable, thin brash ice, and open water) and extracting several features from the data (distance to open water, distance to thin brash ice, size and location of ice blocks (area and perimeter)). These general methods can be used to process other satellite images in the region.

After testing, we found that the pre-trained CNN developed to count seals from the $30\mathrm{cm}$ resolution images does not perform well when applied to the $50\mathrm{cm}$ resolution satellite images, detecting less than 5% of seals.

We constructed a spatial point process model which incorporates seal locations, spatial covariates (which can be determined from the ice classification sub-project) and spatial effects. We analysed a variety of models including Generalised Linear models, Poisson and Strauss processes and Bayesian Hierarchical models, and found that while ice type was an important determinant of seal presence in our analysis, it only partly explained the clustering observed in the seal locations.

It is important to note that the pipeline and analysis performed is preliminary, and as stated below, faced several limitations. We recommend that the robustness of our findings and their generalizability to further datasets be tested as part of any future work.

1.6 Limitations

1.6.1 Limitations in the data

As we do not have image scenes of locations in which counting has been attempted but zero seals have been observed, we cannot generalise our results to other regions of ice. However, these analyses can identify environmental features that may be important to seal colonies in Crystal Sound. It is also likely that seal location depends on covariates unavailable to us, such as memory of nursing locations and marine food availability.

1.6.2 Limitations of the results

The spatial point process models considered do not include the more detailed features constructed in this project, relying only on local pixel values as a feature. Furthermore, the utility of the spatial point process models explored here are *not* for prediction but validation/simulation. If we assume that Crystal Sound represents a typical seal community and compare seal count estimates from a new region outside, we can check if the estimates are accurate or reliable, adjusting for spatial covariates in the new region.

1.7 Recommendations and future work

For feature construction, some parameter tuning is still required to optimise the result. Further investigation into image smoothing methods, edge detection, and non-Euclidean distance calculations (which for example include penalties dues to local terrain) could also improve the accuracy of the feature classification.

By using fine-tuning methods, we may be able to extend the pre-trained CNN to work effectively on lower-resolution images.

The statistical models constructed can be improved by using the more intricate features of the ice, for example the local brash intensity. A further, more detailed statistical analysis will also help improve the model.

2 Data overview

2.1 Dataset description

The data consisted of 37 panchromatic (greyscale) geoTIFFs covering 2 survey regions at $50\mathrm{cm}$ resolution (taken using the WorldView-2 satellite), and 15 geoTIFFs in both colour and panchromatic at $30\mathrm{cm}$ resolution (from the WorldView-3 satellite), covering one additional survey region. The colour imagery contains 8 spectral bands. 2023 manually annotated seal locations were also provided as a point shape-file with coordinates for each seal.

Due to the constraints of the project we chose to focus on the data from Crystal Sound, one of the two locations sampled from the Antarctic Peninsula. This region covers $193.9\mathrm{km}^2$, and has 708 manually annotated seals. An example image from the satellite data is given in fig. 1.

2.2 Data quality issues

One of the challenges is the variation in resolution and colour content of images taken by different satellites. WorldView-2 provides black and white images from 2011-2016 with a resolution of $50\mathrm{cm}$ per pixel, whilst WorldView-3 data are multispectral images from 2017-2019 with a resolution of $30\mathrm{cm}$ per pixel.

More detailed image information is available within $30\mathrm{cm}$ resolution images, but BAS have more $50\mathrm{cm}$ resolution data available that spans a longer time span. The differences in resolution and colour mean that any analysis on the multi-spectral images has to be adjusted before being applied to monochromatic images.

3   Experiment - Ice Classification

Several environmental features may influence a seal's location on the ice, including the distance to open water or brash (*i.e.* where the ice is broken up into small pieces), the size of pieces of ice in the local area to a seal, and features on patches of ice, for example ridges which can be used for shelter.

3.1   Overview

In this section we discuss how we have converted the qualitative features above into quantitative properties on the image. We provide examples using the image in fig. 1, one of the `geoTIFF` files from the Crystal Sound area.

We have built a preliminary python tool ice/pipeline that can be used to batch process a given `geoTIFF` into our engineered feature layers. The tool works by reading a given file, applying the developed methods and then writing the features back to disk as `geoTIFF` with spatial meta data adjusted accordingly. In future we hope that the tool will be used to batch process satellite data into features, either for use in other software or as
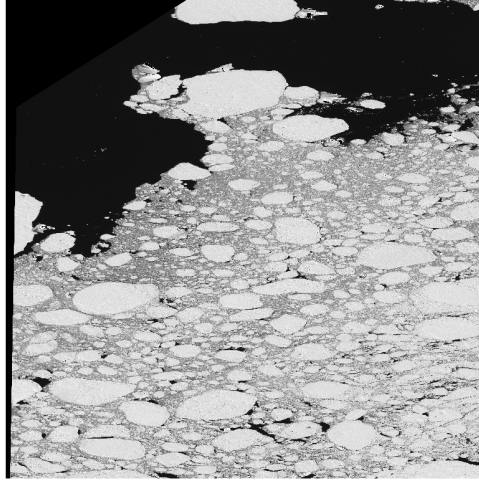
Figure 1: An example of one of the `geoTIFF` tiles considered in the feature detection.

part of a ML pipeline. Additionally the pipeline is designed to be modular so that additional methods can be added as required.

