**Prey Lang reference data procedure**

**1. Introduction**

This document outlines a proposed methodology for generating the reference data for an automated disturbance detection algorithm for the Prey Lang Wildlife Sanctuary. This document will first explain the reference datasets that are available and will then clarify how these reference data will be used within the training of the automated alert system. Additionally, this document describes additional validation that must be acquired and will provide a proposed procedure for generating the reference data.

**2. Reference datasets.**

**2.1. Hand digitized field data**

A dataset of 777 reference polygons were collected across the Prey Lang Wildlife Sanctuary. The reference polygons were digitized from Sentinel-2 imagery. Only a subset (n=XXX) of the original dataset contained valid metadata which indicates the day the disturbance was observed. The dates of the recorded disturbance are given in Figure X.

**2.2. GLAD forest alerts.**

**2.2.1 GLAD Forest alert system summary**

<https://iopscience.iop.org/article/10.1088/1748-9326/11/3/034008/pdf>

GLAD forest alerts were developed to do XYZ by CITATION. INSERT BREIF SUMMARY OF THE ALERT SYSTEM. FROM WHAT DATES DOES DOES THE GLAD FOREST ALERT SYSTEM SPAN?

**2.2.2. Preprocessing for forest alert training data**

The Forest alert data are distributed as multiband GEOTIFF files that contain the alert information for 2019 and 2020. For each year, the alert system produces two layers of interest: 1.) a layer containing the confidence estimate for each mapped disturbance and 2.) a layer where each pixel contains either the first mapped date of a forest disturbance or a background value (0). Each image were processed into a time-series of 365 disturbance/no-disturbance images. For example, a pixel labeled “195” in the 2019 disturbance occurrence layer would be labeled a 1on the 195th layer of the output time-series. To control the amount of commission error in GLAD alerts, a 2-pixel minimum mapping unit was applied *initial* forest alert system; this preserves the inclusions of disturbances which were initially small but later grew into larger patches. The two time-series of forest alerts were then combined into a single time-series of XXX images.

**3. Proposed training dataset**

The following procedure is designed to take the following datasets 1.) a time-series of daily GLAD forest alerts and 2.) a dataset of sentinel 1 imagery and process them into a training and testing data suitable for training a fully convolutional neural network capable of identifying disturbances.

Machine learning development and validation is generally conducted in two phases. The entire dataset set is split into the following sets for model development: the training set, the validation set and testing set. The model is developed using the training and validation sets. The model is then fit to the training set and the validation set is used to provide an unbiased estimate of model’s generalization error. We then iteratively optimize the model’s architecture and hyperparameters of the machine learning model improve how the model learns and generalizes. This process biases our ability to the correctly estimates model’s generalization error using the validation set. Therefore, we get our final assessment of model’s accuracy from the testing dataset.

In this analysis, the training and validation dataset will consist of Nx256x256 image chips with N-1 bands of features and a binary disturbance/no-disturbance label. The test set will be created using Collect Earth Online (Saah et al., 2019) and will consist of photo interpretations over a spatially, temporally random sample of disturbance classifications. The details about the proposed training, validation, and test sets are given below.

**3.2 Training and validation sets**

Training data will be gathered across the area detailed in FIGURE X. The study area will be partitioned into 10 areas of equal size. Data drawn from 3 randomly selected areas will be used to create the validation set; the training dataset will be drawn from the remaining partitions. Each partition will be subdivided into 256x256 subdivisions from which the training and validation features will be aggregated, these subdivisions will be allowed to overlap (FIGURE X).

**3.3 Testing set details**

To determine the number of samples in the validation set we require, we use the formula given by Cochran 1977:

This

**4.1. Validation data characteristics**

**4.2. Data collection methodology**

**4.3 Criticism to address:**

* *Sampling of disturbance agents*
* How do you account for systematic bias due to cloud coverage?

**5. Scripts used:**

Figure 1: