Enabling weather-based decision making for forestry pest and disease management: An Exploratory Data Analysis

Connor McDonald
University of Pretoria
Department of Computer Science
u16040725@tuks.co.za

Gené Fourie
University of Pretoria
Department of Computer Science
u20797274@tuks.co.za

Keywords

Sirex noctilio, Leptocybe invasa, Forestry pest and disease management

1. DETAIL ANALYSIS

This report focuses on analysing weather and pest-prevalence data in South Africa to identify possible relationships between changing climatic conditions and pest populations. Three predominant datasets were received from the Forestry and Agricultural Biotechnology Institute (FABI) for contribution to this project, namely:

1. Temperature and rainfall records of roughly 6000 forestry plantations across South Africa.
2. Sirex noctilio (Sirex) pest inspection samples
3. Leptocybe invasa (Leptocybe) pest inspection samples

1.1 Weather Dataset

This data consists of approximately 107 million daily temperature and rainfall readings from 6137 weather stations scattered across South Africa. The data were collected over a 70 year timespan between 1950 and mid-2019. There are three types of stations namely: RAIN, TEMP and TEMP_RAIN, which indicate the type of data collected by the station. The proportion in which the stations were distributed was relatively even until 2001, when government funding was reduced. This led to a drastic decline in the number of stations as well as a shift in the proportion of station types as seen in Figures 1 and 2.

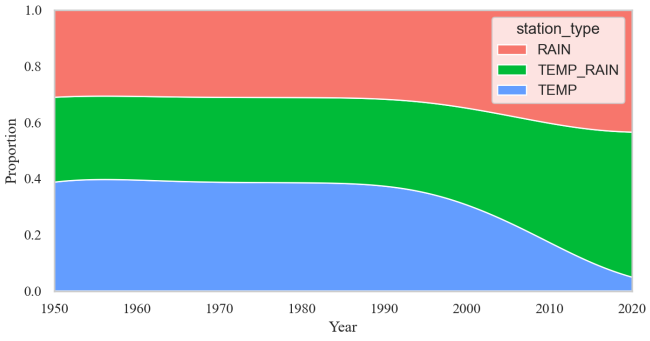

Figure 1: Station Type Distribution

This decrease in stations happens to coincide with a period of missing data for both maximum and minimum daily temperatures between 2001 and 2004 shown in Figure 3, which resulted in null value temperatures for 2.89% of the total temperature data.

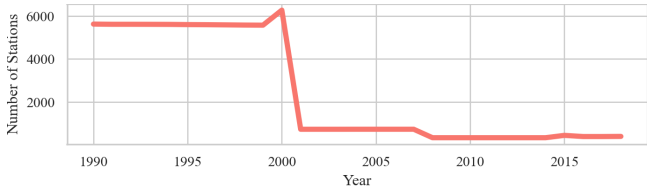

Figure 2: Number of Stations per Year

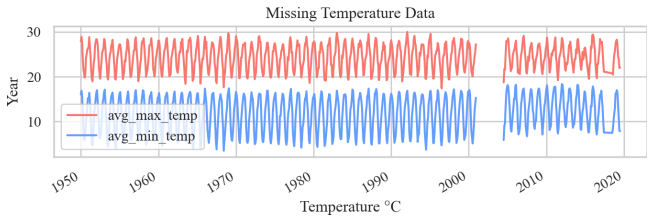

Figure 3: Average Monthly Temperature

The rainfall data only had a small percentage of null values (approximately 0.08%), however, a faulty station was detected due to its annual rainfall exceeding 150000mm for seven consecutive years. Considering that the highest average rainfall for any region in South Africa is around 3000mm annually [1], this station was subsequently filtered out of the dataset as it drastically impacted any statistical metrics that were derived from the data.

The remaining records were visualised with box plots in Figure 4 to highlight the distribution of annual rainfall readings per decade. There is consistency of the readings per decade in terms of magnitude, however, changes in the median and interquartile ranges indicate movement in the data and potential underlying trends in annual rainfall patterns.