We have chosen to engineer highly interpretable spatial features so that further statistical modelling is explainable. Our features also have potentially broad application as they are, for example for automated classification of ice in an area or querying distance to open water.

3.2 Preprocessing

Before analysing the data, we first convert the `geoTIFF` files to a raster, *i.e.* an array of numbers.

Before considering the features of the image, we need to remove artificially constructed areas of the image; the satellite data is fitted into square image files, where areas which are not part of the original photograph are given by pixels with zero value. We identify these areas by thresholding [13] the

image; any pixels with an intensity below 1 are classed as 'void' pixels, all others are part of the original satellite image, as in fig. 2.
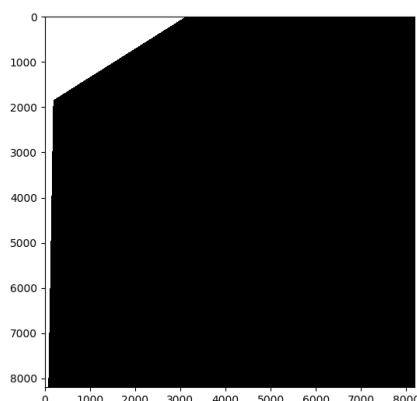

Figure 2: The image shows the mask used to remove pixels that are not from the original satellite image. Invalid pixels here are coloured white, pixels which are from the original satellite image are coloured black.

Furthermore, to use some of image classification methods available in `numpy` we require that pixel values are converted to `uint8` (values 0 to 255), from `uint16` (values 0 to 2047). After a direct comparison of results before and after converting to `uint8`, we verified that minimal detail is lost in the pixel intensity. We produced histograms of pixel intensities before and after image conversion, shown in fig. 3.

The histograms have the same qualitative features, which shows that there is a minimial loss of detail due to image conversion. As these satellite images are large, it is more convenient to use `uint8` pixel intensities to reduce computational cost.

3.3 Methodology and Results

Most of the methods rely on thresholding pixel values, to distinguish ice, water and brash quantitavely. To identify the correct values to threshold the data, we use the histogram of pixel intensities in fig. 3. Two spikes are
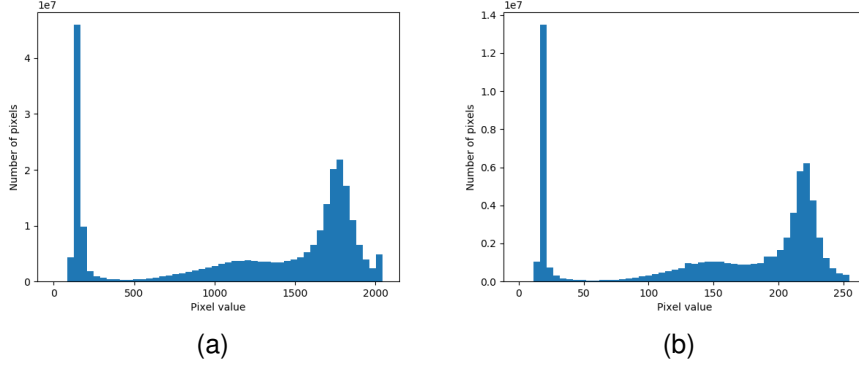
9

Figure 3: Histograms of pixel intensities for different resolution images, each using 50 bins. Left: original raster. Right: pixel intensities after the image has been converted to `uint8` format.

present in both histograms, representing open water (pixel values between 0 and 100 in fig. 3a) and thick ice (pixel values between 1500 to 2000 in fig. 3a), and there is a slight bump for brash ice in the region of pixel values 600 to 1500 in fig. 3a. We use this information in thresholding to segment the image in an appropriate way.

Furthermore, to segment the image we also rely on image smoothing methods [4, 3]; we apply a smoothing convolution to the image, which removes noise from the image and allows for more accurate edge detection.

Distance to open water   To determine the distance to open water, we first need to determine where the open water is. We use a median blur on the image to ensure that small areas of water in areas of brash are not identified as open water. Following this, using the histogram information we choose the appropriate threshold value which identifies the open water area. We then measure the Euclidean distance from each 'ice' pixel to the nearest 'water' pixel, using the `scipy` function `scipy.ndimage.distance_transform_edt` [15]. This then produces a heat map of distances to open water, shown in fig. 4. The computation time for one raster is approximately 19 seconds, showing that the method can be
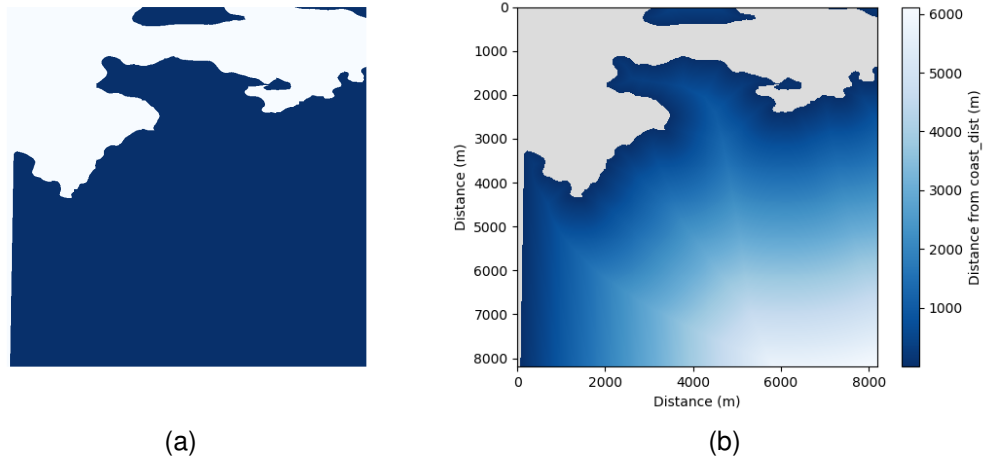
10

efficiently applied to the full set of images.

Figure 4: Distance to open water. Left: smoothing and thresholding is applied to distinguish between open water and ice or brash. Right: The calculated distance to open water from ice or brash.

Distance to brash   By choosing a higher threshold value, we can apply the same methodology to identify the Euclidean distance to brash, as in fig. 5.

How ice is broken up locally   We introduce the notion of brash intensity as a measure of how broken up the ice is locally. To quantify this, we identify the edges using the Laplacian [18] of the image. We then use a convolution with 64×64 averaging filter kernel to determine the intensity of the texture in the local area, given in fig. 6.

The size the piece of ice the seal is on   The result of the Laplacian in fig. 4a is also used to calculate the area of each piece of ice. We apply a convolution with a 16×16 kernel matrix of ones to the result. The output of the processed images (with edge detection filter) was used as an input

11
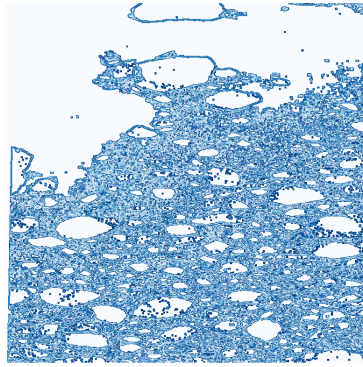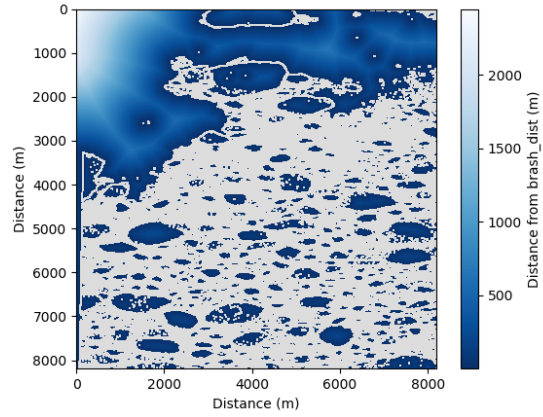
Figure 5: Distance to brash. Left: smoothing and thresholding is applied to distinguish between open water or ice and brash. Right: The distance from ice or open water to brash.

to a polygon generation method [21]. The method uses a mask filter to pick out black areas (*i.e.* where the pixel value is 0) in the image and creates a polygon for region with same pixel value clusters. The pixel (x, y) coordinates are then transformed into CRS spatial (x, y) coordinate (added as meta data). 50 random points are sampled from within the polygon, and the (x, y) pixel position value is extracted from the original TIFF file. The mode of the 50 points pixel values are then used to classify polygon as ice block or water block. The classification is based on a threshold rule. We discard any polygons that are too small to support seals. The result is shown in fig. 7.
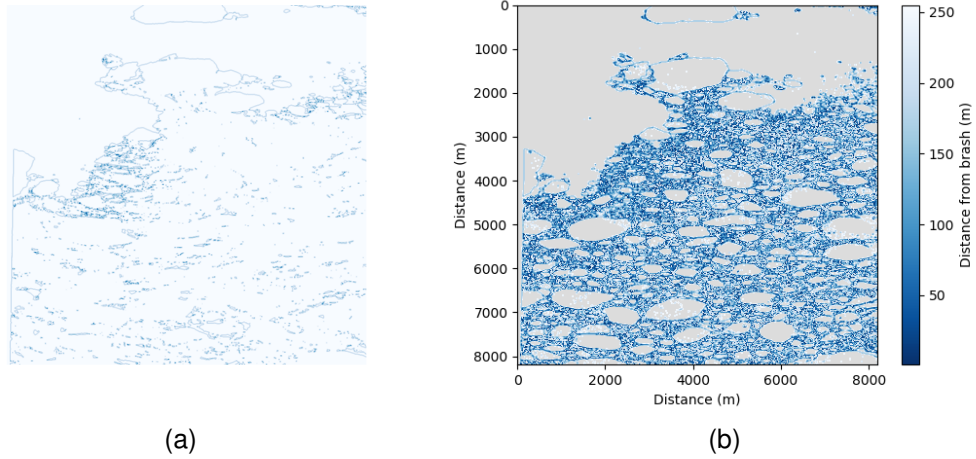
(a)   (b)

Figure 6: A measure of ice texture in the example image; constructed by using the Laplacian (left), followed by median smoothing (right), to produce a brash intensity value.

