Deriving Data Stories from the GAEA Geospatial Tool

Joost ter Braake

BSc. Creative Technology Thesis

Feb — Jul 2024

Supervisor, Client Representative — dr. A. Kamilaris

Critical Observer — dr.ir. J. Klein Brinke

Abstract

This thesis describes how data from an Environmental Digital Twin of Cyprus (GAEA) can be turned into an insightful data story. By using an explorative method typical in data journalism, a subset of geospatial data was selected for study. These data were preprocessed via Python and used in QGIS to spatially align AI-derived Tree Classification data with the underlying soil, aspect, slope, and elevation.

By achieving alignment between multiple data layers they could each be joined to the Tree Class data and prepared for further study. After obtaining descriptive statistics, various tests including Logistic Regression and Chi-2 were performed in order to identify and validate patterns in the data for the data story.

These analyses showed highly significant result across all studied correlations, indicating that all classes of trees tended towards very distinct enviornmental preferences. In comparing the two types of pines prevalent on Cyprus: Brutia Pine (Pinus brutia) and Black Pine (Pinus nigra), the data story emphasizes their stark differences and points to potential implications for forestry.

Acknowledgements

This research project was suggested by dr. Andreas Kamilaris, who served as supervisor and represents the client: the CYENS Research Centre on Interactive Media Pervasive Group, University of Twente

I would like to thank dr. Andreas Kamilaris for being an enthusiastically supportive supervisor. The feedback obtained from our discussions was always on point, I felt our meetings were very productive and engaging, which helped a lot. I would also like to thank dr.ir. J. Klein Brinke for his contribution during presentations. I should also like to thank Asfa Jamil from Cyens for providing the data that made this project possible, and for patiently answering my questions on how to properly align these data.

I would be remiss if I did not thank my great friends and family for rubber ducking for me on many an occasion, and for regularly checking in on me.

Table of Contents

List of Figures

Chapter 1 — Introduction

These past years have seen rapid advances in the fields of Artificial Intelligence and Big Data, and a newfound ability to analyse large swaths of it with machine learning [1], [2]. This bridging of the gap between a large availability of data with the prospect of training AI agents to analyse it for one's purposes can have far reaching consequences in many domains of research.

One such domain is that of *Environmental Digital Twins*. These are systems and tools based on geospatial data which aim to produce models of environments across many data dimensions to grant insight into the workings of the environment in question. Since the modeling of any complex system requires accurate input, and the modeling of a real environment requires a large amount of real data about that environment, AI methods may be used in the classification and analysis of environmental big data [1], [2].

Remote Sensing of the Earth's surface, especially by means of satellites, has been a prolific source of environmental big data. The modern analysis thereof, especially by means of machine learning, accurate modeling, and the layering of data dimensions, has lead to the creation of Geosptatial Tools which may be described as EDTs. GAEA, a tool developed by Jamil et al. which aims to be an EDT of Cyprus, is one of the earliest examples of such a tool [2], and is the impetus behind this thesis. Since an EDT can provide many different dimensions of geospatial data, there are likely latent insights in GAEA's particular aggregation of data, should it be novel. In consultation with the client representative working on the GAEA EDT it was determined that GAEA's data should be investigated, and that if interesting correlations could be found that these should be communicated to a general audience via a data story, as a means of combining data communication and visualisation.

In order to produce a data story based on this EDT, the capabilities of GAEA and its peers must first be understood, and found to be novel. Further, if an interesting combination of data can be found, an approach to narrativising the data is needed. To facilitate these tasks, the following research questions were considered:

RQ 1. *"How can GAEA be used in the making of interesting and insightful data stories on the physical characteristics of Cyprus?"*

RQ 2. *"How can such data stories be feasibly identified, produced, and made suitable for the appropriate audience?"*

As such, this review shall be split in two parts. First, the potential of GAEA as an EDT must be understood, and thereafter theory on the proper production of data stories may be covered to facilitate communicating to a general audience.

Chapter 2 — Background Research

2.1 Context

In keeping with the research questions, the contents of this review are informed by the twin needs of understanding the fields in which GAEA is situated and the tools available therein, as well as finding out how GAEA may be used in the making of interesting and insightful data stories on the physical characteristics of Cyprus.

In order to produce a Data Story based on GAEA and other geospatial tools, its capabilities and those of its peers must first be understood. With an understanding of what kinds of data and tools are out there, and what dimensions of data these tools and GAEA offer, we may get a picture of what we can work with in producing narratives from these data and with these tools, and finally how we may identify these data stories, and present them to an interested audience in a compelling way.

When considering GAEA's position in the field of geospatial tools and environmental digital twins, comparisons may be made by considering the type of data and methods of analysis used, the area covered, the environmental services offered, as well as the ability to generate data visualizations. Particular interest must be paid to challenges faced in this field, so that roadblocks may be avoided and gaps in current knowledge may be found.

In looking to produce a good data story, it is important to to consider the state of the art in techniques for data visualization and data storytelling. We need theory to structure the search for potential stories, and to later make judgements about what approaches and elements are desirable, and which are not. Further, examples produced with typical modern tools can give some inspiration for what is possible, and an overview of available data can support novelty.

2.2 GAEA & other Enviornmental Digital Twins

Digital Twins need to accurately represent what they are modeling, and so their scope is often limited. As GAEA aims to be a digital twin of the whole island of Cyprus, its services cover the entire breadth of the island at some 9,000 km2, and aims to do so at a granular resolution. GAEA's services cover the categories of Land Use Monitoring, Geohazards, Proximity to locations, Climate Monitoring, and Geomorphological Characteristics. Data from various sources is combined to produce these services, including from the Land registry of Cyprus, weather stations, EU soil data, and satellite imagery up to a resolution of 50 centimeters [2].

This high resolution remote sensing data may be used with machine learning to perform tasks such as the counting of the number of trees in a given area, and these types of output may be used in internal models to make predictions, such as the risk for wildfires. An overview of all the available dimensions is given in figure 1 [2]. The breadth of geospatial data dimensions covered by GAEA, together with the high resolution of critical data lends credence to the claim that GAEA constitutes an environmental digital twin