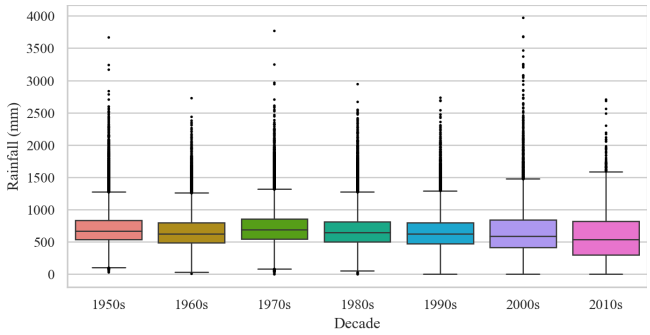

Figure 4: Annual Rainfall Per Station Grouped by Decade

1.2 Sirex inspection samples

The Sirex dataset includes 3780 inspections taken across five South African provinces. An average of 473 inspections are recorded per year, with approximately 24% of inspections indicating a positive pest finding. The attributes of note include: location, pest presence, severity, and affected tree species.

1.2.1 GPS inspection location

The dataset provides approximate centroid coordinates of plantation compartments in which inspections were performed. Several errors - including interchanged latitude and longitude coordinates, null coordinates and GPS inaccuracies - are seen in the data, however, inference with like site numbers and compartments can be performed to correct faulty coordinates.

1.2.2 Sirex presence binary indicator

The binary indicator represents the outcome of an inspection, with 920 inspections indicating a positive Sirex presence. Figure 5 displays the spatial distribution and inspection results.

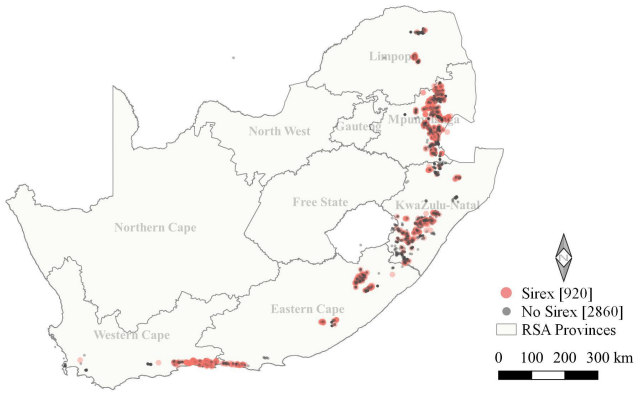

Figure 5: Sirex inspection distribution

1.2.3 Condition of trees and pest severity

Table 1 indicates the condition of the trees inspected. The data indicates that only the samples with a Sirex positive result have trees that have been declared as dead or dying. However, the high percentage of unclassified stems provide uncertainty in the analysis of living versus dead or dying stems due to the Sirex pest. Consequently, the severity of the Sirex pest is poorly represented by the condition of trees within the sample inspections.

Table 1: Sirex condition status

Presence	Living	Dead/dying	Unclassified	Total stems
No Sirex	88.7%	0.0%	11.3%	**100%**
Sirex	88.1%	3.5%	8.5%	**100%**

1.2.4 Tree species and Sirex prevalence

Figure 6 provides the Sirex prevalence per pine tree species. The three predominant species are: P PAT (*Pinus patula*), P RAD (*Pinus radiata*) and P ELL (*Pinus elliottii*), with the P PAT and P RAD species having a higher positive Sirex prevalence.

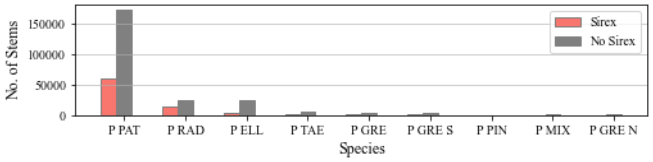

Figure 6: Sirex prevalence per tree species

1.3 Leptocybe inspection samples

The Leptocybe dataset includes 747 inspections, concentrated in four South African provinces. Inspections vary largely per year, from a minimum of 46 inspections in 2017, to a maximum of 187 inspections in 2019. Approximately 40% of inspections indicate a positive finding of the pest.