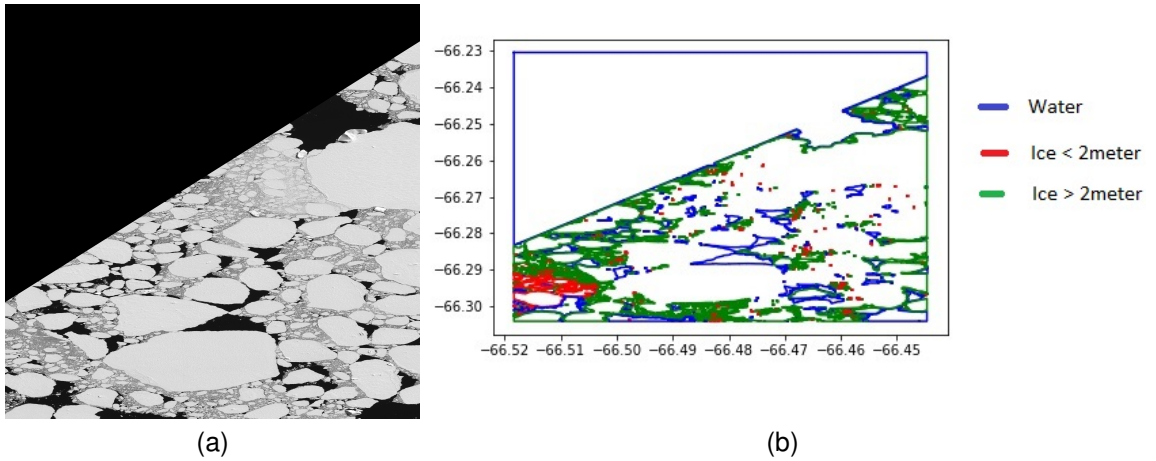(a)   (b)

Figure 7: An example of ice shapes calculated from an image. Left: the original image considered. Right: polygons identified in the image, where different colours depict different size polygons.

4 Experiment - Seal Counter

4.1 Overview

This part builds on a pre-trained (CNN) [14], available in a public GitHub repository `github.com/iceberg-project/Seals`, which was trained on

$30\mathrm{cm}$ resolution satellite images. As part of this task, we have studied whether the parameters learned for the CNN are suitable when applied to $50\mathrm{cm}$ resolution images, and whether simple transformations to the low-resolution images aids the performance of the CNN.

We used the branch `paper/SealNet_code` from `iceberg-project/Seals` in the GitHub repository. The code include scripts for generating tiles, training a model, including different options for the model architecture, and different outputs. Outputs include heatmaps, counts of seals for a given tile, and locations of the seals. There is a notebook with instructions to generate the training dataset. Shape files are used, but we note that it is unclear however how the CSV file for mapping tiles to shape files with ground truth is built. We found the libraries `rasterio` (`https://rasterio.readthedocs.io`), which allows rasters to be read in `python`, and provides transforms between pixel locations and geographic coordinate systems and `fiona` (`https://pypi.org/project/Fiona`), for reading shape files, to be very useful.

4.2 Methodology

The code from the repository provides trained models, trained on $30\mathrm{cm}$ resolution images (or rasters). We attempted to run the pre-trained models on the $50\mathrm{cm}$ resolution rasters off-the-shelf. For this, we separated the satellite images into 224×224 pixel sub-images, to match the size of the $30\mathrm{cm}$ inputs. All images are monochromatic, and are normalised across the channels with hard-coded means and standard deviations.

4.3 Results

The expected performance of the CNN on high-resolution images is around 90%. The table in fig. 8 compares the accuracy of the model for the different resolution images.

Unfortunately, the results show that the model was ill-fitted to be deployed on the $50\mathrm{cm}$ resolution images. We believe this is because each individual seal is identified by a larger number of pixels in the higher

Resolution	30cm	50cm
Performance	90%	2.8%

Figure 8: Performance of the CNN for different resolution images. The results show that the CNN performs poorly on the lower resolution images.

resolution images. In general, the camera specifications for the satellite images for the area of Crystal Sound differ from those that were used to obtain the images that the model has been trained on. While satellite images from different sources look similar to the human eye, they differ significantly enough to confuse the machine learning model.

5 Experiment - Ecological Analysis

5.1 Overview

Based on the results in sections 3 and 4 on ice classification and seal counting, we want to identify which environmental features influence the density of seals on the ice. Furthermore, we want to test if the clustering pattern of the seals locations is caused by the geography or by individual preferences of seals.

We propose several methods based on spatial point processes [10] (which look at the GPS locations of seals) and provide some preliminary results, using the count of seals and images taken of the Crystal Sound. For this analysis, we produced code using `R`.

Firstly, we look at the distribution of seals on the ice. In this analysis, we fixed the observation window `spatstat::owin()` to the convex hull around the seal locations, rather than the default rectangular bounding box, as this will affect estimates and goodness of fit. Our goal is to understand the process in regions where seals are observed to support the seal count methods from earlier sections by providing precision estimates.

Using this information we fit a statistical model to the set which accounts for any clustering that may occur between seals, to include any features constructed in section 3.