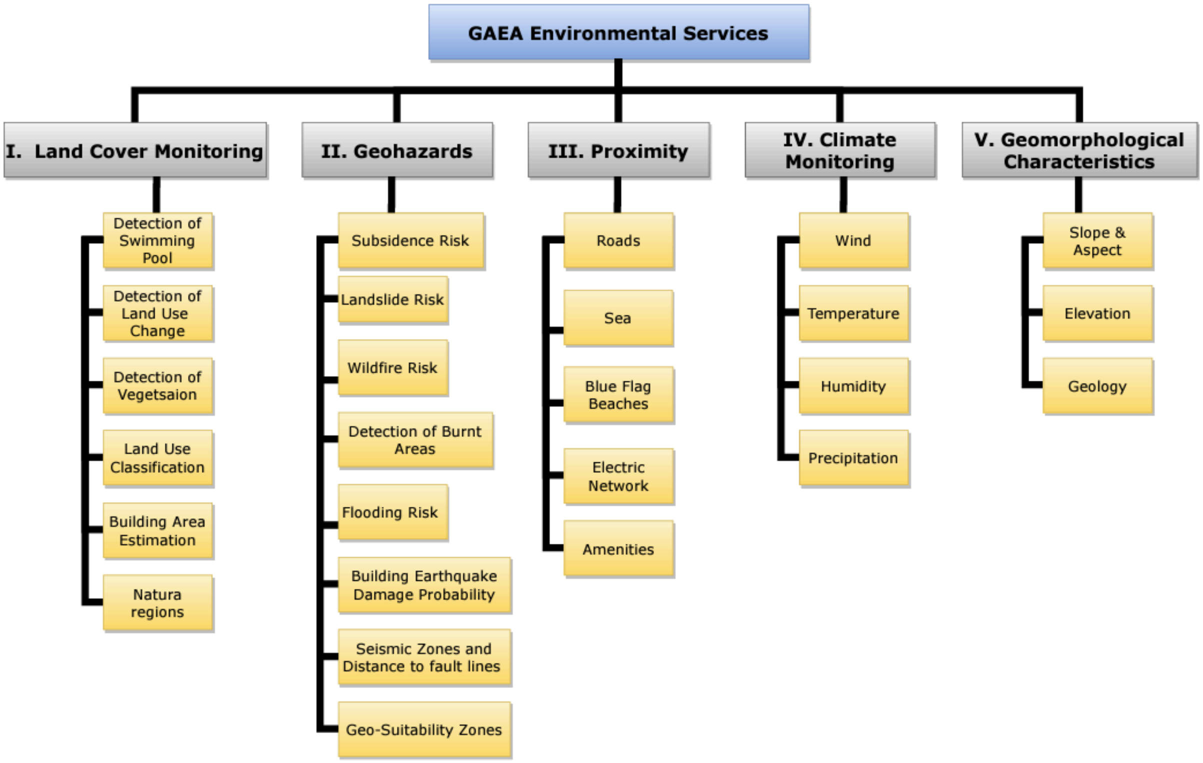


Figure 1: GAEA's services / geospatial data dimensions — via Jamil et al. [2]

The combination of the kind of geospatial data available in Geographic Information Systems (GIS) with machine learning to produce models of risk has earlier been applied by Palaiologou et al. to simulate wildfires and produce risk assessments thereon [3]. Further, tools such as Earth Map as presented by Morales et al. can show this kind of output, or past burn evidence data, as a map overlaying selected regions across the globe [4]. By combining risk assessments with clear to read visulizations on maps a user of a geospatial tool may inform themselves about risks they care to avoid, and a researcher or producer of such visualizations may leverage this method to inform those who might be affected.

Both GAEA's and Palaiologou et al.'s wildfire risk assessment tools offer georisk information based on machine learning models, but differ in intended mode of presentation.

Instead of aiming to produce a national registry, resulting in a static map, GAEA offers a web-based dashboard service for user accessibility. This approach to geospatial data presentation is not only seen in Earth Map [4], but more closely in Zhang et al.'s "City appearance environment management system", which aims to make geospatial data readily and clearly available to support the governance of Changchun [5]. There is a congruence in the method by which the user interacts with the tool and the type of output produced between Zhang et al.'s dashboard tool and GAEA's own web-application. GAEA does not however offer the type of map-overlay layer as seen in these other tools [4], [5], while it does possess the necessary type of data to visualize dimensions such as wildfire risk in this manner [2], [3], which points to a potential opportunity for data visualizations to be produced based on these outputs.

All these tools have in common that in order to process data and arrive at any visualizations, they must work with GIS tools. Yang et al. contend that Geospatial big data, such as remote sensingis challenging to work with, especially for conventional GIS systems [6]. They argue this is to a great extent caused by the sheer amount of data, and the high resolution of the data, which besides the difficulties in processing also gets in the way of AI training. Verma et al. see these challenges also, and offer an "integrated" tool for managing these data in the context of machine learning projects [7].

And yet conventional GIS tools are varied and essential tools for processing remote sensing data. Macarringue et al. provide an overview of the various approaches to land use/change data and remote sensing generally, and compare the functionality of different GIS options for these uses [1]. From their tabular comparison especially it may be noted that among free, open source GIS tools capable of using and manipulating remote sensing data, QGIS stands out for the combination of its features, usability, and possible integration with languages such as Pyton.

Besides the types of data and how they are processed the granularity is also of interest. Whilst Morales et al. with Earth Map [4] make an impressive set of open data covering the globe available easily to users, and visualize this data well on map layers, much of the data available is not of great resolution. Ignatius et al. in their highly detailed simulation of the climate of Singapore and how it is affected by buildings and energy infrastructure show that given a high enough resolution of data and a sophisticated model, useful information may be provided to experts [8]. This implies a gap between the local and global scale which GAEA's data fills, even as it also models aspects of buildings. GAEA's 50cm resolution image data combined with its ML models might prove to show a gap in interesting data which might interest both experts as well as laypeople.

2.3 Data Visualization and Story Techniques

The field of journalism has been one of the most prolific producers of data stories in recent years, as publications move from print to digital storytelling. Weber et al. provide a study based on structured interviews with journalists engaged with this form of storytelling with data to derive insights on approach and elements [9]. A sizable analysis by Segel and Heer identifies many of the same relevant features, and provides an extensive list of specific techniques as compared with the context in which they are most typically used [10].