The dataset attributes are similar to the Sirex data, however, GPS coordinates are more consistent, and pest severity more detailed.

1.3.1 GPS inspection location and pest presence

Figure 7 displays the distribution of Leptocybe inspections, with 314 inspections indicating a positive Leptocybe presence. The number of inspections are fewer than the Sirex inspections, however, inspections overlap in multiple eastern plantations.

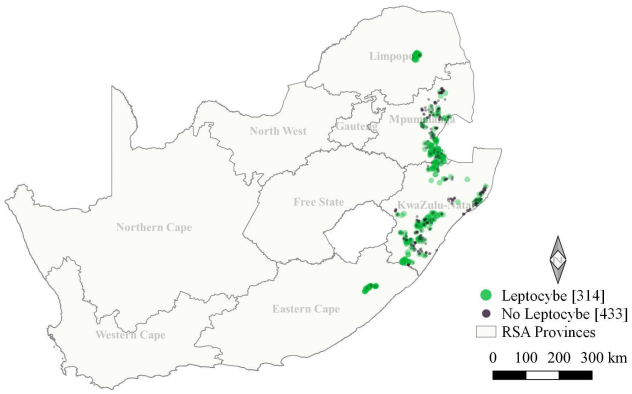

Figure 7: Leptocybe inspection distribution

1.3.2 Leptocybe intensity

The number of trees infected by the pest and the total number of trees inspected per inspection is used to determine the intensity of the pest infestation. Figure 8 indicates that infestation levels are highest in 2017, which follows a higher rainfall season in 2016. Therefore, there is preliminary evidence to suggest that pest intensity is dependent on rainfall patterns.

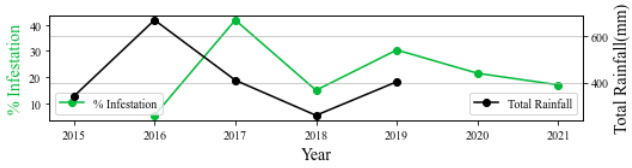

Figure 8: Leptocybe intensity vs total rainfall

1.3.3 Tree species and Leptocybe prevalence

Figure 9 provides the Leptocybe prevalence per Eucalyptus (E.) species. Through considering the four predominant species, Leptocybe is most prevalent in EGXN (*E. grandis x E. nitens*) and EGRA (*E. grandis*), and least in EGXU (*E. grandis x E. urophylla hybrid*) and EDUN (*E. dunnii*), suggesting that certain species are more prone to pest infestation than others.

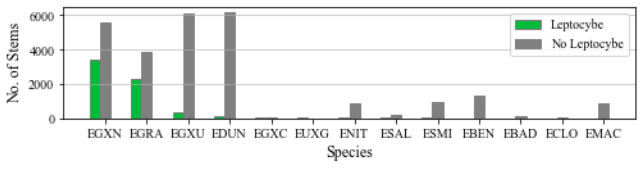

Figure 9: Leptocybe prevalence per tree species

2. DISCUSSION

The first evident problem uncovered by this exploratory data analysis is that the pest inspection data (recorded annually) does not have the same level of granularity as the weather data (recorded daily). This complicates the process of identifying relationships between weather conditions and pest prevalence due to the absence of seasonal differences in pest populations. However, the preliminary analysis outlined in Section 1.3.3 above indicates that possible trends may be identified (albeit the difference in granularity) through considering, for example, a higher rainfall period prior to an increase in pest intensity in a subsequent year.

On top of this, the number of pest inspections varies per year, this makes absolute metrics, such as total pest presence, invalid for year on year comparison. To address this issue, one can make use of relative metrics such as percentage pest presence. Fortunately, weather stations are thoroughly distributed, and cover all regions where pests are recorded. The proximity of pest inspections to weather stations is shown in Figures 10 and 11, which highlight that most pest inspections are within 10 km of a weather station.