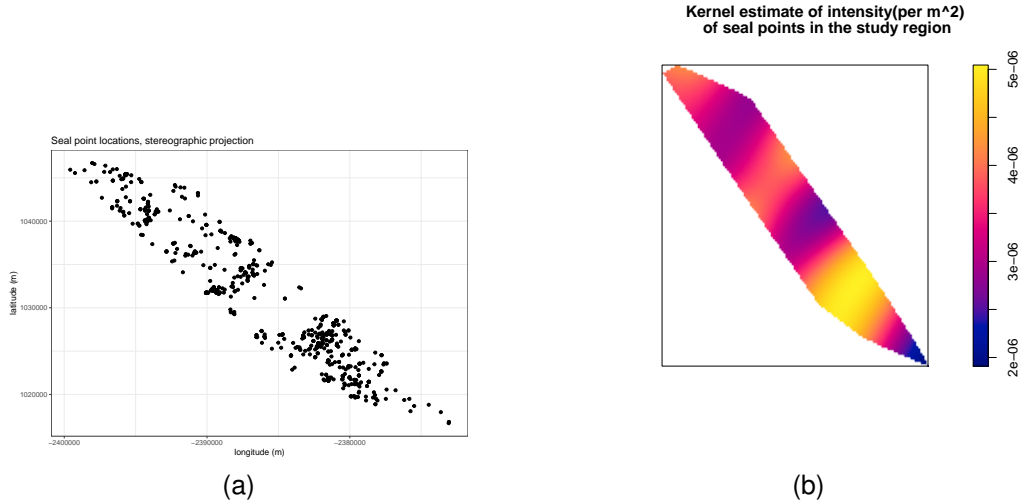(a)   (b)

Figure 9: (Left) the location of seals identified by human visualisers. It doesn't show the teardrop convex hull around the seals but it was there for the analyses. This shows there is a degree of spatial clustering. There are also areas where there are no seals. (Right) This figure shows the spatial structure of the seal's locations far better than the simple point plot. There are three main clusters identified. The range shown in the graph ranges from $2 - 5\mathrm{km}^2$. Note the bandwidth obtained is 13m.

5.2 Methodology and Results

Seal intensity We can produce a non-parametric Kernel-smoothed estimate [16] to allow visual assessment of the spatial structure, given in fig. 9b. The bandwidth of the kernel is objectively selected assuming the seal process is an inhomogeneous Poisson process using Berman & Diggle's method via `spatstat::bw.diggle()` [5].

By constructing a histogram of pairwise distances we can assess the number of clusters of seals (fig. 10). Furthermore, construct Ripley's K function [11] to test for clustering (fig. 11).

We can use a quadrat test [9] to test the homogeneity of the location of seals. A quadrat test shows extremely strong evidence against the null hypothesis of homogeneity ($p < 2.2 \times 10^{-16}$) so inhomogeneous point
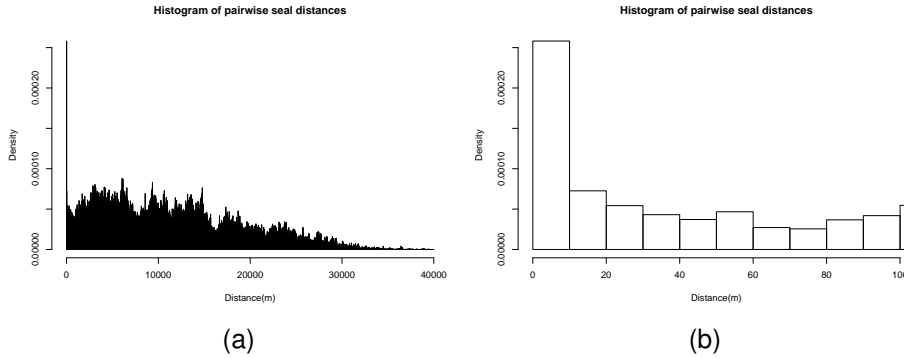
Figure 10: (Left) This distance histogram shows up to possibly five major clusters. The biggest peak is close to distance=0. (Right) This distance histogram is zoomed into the first ten metres from a seal and is probably picking up parent(s?) and their young since this is birthing season.

process models would provide a better fit.

It is also of interest to consider the seals location distribution as a function of pixel colour, which will be related to the type of sea ice the seal is located around (stable ice, thin-brash ice, and open water). We use the kernel intensity estimate to include the location of seals. We display the pixel intensity near each seal versus the pixel intensity of the image in fig. 12, which can be used as a spatial covariate in the point process model.

5.2.1 Spatial point process model

If seals determine their location on the ice independent of environmental features we would expect their locations to be randomly distributed across the study space. We can identify environmental features that may influence the seal's location using a generalised linear model.

Generalised linear model. For some intensity of seals μ at any location (x, y) in the study region:

$$\log \mu(x, y) \sim S(x, y) + X_i(x, y) + Y_j(d)$$
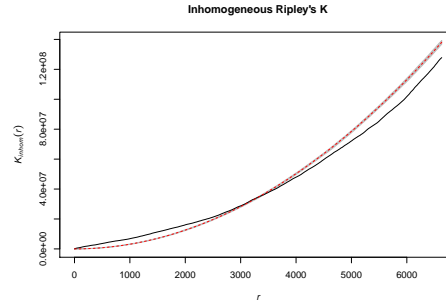
17

