Plan Overview

A Data Management Plan created using DMPTool

DMP ID: https://doi.org/10.48321/D13366

Title: DMSP for "Monitoring Abundance of Bumblebees and Honeybees Using Optical Sensors"

Creator: Benjamin Thomas - **ORCID:** 0000-0002-8880-0795

Affiliation: New Jersey Institute of Technology

Principal Investigator: Benjamin Thomas, Gareth Russell

Contributor: Cristo Leon

Funder: United States Department of Agriculture (usda.gov)

Funding opportunity number: USDA-NIFA-AFRI-009755

Grant: https://www.grants.gov/web/grants/view-opportunity.html?oppId=345796

Template: USDA-NIFA: National Institute of Food and Agriculture

Start date: 02-01-2024

End date: 01-31-2028

Last modified: 01-19-2024

DMSP for "Monitoring Abundance of Bumblebees and Honeybees Using Optical Sensors"

The proposed work includes three experiments, each using an eBoss instrument (1 semi-controlled experiment, 2 field experiments). The primary raw data generated by the eBoss is a digital 1D vector recoding the measured voltage across the optical detector of the instrument. Transit signals will be collected every time a flying insect crosses the field of view of one of the instruments, it is the primary data used for analysis. Additional meta-data from co-located weather stations will also be recorded.

Across all three experiments, every individual transit signal is systematically stored as a distinct .csv file. Each file captures a span of 1 second of recording at a rate of 100 kHz, equating to a 1D vector of 100,000 data points per transit signal. Essential metadata pertaining to the signal is added to this file, encompassing details such as the date, time, instrument identifier, temperature, and incident solar radiation level.

In addition, the predictor variables obtained from the analysis of each captured signal are add to each .csv file. These variables, integral for subsequent analyses, encapsulate significant attributes of the transit signals. The extracted predictor variables include:

Retrieved Wingbeat Frequency: This key metric provides insights into the oscillation frequency of wing motion.
Wing and Body Optical Cross Section: These metrics quantify the cross-sectional dimensions of both wings and bodies.
Mel Frequency Cepstral Coefficients (MFCC): A set of 13 coefficients representing the spectral characteristics of the signal.
Background Level: This variable gauges the ambient noise level during signal acquisition.
Transit Duration: The temporal span of each individual transit signal.

The data collected during the semi-controlled experiment (Aim 1) will constitute the basis of the training dataset, used by the machine learning classifier. This training dataset is composed of similar transit signals (.csv file) that are labeled (i.e. the insect from which the transit signal has been obtained is known).

The project's data management strategy ensures the systematic collection, storage, and accessibility of generated data, adhering to best practices for data preservation and sharing. The following outline delineates the comprehensive data management workflow:

1. Data Acquisition and Initial Storage: Data is initially created and recorded on the hard drive of the designated field acquisition computer. To safeguard against loss and ensure long-term storage, data is subsequently duplicated onto an external hard drive, serving as the primary local storage repository. This dual-storage approach provides immediate access to data while mitigating potential hardware failures.

2. Data Transfer and Verification: The recorded data is then transferred to a laboratory computer for subsequent analysis. A meticulous verification process is undertaken to confirm the accurate transfer of data. Upon successful verification, the data stored on the field acquisition computer is deleted to free up resources.

3. Backup and Accessibility: Both raw data and secondary data derived from analysis are duplicated onto Dropbox, facilitating secure backup and seamless accessibility for all members of the research team. This cloud-based backup ensures redundancy and collaborative access, supporting ongoing research efforts.

4. Long-Term Storage and Public Access: To ensure comprehensive preservation and public accessibility, all data is deposited in the "Open Science Framework" data repository. This repository is selected due to its robust

infrastructure and commitment to open access. Data is organized, annotated, and curated for long-term storage, enabling its use by the broader scientific community.

5. Storage Duration: External hard drives containing data are maintained indefinitely within the Principal Investigator's laboratory, ensuring that a physical copy is readily available for reference and verification. Similarly, data stored on Dropbox and the Open Science Framework are intended to be preserved indefinitely.

6. Data Size Estimation: The eBoss instrument generates approximately 0.5 Gbyte of data daily. Over the course of the project, two instruments will be employed for 9 months in each of year 2, 3, and 4. Additionally, the semi-controlled experiment is anticipated to run for a maximum of 30 days. Given the projected experiments, the project's data generation is estimated to amount to approximately 1 Tbyte of data throughout the project's duration.

All collected data will be stored in the Open Science Framework data repository. This includes the raw data, all transit signals as well as the training dataset. This platform ensures long-term preservation and public access to the data. We will organize the data in a structured manner, provide detailed metadata, and encourage proper citation through OSF's built-in citation mechanism. Raw data as well as all .csv files generated for each transit signals will be made fully available to the public, without any restriction.

In all publications resulting from this research, we will include a data availability statement indicating where and how the data can be accessed, fostering transparency and encouraging reproducibility of findings.

The PI Benjamin Thomas will be responsible for the implementation of the DMP. Co-PI Gareth Russell will take this role if needed. The funds needed for this DMP are low and included in the proposed budget.

Planned Research Outputs

Dataset - "Training Datasret"

The training dataset is obtained from the analysis of the data recorded during the semi-controlled experiment (Aim 1). It is used for training of the machine learning classifier algorithm (SVM)

Dataset - "Raw data and Transit signals"

All raw data recorded during the two field experiments (Aim 2) and transit signal files (.csv) obtained from the analysis of the raw data.

Text - "Articles"

Publications in peer-reviewed journals of the scientific findings.

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
Training Datasret	Dataset	2025-01-30	Open	Open Science Framework Dropbox	1 GB	Creative Commons Attribution Non Commercial No Derivatives 4.0 International	Dublin Core	No	No
Raw data and Transit signals	Dataset	2025-02-02	Open	Open Science Framework Dropbox	1 TB	Creative Commons Attribution Non Commercial No Derivatives 4.0 International	None specified	No	No
Articles	Text	2026-01-16	Open	None specified	10 MB	None specified	None specified	No	No