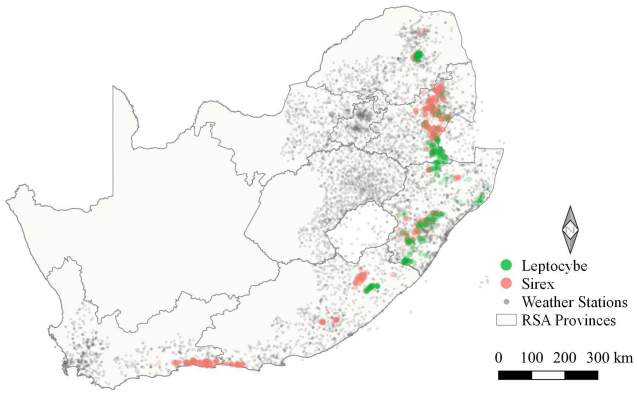

Figure 10: Weather Station Distribution vs. Pest Distribution

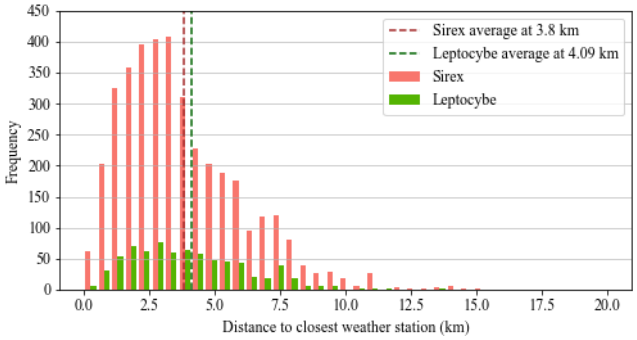

Figure 11: Weather Station Distribution vs. Pest Distribution

3. OTHER SCOPING MATTER

An efficient machine learning model requires a linking component between the independent variable and the dependent variable. In this research, the linkage exists between the spatial location (specifically, geographic coordinates) of weather stations and the location of pest sightings. However, an apparent problem is that the coordinates of weather stations do not align with the coordinates of pest sightings. To address this issue, one needs to assign pest sightings to their nearest weather station in a computationally efficient manner.

This process can be performed with a concept from computational geometry known as *Voronoi Diagrams*. These diagrams partition a plane with *n* points into convex polygons, such that each polygon contains exactly one generating point and every point in a given polygon is closer to its generating point than to any other [2]. The application of Voronoi Diagrams to weather stations can ultimately assign the annual weather records for a given station to all the pest sightings that fall within that station's polygon (Figure 12).

However, the Voronoi Diagram analysis must be cognisant of the type and period of weather data available per station, due to the variability of data mentioned in Section 1.1. This consideration may call for layering of Voronoi Diagrams per station type and per time period, or a consolidation of several nearby stations, to allow for improved representation of weather conditions at pest sampling sites.

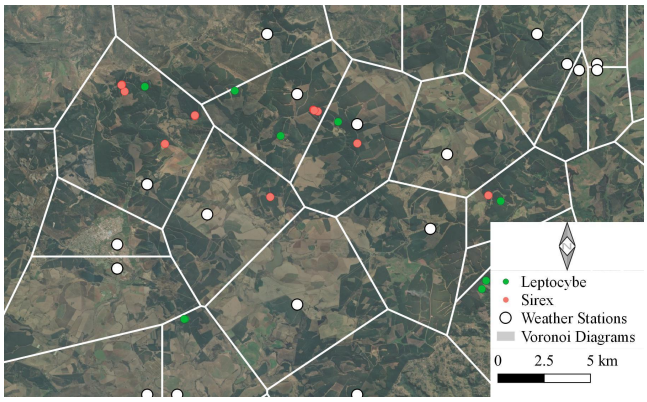

Figure 12: Voronoi Diagram

4. REFERENCES

[1] Rainfall in South Africa. Climate and Farming in South Africa. Water Research Commission. . https://southafrica.co.za/rainfall-south-africa.html

[2] Weisstein, Eric W. "Voronoi Diagram." From MathWorld--A Wolfram Web Resource. https://mathworld.wolfram.com/VoronoiDiagram.html