Inhomogeneous Ripley's K

Figure 11: This graph shows a black solid line as the estimate of the inhomogeneous Ripley's K function and another line with grey banding as the theoretical inhomogeneous Ripley's K function (*i.e.* no clustering). There is clustering between $0 - 3\mathrm{km}$ and then inhibition from $3 - 7\mathrm{km}$. Presumably the major clusters seen in previous plots are dominating the signal as their radius would be similar to $3\mathrm{km}$ in radius. The inhibition process may be linked to the maximum food availability available for a typical seal cluster.
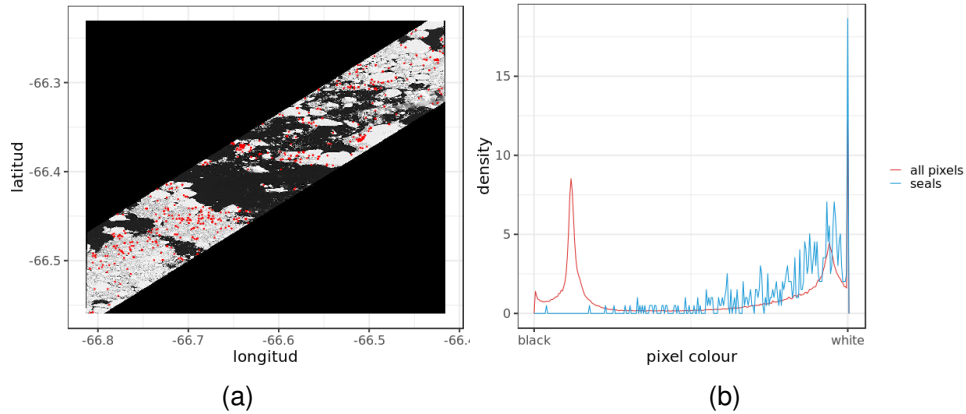

(a)   (b)

Figure 12: Left: Red points indicate the location of seals. Right: the density of colours for all pixels surroundings vs. the colours of the pixels around a seal location.

where S is the spatial process solely governing the seals interactions, $X_i(x, y)$ are spatial covariate(s) at the point (x, y), or spatial covariates(s) $Y_i(d)$ for point, line or polygon features that were a distance d away from

18

the location (x, y) [12]. This model structure allows us to compare the relative contributions between seal-seal interactions (*e.g.* spatial clustering or inhibition), and seal-environment interactions.

5.2.2 Models without spatial covariates

An initial simple model included only the seal process S, and ignored the spatial covariate terms X, Y. S was changed to minimise the Akaike information criterion (AIC) [1] to obtain the best fit. Inhomogeneous is defined as a linear trend in a direction in (x, y) space. We considered 4 different models for S:

1. Homogeneous Poisson process

2. Inhomogeneous Poisson process

3. Homogeneous Strauss process, with $r = 3000m$

4. Inhomogeneous Strauss process, with $r = 3000m$

where r was obtained from the clustering scales observed in the earlier descriptive analysis. The results are given in fig. 13

Model	AIC
Homogeneous Poisson	19146.91
Inhomogeneous Poisson	19146.61
Homogeneous Strauss	130.353
Inhomogeneous Strauss	134.557

Figure 13: The AIC for different models. The results show that the homogeneous Strauss model provides the best fit for the data.

Plotting the fitted model's intensities on a map and comparing with the kernel-smoothed map from the descriptive analysis visually shows how well the model could predict the observed intensities, for example we

19

consider the Strauss model in fig. 14. These results show that a simple model is not sufficient to simulate seal locations.

**Fitted values of intensity,
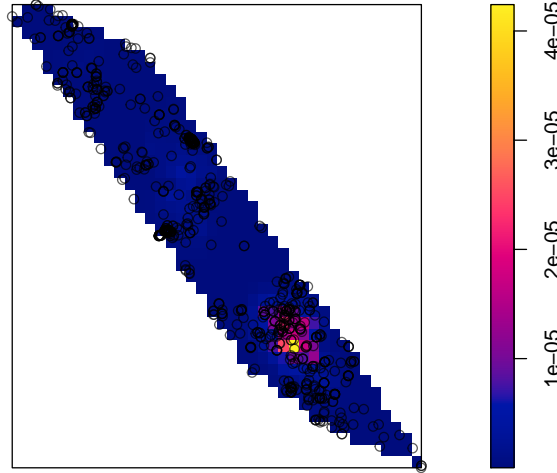conditional on a Strauss clustering process (r=3000m)**

Figure 14: The plotted fitted values show a poor fit of the homogeneous Strauss model to the kernel estimated intensity.

5.2.3   Bayesian hierarchical models

We propose a point process to relate the seal counts and any feature of the geography. The model is constructed as a hierarchical model [2], where the risk of observing seals in a given location is the main parameter to be estimated. The latent field should capture the features of the ice, thickness, and any measure taken by the ice classifier, as covariates. A spatial term should capture any spatial effect not explained by the covariates, giving us insights about the nature of the seals distribution (*e.g.* agglomerative trend, random behaviour).

20

Spatial effects can also be incorporated using, for example, a Besag model, a Besag-York-Mollie model [6] or a Gaussian Markov Random Field [19]. Here we implement a Besag model. The output of the model includes the posterior distribution of all the parameters included. Other outputs include a map of the probability of finding a seal given the geography of the region studied, and a map of the spatial effect.

A Bayesian hierarchical model was implemented using `INLA`. The area was divided in 60000×60000 squares. The logarithm of the counts on each lattice was modelled as the combination of one covariate and a spatial effect. An ice/sea classification was included as a covariate while an spatial effect was designed using a Besag model on a lattice.

We refer back to fig. 9 for the input of the model. The mean and the standard deviation of the posterior of each spatial effect is shown in fig. 15. The prior and posterior of the covariate (ice/sea class) is shown in fig. 16.
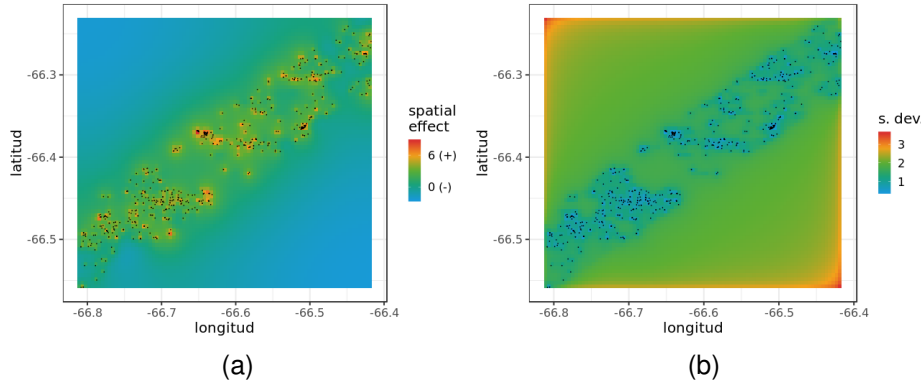

(a)   (b)

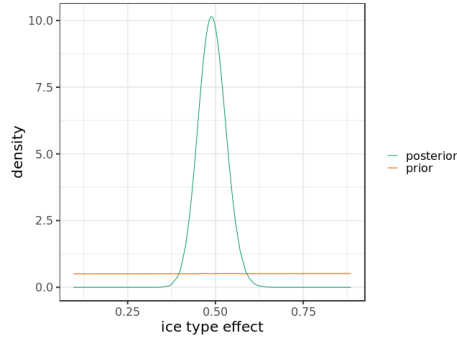Figure 15: Mean (left) and standard deviation (right) of the posterior of each spatial effect.

Figure 16: Prior and posterior of the ice/sea class.

6 Future work and research avenues

6.1 Ice classification

By using smoothing and thresholding in a similar as in this report, one could also detect features in the ice such as ridges or gradients. There are many different available methods for edge detection and image segmentation (*e.g.* the constrained graph Laplacian [8] and Canny edge detection [7]), a detailed investigation into alternative methods for this application would be beneficial.

Improvements in the distance calculations are also necessary to make them more robust. Using the methods constructed in this report, features are only constructed for each tile of the `geoTIFF` file. This may mean inaccuracies arise, where the smallest distance to, for example, open water, covers two tiles, or when the smallest distance is given by the distance to the edge of the satellite image. By considering the entire image we can resolve distances between adjacent tiles, however this would greatly increase the computational cost. This problem requires more investigation.

The way distance is measured could also be improved. In this project, due to time constraints we considered Euclidean distances to open water and brash. However, movement of seals is expected to depend on local terrain. For example, the Euclidean distance doesn't take into account ridges in the ice which can't be traversed by seals. Investigation into

weighted distances [22] and particle approximation models may be beneficial to improve distance calculations.

6.2 Seal counter

One may learn a model that gives more precise information, such as where seals are located. This requires modifications to the labels on the data, as well as some modifications to the output layer of the CNN.

To go forward, we suggest to train the model on the $50\mathrm{cm}$ resolution images. The pre-trained parameters of the model are likely a good initialisation for this, and we therefore suggest a fine-tuning [17] of the parameters (which are downloadable from the Github repository), using the lower-resolution imagery. Fine-tuning should require substantially less labelled seals than were required in training the machine learning model from scratch.

6.3 Ecological analysis

The more detailed features constructed in this report should be implemented as additional spatial covariates in the point process model. Following this, a detailed statistical analysis of the results would provide feedback for the features and help improve the model. The model should then be applied to more satellite images, using seal counting methods from the CNN results, to verify the model.

7 Project Contributors

Aditya Acharya is a Research Associate at the Alan Turing Institute. He works on modelling city transport and spatial analysis. His past research revolved around behaviour modelling and computer vision research. His contribution in this project was to extract ice features from given satellite images. He also helped to re-engineer CNN model for seal detection on different resolution images.

Anna Laws is a PhD student in Astrophysics at the University of Exeter, where she takes, processes, and analyses observations of planets forming around other stars. Her main role in this data study group was as the facilitator. She also contributed to the ice classification methods sub-project.

Christopher Tegho studied machine learning at the University of Cambridge and is currently working as a machine learning engineer at Calipsa, focusing on research for video classification. Chris investigated previous work for automatically counting seals from high resolution imagery, and evaluated the performance of the models trained on the $30\mathrm{cm}$ resolution imagery and tested on the $50\mathrm{cm}$ resolution imagery.

Eleanor Miller is a final year PhD Student at the University of Cambridge where she is studying population genetics, primarily focusing on changes to population size in Northern Hemisphere bird species since the Last Glacial Maximum. She primarily worked with the ice-classification team as well as contributing to group discussions and report writing.

Fred Shone is a Research Data Scientist at Arup London, primarily working on massive simulations using Agent Based Models and new algorithms to build real-time transport networks. His past research has been focused on pedestrian analytics using computer vision. Fred contributed to the image classification sub-project, in particular in data Engineering and pipe-lining raster processing and 'polygonizing'.