One principle bifurcation in method lies in where to begin. To start with a story, and to then find data to support that story, or to start with a dataset and to find notable patterns in that set which then lead to a potential story [9]. The former approach allows one to have a good grasp of the intended narrative ahead of time, and to already identify any stakeholders to which this story might be relevant. The alternative, exploratory approach however is more typical for truth-finding, relevant to both research and journalism. As our starting point is a collection of geospatial data and geo-analysis, the exploratory approach detailed by interviewees in Weber et al.'s study may be more suitable to producing data stories from EDTs.

Commentary from these studies imply that data suitable for narrativization must reveal some inherently interesting pattern, and bring some novel insights to whom it is presented. Further, though many means of executing data visualisations and stories are covered, consistent emphasis is put on the need for aesthetic appeal, which is corroborated both by near unanimous agreement among Weber et al.'s interviewee's [9] as well as in the cases examined by Segel and Heer [10].

2.4 Literature Review Conclusion

The purpose of this review has been to lay a foundation of understanding of GAEA's data and systems, as well as how data should be approached in producing an effective data story. By seeking to understand GAEA in its peer context, the challenges of working with big data sets in training AI models and producing client-facing tools were especially highlighted. As GAEA's client cannot easily be used in producing data stories, a GIS tool should be utilized on extracted data layers to produce the visual foundation for a data story, utilizing the results of AI models where data accuracy is well validated. Furthermore, the data aggregates found in this review allow for novelty testing, as exported data layer may be compared for differences.

By combining these insights and this knowledge to data storytelling with an exploratory methodology, a data story may be produced by extracting GAEA's data as layers into a GIS, and applying the type of visualizations appropriate to those data. Whilst applying the exploratory method to GAEA's data one must consider patterns that result from these data, the novelty of the data, and the aesthetic potential afforded by the dimensions one is working with.

Big geospatial data has the potential to grant crucial insights into our environment, and EDTs may play a large role in accurately representing our world to us. At present however, it is hard to work with the tools that produce them, and those that exist cannot derive from themselves those data and narratives most pertinent to an average user. That this review seeks to aid in the manual production of this value is demonstrative of the present limitations of the field. From what point these tools can even be considered an EDT has no clear answer in the literature, a taxonomy of features and data quality requirements might bring more objectivity to this determination.

Likewise, though data storytelling has exploded as a field, few bridges exist between its approach and EDTs, a gap which this review has hoped to begin to address. The scope of this review ultimately centered around one EDT, and found that literature on the 'How?' questions directly was sparse. As the literature on data stories and the results of GIS use both show, practitioners of these tools and methods are currently leading the way. Until theory catches up to practice, what makes a data story interesting or insightful is part art, part science.

2.5 State of the Art

2.5.1 State of the Art: Data Visualization and Story Examples

Firstly, the client provided an example data story by Eftychiou under internship at CYENS, utilizing data dimensions now present in GAEA [11]: Natura 2000 areas * Land Use Change. By means of this map based visualization as seen in Figures 2, 3 and an accompanying textual narrative, the argument that recent construction work is occurring very nearby Natura 2000 areas can be made clear. An effective and commonly used technique with visualizations made on geospatial data (via GIS tools) employwed here is the introduction of subsequent layers. By first clarifying where Natura 2000 areas are present, and only then as the reader scrolls further progressing to showing LUC as a layer on top (whilst fading harsh borders for visual clarity), the added layer is automatically going to be considered in relation to the one previous. The argument is made visually and implicitly in support of the explicit textual argument that follows.

