# Data Science MSc Project Proposal

## Developing a forest canopy change detection model for GB using Sentinel 1 Synthetic Aperture Radar

The aim of the project is to examine machine learning techniques for detecting change in woodland cover in Britain using Sentinel 1 imagery, which is freely available from the European Space Agency (ESA). The Sentinel 1 mission captures C-band radar, which allows data to be captured day or night, irrespective of cloud cover [1]. The Sentinel 1 mission has been running since 2014 and data is available at least every 12 days for the same location [2]. Traditional optical satellite imagery is available from ESA’s Sentinel 2 program, as well as NASA’s Landsat missions and may be used in this project in conjunction with Sentinel 1 if required to improve model accuracy. For Forest monitoring, techniques using optical imagery are well developed, using techniques such as a Normalized Difference Vegetation Index (NDVI) [3]. For this reason this project is particularly keen to explore what can be achieved with Sentinel 1 Radar data, which may have specific benefits for forest monitoring:

* It may detect ground disturbance from planting of new trees, where trees are not yet large enough to be visible on optical imagery or aerial photography.
* It may be possible to detect changes in canopy cover by forest thinning where the canopy still looks relatively undisturbed on optical imagery.
* It can work regardless of cloud cover, so it may be possible to spot deforestation sooner than waiting for a cloud-free optical image, which is big challenge when using optical imagery in Great Britain [7].

The imagery from the Sentinel Satellites may be regarded as Big Data because of the imagery’s size and frequency of capture. Sentinel captures approximately 156 TB globally per annum [4] and it is estimated one flight over Great Britain captures 30 GB of imagery, so approximately 1 TB per annum. The Sentinel datasets may also be described as Big Data due to the structure and type of data, consisting of pixel values in an image with the values representing radar backscatter in decibels [6]. One of the challenges will be handling this volume of data. Although the Sentinel 1 data is freely available to download from the ESA website, numerous pre-processing steps are required before it can be used for analysis [5]. Tests have already been performed on the pre-processing steps using ESA’s SNAP Toolbox and these steps involve deriving copies of the image, further increasing the storage space required.

To mitigate the challenge of storing this volume of data and to enable the focus to be data analysis and model building, rather data preparation, it is hoped that the Sentinel 1 imagery Google currently serves up as part of its Earth Engine can be used. Google have already implemented a specific set of pre-processing steps on the Sentinel 1 data they provide [6]. The two challenges with using the data in Google Earth Engine will be:

1. Is the pre-processing that has been applied to the Sentinel 1 data, appropriate for GB and for Forestry?
2. Can analysis be carried out in Google’s cloud based environment, without incurring huge charges in using the Google Cloud Platform?

From a proof of concept project the Forestry Commission was involved in using Sentinel 1 data [8], it is known that a multitemporal analysis is required, i.e. a stack of the images captured every 6-12 days need to be analysed together because little insight is gained from one image analysed in isolation. This will add to the challenge in data processing and creating features to train a machine learning model.

The aim is to make use only of open source software and tools and to take a ‘Data Science’ approach to the problem, rather than using specialised non-open source remote sensing software (specialist software such ERDAS Imagine, ENVI, eCognition). The following tools will be used to begin with:

* ESA SNAP Toolbox for initial image familiarisation and analysis techniques
* QGIS for reviewing and preparing training features (vector data)
* Python Google Earth Engine API
* Python spatial libraries (GDAL, RasterIO)
* Python machine learning libraries, such as Ski-Kit Learn

The aim is to produce polygons showing where change has occurred in woodland over the period in which the Sentinel 1 data has been available (2014 to 2018). The publicly available National Forest Inventory (NFI) Map [9] will be the constraining extent which is monitored, although some thought will be given to supplementing this with the LandCover Map 2015 [10] to see if it is possible to detect woodland in land parcels outside of the NFI Map.

Ultimately results would need to be validated by field visits, but this is not deemed necessary for this project. Using the attributes in public datasets (NFI and LandCover Maps and the Sub-compartment database [11] covering the National Forest Estate), it will be possible to assess results in the manner of a supervised machine learning task, by holding out a test set of features and using separate sets for model training and validation. The success rate on the test set of features will be the measure of success in the results that will be presented.

A project plan with timescales is provided in a supplementary document. These are the tasks which will need to be accomplished:

1. Familiarisation with Sentinel 1 SAR imagery properties and analysis techniques, following online tutorials from ESA.
2. Understand what can be achieved with Google Earth Engine’s source of Sentinel 1 imagery and if it will be suitable for data access and analysis.
3. Develop Python scripts to build features suitable for training machine learning models, incorporating Sentinel 1 pixel properties, training site properties (woodland type, age), auxiliary information as discover necessary such as rainfall data for the date of image capture.
4. Experiment with different supervised machine learning modelling algorithms (Random Forest/ XGB Boost have success in similar use cases, is a deep learning technique applicable?)
5. Experiment with different input training feature attributes to improve model prediction accuracy.
6. Present results and write up.

References:

1. <https://sentinel.esa.int/web/sentinel/user-guides/sentinel-1-sar/overview>
2. <https://sentinel.esa.int/web/sentinel/user-guides/sentinel-1-sar/revisit-and-coverage>
3. <https://earthobservatory.nasa.gov/Features/MeasuringVegetation/measuring_vegetation_2.php>
4. <https://pdfs.semanticscholar.org/presentation/5352/4af9af17e0acb2ece18183bfe027082450be.pdf>
5. <https://sentinel.esa.int/web/sentinel/level-1-pre-processing-algorithms>
6. <https://developers.google.com/earth-engine/sentinel1>
7. <https://www.gislounge.com/methods-creating-cloud-free-satellite-imagery-sentinel-2/>
8. <https://www.capgemini.com/2017/08/how-to-do-machine-learning-on-satellite-images/>
9. <http://data-forestry.opendata.arcgis.com/datasets/national-forest-inventory-woodland-gb>
10. <https://www.ceh.ac.uk/services/land-cover-map-2015>
11. <http://data-forestry.opendata.arcgis.com/datasets/national-forest-estate-subcompartments-gb>
12. <http://step.esa.int/main/toolboxes/snap/>