Laura Marcela Guzman Rincon is a PhD student in Mathematics of Real-World Systems at the University of Warwick. Her work is aimed to detect outbreaks of Campylobacter infections using epidemiological and genetic data. Laura worked with Timothy Pollington on the ecological model, in particular designing and coding a Bayesian hierarchical model to relate the seal locations and any geographical features.

Laurie Baker is an MRC Transition Fellow at the University of Glasgow where she is studying wildlife epidemiology using Bayesian spatial modelling techniques. Her past research focused on grey seals on the Scotian Shelf in Canada. She contributed to the problem formulation for ice-classification and the environmental modelling and report writing.

Lucas Deecke is a PhD student in machine learning at the University of

Edinburgh; his research broadly focuses on deep learning, continual learning, and computer vision. Lucas contributed an evaluation of convolutional neural networks for different resolutions of satellite imagery.

Prem Gill is a polar conservationist and PhD student with British Antarctic Survey (BAS), Scott Polar Research Institute (SPRI) - University of Cambridge and the World Wildlife Fund (WWF). His main research focus is the study of ice seals via state-of-the-art remote sensing techniques. Outside of this, Prem is heavily interested in providing opportunities in conservation and polar science for under-represented groups. As the DSG Challenge Owner, Prem set the challenge and supported participants by providing background domain expertise / insights into seal ecology, sea ice dynamics, the dataset and application of previous techniques. He also used this opportunity to showcase potential careers for Data Scientists within polar research to the participants.

Tim Hurst is a final year PhD student in applied mathematics at the University of Edinburgh and Heriot Watt University. He works on constructing mathematical models for various problems with multiple scales, and including information from real world or synthetic data. In the data study group, Tim focused on image classification methods and feature engineering, including edge detection methods and texture identification to help classify ice in satellite images.

Timothy Pollington is a University of Warwick PhD student based at the Big Data Institute, Oxford, focusing on the spatial analysis of infectious diseases. Timothy worked with Laura on the ecological model, focusing on descriptive analysis (including distance histograms, kernel-smoothed intensity estimation, and analysis using Ripley's K function) to better understand how observed seal colonies are distributed spatially, to provide a measure of precision when future seal counts are made in similar settings. Timothy also produced code for a statistical point process model to fit as a generalised linear model to the seal location data and spatial covariates.

Victor Sanchez-Silva is the challenge PI, and is an associate professor with the Department of Computer Science, University of Warwick, U.K. His main research interests are in signal and information processing with

applications to multimedia analysis, security, image and video coding, and communications. He has published several technical papers in these areas and coauthored a book [20].

8 Acknowledgements

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

[2] G. M. Allenby. Hierarchical bayes models. *The handbook of marketing research: Uses, misuses, and future advances*, pages 418–440, 2006.

[3] G. R Arce. *Nonlinear signal processing: a statistical approach*. John Wiley & Sons, 2005.

[4] G. Aubert and P. Kornprobst. *Mathematical problems in image processing: partial differential equations and the calculus of variations*, volume 147. Springer Science & Business Media, 2006.

[5] M. Berman and P. Diggle. Estimating weighted integrals of the second-order intensity of a spatial point process. *Journal of the Royal Statistical Society*, B 51:81–92, 09 1989.

[6] J. Besag, J. York, and A. Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20, 1991.

[7] J. Canny. A computational approach to edge detection. In *Readings in computer vision*, pages 184–203. Elsevier, 1987.

[8] W. Casaca, L. Gustavo Nonato, and G. Taubin. Laplacian coordinates for seeded image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 384–391, 2014.

[9] N. Cressie and T.R.C. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3):440–464, 1984.

[10] D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media, 2007.

[11] P. M Dixon. Ripley's K function. *Wiley StatsRef: Statistics Reference Online*, 2014.

[12] P. K Dunn and G. K Smyth. *Generalized linear models with examples in R*. Springer, 2018.

[13] R. C. Gonzalez. *Digital image processing*. Pearson Education, 2009.

[14] S. Khan, H. Rahmani, S. A. A. Shah, and M. Bennamoun. A guide to convolutional neural networks for computer vision. *Synthesis Lectures on Computer Vision*, 8(1):1–207, 2018.

[15] J. Kinser. *Image operators: image processing in Python*. CRC Press, 2018.

[16] Q. Li and J. S. Racine. *Nonparametric econometrics: theory and practice*. Princeton University Press, 2007.

[17] F. Radenović, G. Tolias, and O. Chum. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In *European conference on computer vision*, pages 3–20. Springer, 2016.

[18] M. Reuter, S. Biasotti, D. Giorgi, G. Patané, and M. Spagnuolo. Discrete Laplace-Beltrami operators for shape analysis and segmentation. *Computers & Graphics*, 33(3):381 – 390, 2009.

[19] H. Rue and L. Held. *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC, 2005.

[20] B. Sobolev, V. Sanchez, and L. Kuramoto. *Health care evaluation using computer simulation*. Springer Science & Business Media, 2012.

[21] S. Wu and M. R. G. Marquez. A non-self-intersection Douglas-Peucker algorithm). In *16th Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI 2003)*, pages 60–66, Oct 2003.

[22] Z. Xu. Hybrid weighted distance measures and their application to pattern recognition. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 17–23. Springer, 2008.