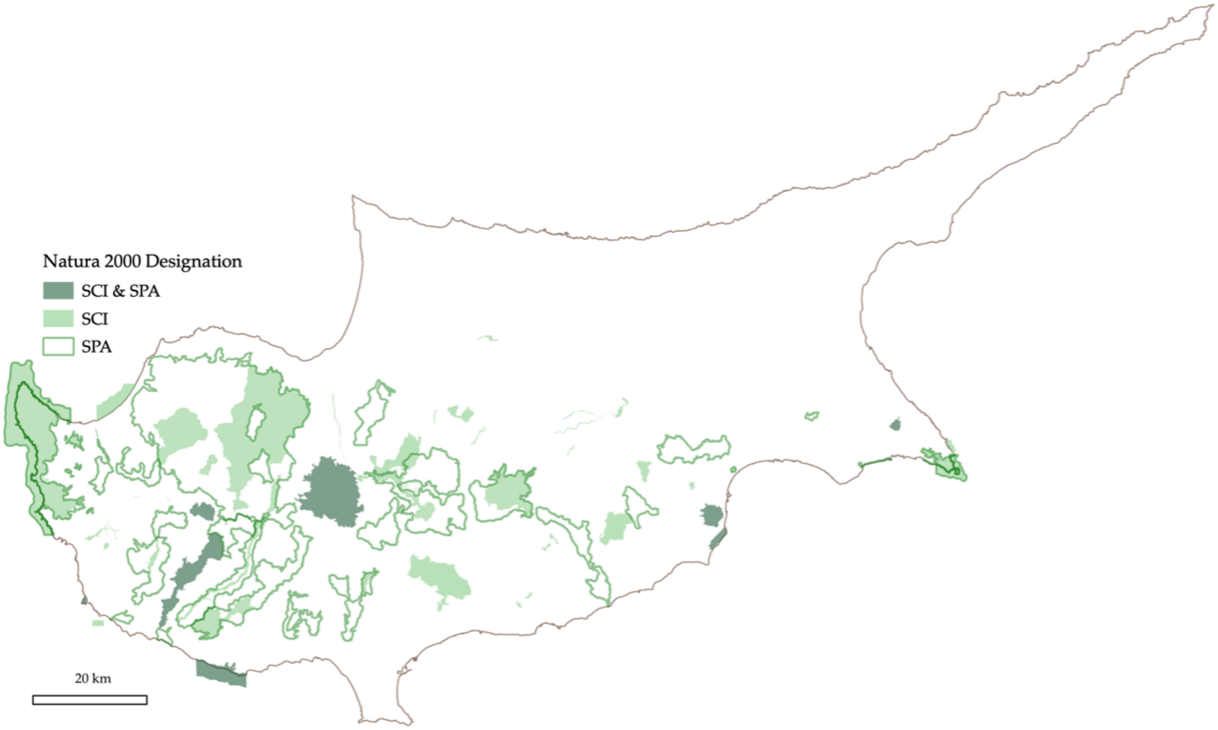


Figure 2: Map of Natura 2000 areas in Cyprus — via Eftychiou [11]

