

Data Science Project Scoping Worksheet

Team Information	2
Team Name:	2
Names of team members:	2
Student number of team members:	2
Signatures & Date:	2
Project Information	3
Project Name:	3
Project Lead:	3
Project Lead Signature & Date:	3
Organisation Name:	3
Project Description:	3
Who are the agencies/departments that will need to be involved?	4
Who are the individuals in these organisations that are stakeholders? What is their role?	4
Project Goals	5
Project Actions	6
Data Information	8
A. Internal Data	8
B. External Data	10
C. Additional Data Requirements	11
Analysis Goals and Needs	12

Team Information

Team Name:

Significant Outliers

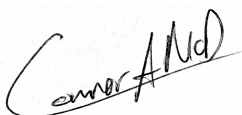

Names of team members:

- Connor McDonald
- Gené Fourie

Student number of team members:

- Connor McDonald - u16040725
- Gené Fourie - u20797274

Signatures & Date:

Name	Signature	Date
Connor McDonald		12-03-2022
Gené Fourie		12-03-2022

Project Information



Project Name:

Enabling weather-based decision making for forestry pest and disease management

Project Lead:

Prof Bernard Slippers and Dr Ilaria Germishuizen

Project Lead Signature & Date:

Name	Signature	Date
Prof Bernard Slippers		21/03/2022
Dr Ilaria Germishuizen		18/03/2022

Organisation Name:

Forestry and Agricultural Biotechnology Institute (FABI)

Project Description:

The biggest threats to South African Forestry and Agricultural sectors are climate change and invasive pests. The aim of this project is to work closely with Forestry and Agriculture Biotechnology Institute (FABI) to optimally integrate their data sources, specifically focusing on temperature, rainfall and pest prevalence. This is to be done with the development of a user-friendly web-based platform to ensure easy access by industry partners. This tool is to be part of the existing Information Hub. The application will be able to perform basic analysis and visualisation, and allow for some degree of scenario modelling.

Insects /pests are directly dependent on external weather for physiological activity. By accurately understanding how weather fluctuations correlate to pest distribution, we can cater our response to changing climatic patterns in support of management strategies to reduce the spread of pests across South Africa. This project will focus on two of the most common insect pests in South Africa, namely *Sirex noctilio* and *Leptocybe invasa*.

Modified from University of Chicago Center of Data Science & Public Policy Scoping Sheet.

Who are the agencies/departments that will need to be involved?

- Forestry and Agriculture Biotechnology Institute (FABI)
- Institute of Commercial Forestry Research (ICFR)
- Advance.io

Who are the individuals in these organisations that are stakeholders? What is their role?

Institute	Name	Role
FABI (Forestry and Agricultural Biotechnology Institute)	Prof. Bernard Slippers	Project Coordinator
FABI (Forestry and Agricultural Biotechnology Institute)	Hannes Strydom	Project Liaison
ICFR (Institute for Forestry Research)	Prof. Ilaria Germishuizen	Research Adviser
Advance.io	Arné Schreuder	Technical Consultant

Project Goals

<p>Goal 1:</p> <p>Integration and standardisation of weather and pest datasets.</p>	<p>Goal 2:</p> <p>Analysis of weather and pest trends over time. These analyses will aim to detect any correlations between pest prevalence and climatic patterns.</p>	<p>Goal 3:</p> <p>Web-based user interface development for analysis, visualisation and scenario modelling.</p>
<p>Constraints:</p> <ol style="list-style-type: none"> 1. All three datasets are in a different structure. 2. All three datasets cover a different time frame. 3. The weather dataset is split between a Google BigQuery database and textfiles (for more recent years). 	<p>Constraints:</p> <ol style="list-style-type: none"> 1. Pest data is recorded annually compared to the weather data that is recorded daily. The difference in granularity limits detailed seasonal inference. 2. Any models, or code written in this section needs to be relatively lightweight so that scenario modelling is still a viable feature on the final platform. 3. Route most of the processing through Google BigQuery to benefit from the processing power available. 4. Visualisations need to allow for interrogation by the user e.g. zooming in, filtering, grouping. 	<p>Constraints:</p> <ol style="list-style-type: none"> 1. User interface to form part of current Information Hub platform, not a stand-alone user interface. 2. This will require a database to store meta-data and analysis data, this will need to be hosted on the current Information Hub Server. 3. App needs to be easy to update and modify. Therefore, clean and understandable code is imperative.

Modified from University of Chicago Center of Data Science & Public Policy Scoping Sheet.

Project Actions

Action 1: Download of the pest and weather data from Information Hub	Action 2: Structured SQL database design and creation of tables on Information Hub	Action 3: Upload of data to restructured tables on Information Hub
Questions	Questions	Questions
Who is taking the action? Connor McDonald	Who is taking the action? Gené Fourie	Who is taking the action? Connor McDonald
What/Who is it being taken on? Pest and weather data	What/Who is it being taken on? Pest and weather data to be structured into a relational database schema	What/Who is it being taken on? Pest and weather data to be loaded into a relational database schema
How often is the decision to take this action made? Once	How often is the decision to take this action made? Initial design to be updated when loading and analysing data	How often is the decision to take this action made? Once
What channels are/can be used to take this action (in person, digital, etc.) Digital	What channels are/can be used to take this action (in person, digital, etc.) Digital	What channels are/can be used to take this action (in person, digital, etc.) Digital
Other questions about the action Direct access to the database is preferred for data extraction, however, data can be exported from the Information Hub viewer in chunks. sections.	Other questions about the action A SQL database will be created within a project on Information Hub for design purposes. Updates to the structure will require liaison with developers of Information Hub when the solution is implemented.	Other questions about the action None

Action 4: Connection to Information Hub/Big Query through spatial platform, spatial analyses and spatial visualisations	Action 5: Exploratory analyses and machine-learning algorithms on weather and pest data	Action 6: Integration of findings and approaches into the Information Hub, visible through a user interface
Questions	Questions	Questions
Who is taking the action? Gené Fourie	Who is taking the action? Connor McDonald and Gené Fourie	Who is taking the action? Connor McDonald and Gené Fourie
What/Who is it being taken on? Pest and weather data on the Information Hub	What/Who is it being taken on? Pest and weather data on the Information Hub. Different methods and approaches will be considered	What/Who is it being taken on? Results of the analyses in Action 4, and Action 5
How often is the decision to take this action made? As-and-when required throughout project duration	How often is the decision to take this action made? As-and-when required throughout project duration	How often is the decision to take this action made? Once, after prior analyses are performed
What channels are/can be used to take this action (in person, digital, etc.) Digital	What channels are/can be used to take this action (in person, digital, etc.) Digital	What channels are/can be used to take this action (in person, digital, etc.) Digital
Other questions about the action None	Other questions about the action Findings and analyses to be compared to the second MIT808 team working on the FABI project to ensure minimal duplication of work.	Other questions about the action The user interface is to display the findings of the analysis and must allow for upload of new pest and weather data into database structure.

Data Information

A. Internal Data

Data Source Leptocybe pest dataset (2016-2021)	Data Source Sirex pest dataset (2012-2019)	Data Source 1950-2019 weather dataset
What does it contain? Year, sample size of number of trees surveyed, tree species, GPS location of tree compartment, and frequency and location of leptocybe on trees surveyed.	What does it contain? Year, region, company, plantation, plantation compartment, age of trees, tree species, predefined risk of infection, GPS location of compartment, total trees (stems) in compartment, living/dead/dying status of trees, collector username, and binary classifier of the prevalence of Sirex.	What does it contain? Rainfall (mm), and maximum and minimum temperatures per weather station. Between 1950 and 2010, the data is recorded for approximately 6000 weather stations. Over the past 15 years, the South African Weather Services has started closing down weather stations and charging a rate for the data. As a result, the data for latter years originates from approximately 700 weather stations.
What level of granularity? Annual, at a plantation compartment level	What level of granularity? Annual, at a plantation compartment level	What level of granularity? Daily, per weather station
How frequently is it collected/updated after it's captured? The data is collected, updated and captured on a yearly basis between 2016 and 2020.	How frequently is it collected/updated? The data is collected, updated and captured on a yearly basis between 2012 and 2019.	How frequently is it collected/updated? The data is collected, updated and captured on a daily basis.
Does it have unique identifiers that can be linked to other data sources? No, however, the data has a spatial attribute which allows	Does it have unique identifiers that can be linked to other data sources? No, however, the data has a spatial attribute which allows	Does it have unique identifiers that can be linked to other data sources? The weather data matches to the weathers stations

Modified from University of Chicago Center of Data Science & Public Policy Scoping Sheet.

University of Pretoria MIT 808 Scoping Sheet

for spatial link to other data sources.	for spatial link to other data sources.	through the clim_no attribute.
Who's the internal owner of the data? The ICFR is the custodian of the dataset that is available through FABI.	Who's the internal owner of the data? The ICFR is the custodian of the dataset that is available through FABI.	Who's the internal owner of the data? The ICFR is the custodian of the dataset that is available through FABI.
How is it stored? Google BigQuery	How is it stored? Google BigQuery	How is it stored? Google BigQuery
<p>Other</p> <ul style="list-style-type: none"> • The pest data spans across departmental companies (SAFCOL), privately owned companies (Sappi, Mondi), medium growers (farmers). • The pest data is predominantly for the summer rainfall region of South Africa, due to the location of the plantations. • The Leptocybe dataset for 2021 follows an alternate structure to the dataset for 2016 to 2020. 		<p>Other</p> <ul style="list-style-type: none"> • Weather data is included for stations that fall outside as well as inside the forestry areas. • Previous students have worked with the weather dataset and have assisted in cleansing the data.
<p>Other</p> <ul style="list-style-type: none"> • FABI has made the data available through the Information Hub, a cloud-based platform. 		

B. External Data

Data Source Shapefiles of Safcol (ex Komatiland) Forest landuse and plantations	Data Source Demarcation shapefiles	Data Source Contour shapefiles
What does it contain? Spatial representation of plantation compartments	What does it contain? Spatial representation of South African boundaries	What does it contain? Contours across South Africa
What level of granularity? Per plantation compartment	What level of granularity? Demarcation of South Africa per province, district/local municipality and ward	What level of granularity? 10m to 20m intervals
How frequently is it collected/updated after it's captured? Once	How frequently is it collected/updated? As determined by the Municipal Demarcation Board	How frequently is it collected/updated? Once (for this specific dataset)
Does it have unique identifiers that can be linked to other data sources? Yes, compartment indicators	Does it have unique identifiers that can be linked to other data sources? The spatial attribute is of importance within the data preparation stage. Unique identifiers are not required.	Does it have unique identifiers that can be linked to other data sources? The spatial and elevation attribute is of importance within the data preparation stage. Unique identifiers are not required.
Who's the internal owner of the data? Gené Fourie. Data was sourced for a previous project of student.	Who's the internal owner of the data? Gené Fourie. Data was sourced from the Municipal Demarcation Board public website.	Who's the internal owner of the data? Gené Fourie. Data was sourced for a previous project of student.
How is it stored? .shp, .shx and .dbf files`	How is it stored? .shp, .shx and .dbf files`	How is it stored? .shp, .shx and .dbf files`
Other Permission is required from SAFCOL if the shapefile is to be published with results of the project.	Other -	Other -

Modified from University of Chicago Center of Data Science & Public Policy Scoping Sheet.

C. Additional Data Requirements

Weather and pest data is confidential and are for the sole purpose of use in the Information Hub platform and for this project only.

Analysis Goals and Needs

<p>Analysis 1:</p> <p>Exploratory Data Analysis on the provided data</p>	<p>Analysis 2:</p> <p>Detect seasonal and regional fluctuations in pest data through</p>	<p>Analysis 3:</p> <p>Prediction of pest prevalence location based on weather conditions</p>
<p>Analysis type:</p> <p>Descriptive</p>	<p>Analysis type:</p> <p>Detection</p>	<p>Analysis type:</p> <p>Prediction</p>
<p>Which action will this analysis inform?</p> <p>Action 5 from the above Action list</p>	<p>Which action will this analysis inform?</p> <p>Action 5 from the above Action list</p>	<p>Which action will this analysis inform?</p> <p>Action 5 from the above Action list</p>
<p>How will you validate this analysis?</p> <p>Validation by ICFR/FABI to confirm status quo.</p>	<p>How will you validate this analysis?</p> <p>Graphical visualisations to visually inspect the past correlations identified</p>	<p>How will you validate this analysis?</p> <p>Proposal to ICFR/FABI on where to conduct further tests on pest data (in the field) when weather conditions meet certain criteria.</p>