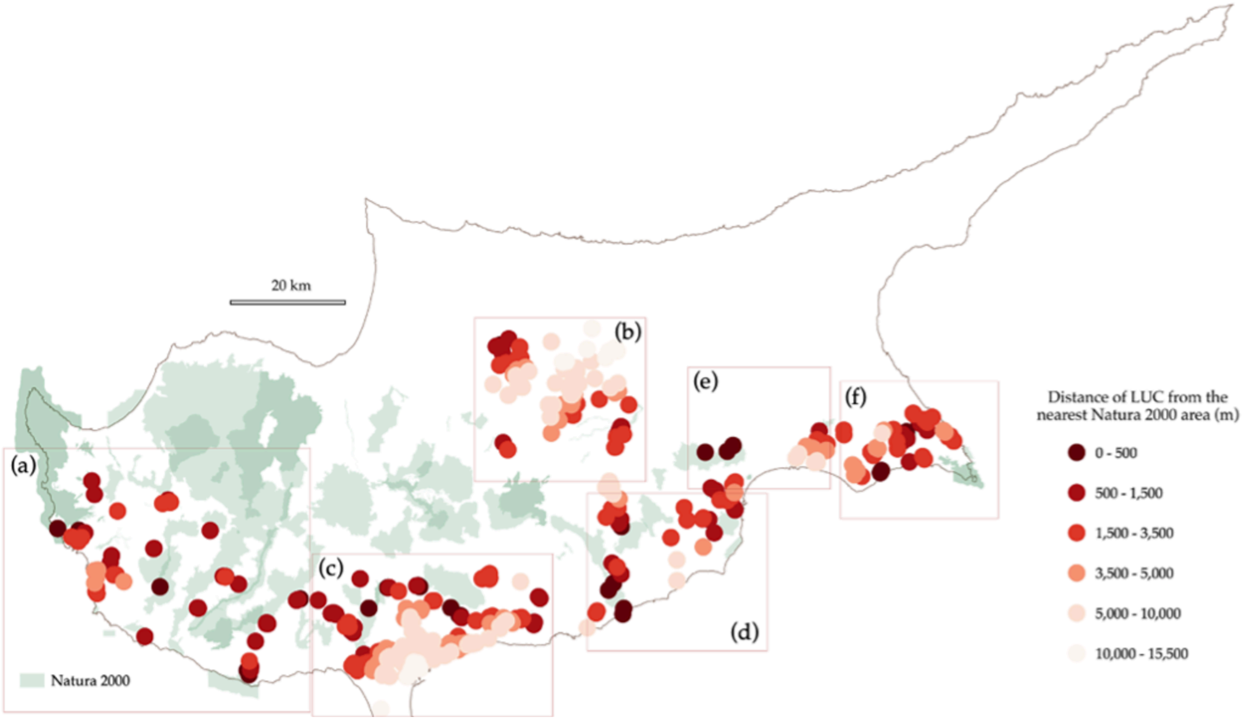


Figure 3: Map of Land Use Change near Natura 2000 areas in Cyprus — via Eftychiou [11]

Castro-Salazar et al. utilized Earthmap to produce the data visualization in Figure 4 [4][12], showing in a few panels a number of climate and vegetation trends affecting lake Chad, together with a model's interpretation of where adjustments might be most effective. This story is notable for demonstrating how an argument can be visually laid out by adding and removing map layers combined with graphs that track a variable across time. A correlation is quickly and effectively implied. The inclusion of the last panel is puzzling, as no pattern appears discernable.

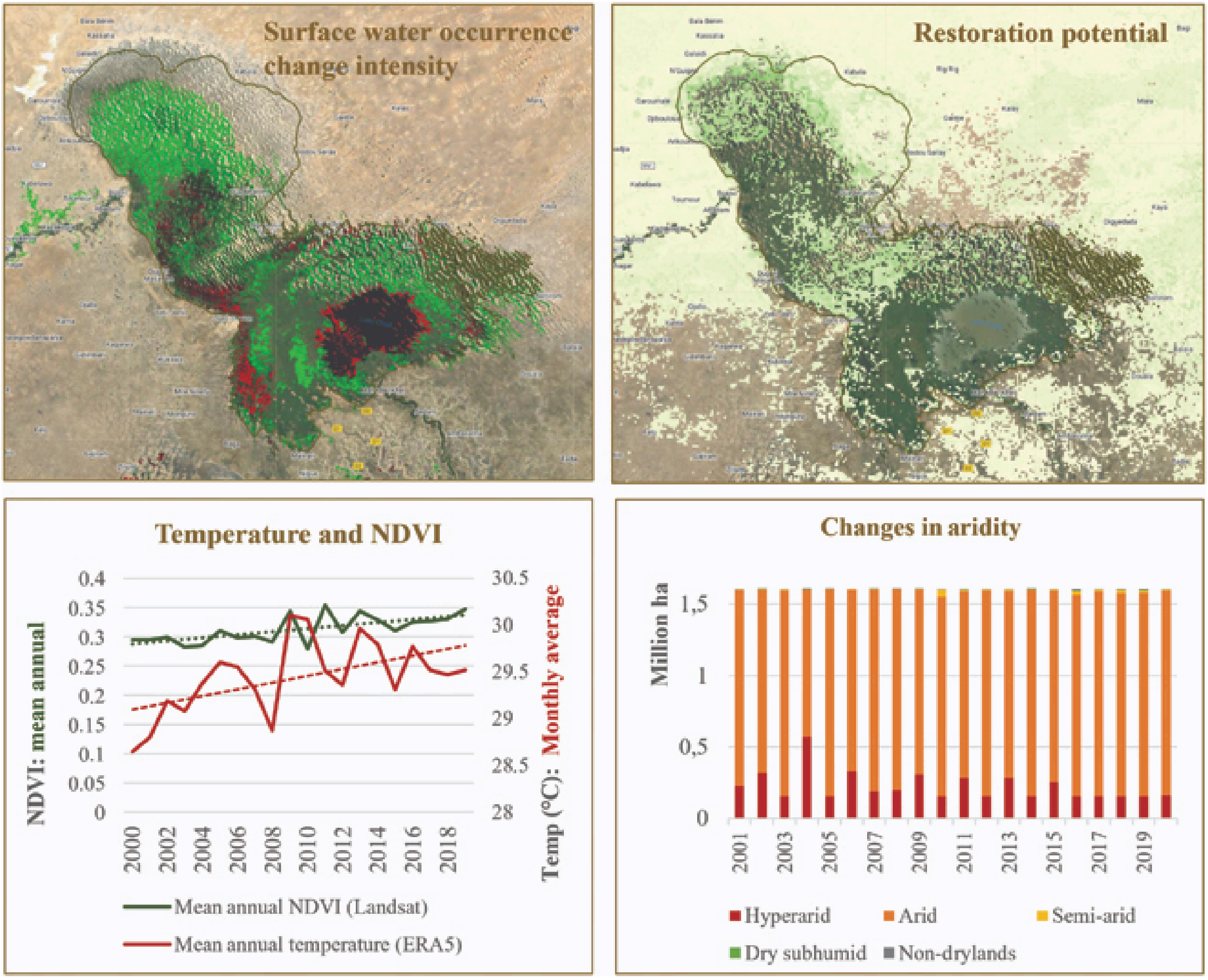


Figure 4: Lake Chad surface water restoration and temp — via Castro-Salazar et al. [12]

Chapter 3 — Methods and Techniques

3.1 Adaptating the CreaTe Design Process

Mader and Eggink provide a design process used extensively in the Creative Technology programme [13]. An adapted version of this process is used throughout this project, and for the structure of this document. The method calls for starting with a design question, here RQ 1, and progresses through four stages to ultimately produce a technological artefact. In this case the product is a Data Story, most concepts inherited from this method apply. The greatest point of deviation is in moving user-stakeholder requirements to specification, as we are also applying the data-exploration method called for by Weber et al. [9].

3.1.1 Ideation: Divergence -> Convergence

Starting from our research questions, GAEA's data dimensions were considered for potential interesting patterns, novelty of these data and these patterns, and on the aethetic potential of any visualisations that might result. Further, practical constraints regarding the data such as resolution were considered. This was carried out in biweekly consultation with the client representative, a domain expert on the GAEA tool and data visualization.

From the outset, the aim was to identify a handful of candidate stories, and to then pick a main candidate from these. In order to converge in this manner, many data combinations were first tried, and initially promising categories of stories were considered.

3.1.2 Specification

In the specification phase, requirements will be defined to in consultation with the client representative, to clarify what elements of the central idea are crucial, and which are simply nice properties to enable the project to run smoothly. This phase is adjusted from its typical implementation within the CreaTe method as the end-product, the data story, shall deliberately not be fully envisioned in keeping with the explorative method, where the data story only comes out of the data after they have been thoroughly studied.

3.1.3 Realisation

The realisation of the data story, which shall be rendered also in this chapter, requires a lengthy much data processing and analysis to arrive at. This is the results phase, and the corresponding chapter shall also render account of the process by which the data story is ultimately created.

Chapter 4 — Ideation

4.1 Divergence: Considering GAEA’s Data

We come armed with three positive qualities to look for: (1) interesting patterns, (2) novelty, (3) aesthetic potential. It is important to note that the only objective elements in these is the mere presence of correlations, the absence of these in existing work, and what is technically possible to render visually. “What is interesting about a pattern?”, ”What is new information to an imagined reader?”, “What is beautiful?”: these are all very subjective questions, which can only be approximated in any objective sense by finding other subjects to discuss with and noticing patterns of answers. In this ideation process such feedback was obtained from the client representative, but subjective judgements nonetheless had to be made about which data dimensions to consider in depth, and which not to pursue. This approach is ultimately in line with the Divergence part of the CreaTe design method [13].

When brainstorming in Divergence, one typically does not want to limit the design space. Limiting factors are usually applied when the concept space is largest in order to thin down on options to continue further with. In the case of this project the number of potential combinations from the data were so large that limiting factors were needed early in ideation. Especially when combining multiple data dimensions, and including data sources outside GAEA’s set, the combinations are endless. As such, the availability and quality of the data weighed strongly, together with the feasibility of working with said data. Also, in keeping with the need to have a designer’s eye for determining what patterns might be interesting, a good number of environmental services could be dismissed from consideration for this project, for being unlikely to combine in interesting ways with any other data dimension. ‘Detection of Swimming Pools’ for instance appeared unlikely to synergize well with other data. Furthermore, at up to 77% accuracy of detection, the data quality is not as high as most other dimensions.

4.1.1 Divergence: Interesting Data Dimensions in GAEA

While considering GAEA’s data dimensions narrative synergy was strongly considered. For example, the prevalence of climate, geomorphological, georisk, and landcover type factors produced many combinations with a shared theme related to climate. Data dimension combinations which seemed to share such a theme were clustered hierarchically. For instance:

- Vegitation type * Factor
    - Vegitation type * Geomorphological Char.

        - Vegitation type * Slope & Aspect

4.1.2 Divergence: External Data

After combinations native to GAEA were made, some data external to GAEA were considered. Most of these external dimensions were aggregated by Earth Map [4], covering similar dimensions as GAEA, though often at a lower resolution. A dataset of Archeo-sites compiled by Crawford was also considered to potentially work with natura 2000 areas [14].

Ultimately, there were few interesting outside sources of data found at this stage, as the vast majority of outside data dimensions regarding the physical characteristics of Cyprus were of similar or worse resolution than a data dimension in GAEA, bringing neither the potential for novely nor ease of work through higher granularity data.

4.2 Convergence: Obtaining Five Potential Data Stories

After a sizable number of potential combinations were made and discussed in consultation with the client representative, the method calls for converging back down to a handful of potential options. This step involved turning the most promising data combinations into potential story concepts, and then testing these concepts against the parameters we care about. A short description of each story concept as considered and discussed in this phase is given below together with the data dimensions it looks to combine:

- *1. "For a given area of land, we can look at the amount and type of vegetation and find how this relates to the soil type & depth and the geomorphological characteristics of the land."*
    - *Vegetation presence & type * Slope & Aspect & Elevation * Soil type & Depth*
- *2. "For a given area of land, we can look at the amount and type of vegetation and find how this relates to the precipitation & humidity""*
    - *Vegetation presence & type * Precipitation & Humidity*
- *3. "For a given area of land, we can look at the amount and type of vegetation and find how this relates to the risk of wildfires."*
    - *Vegetation presence & type * Wildfire risk & burn proof*
- *4. "By comparing various landcover / landuse datasets and overlaying them, differences in their classification of the same area might be found, especially as resolution differers greatly between datasets"*
    - *Various landcover type datasets*

- *5. “GAEA offers many goerisk services, which could be overlaid with natura and archeological site clusters to determine which are most at risk.”*
  - *Archeo-sites * Natura 2000 areas * Georisks*

Out of all categories of data combinations “Vegitation presence & type * factor” yielded the greatest number of potential combinations. As other story options up until this stage could be rejected by some fatal flaw it was not yet necessary to consider each in detail for all the factors we care about. For example, one potential story hinging on a difference in solar irradiance across the island could be rejected on the basis of insufficient difference in this factor across space. The remaining stories were considered for a number of discussed factors, and these judgements were then again considered in consultation.

The table below sums up the result of that process, which each of the options’s strengths and weaknesses compared with the alternatives. As Vegetation type also covers tree types, and as this is a novel data set, the first three similar options performed well. The version testing principally against geomorphological characteristics outperformed because vegetation type as compared to wildfire modeling or water factors was judged as likelier to lead to interesting and novel stories, more space was opened up as a result of this data set than for the alternatives. Though there is a lot of hydromorphological data available, the pattern tends to be quite uniform. Wildfire risk is an interesting comparison, but since underbrush cannot yet be distinguished from trees it would be hard to draw strong causal links. The fourth option would consider different land cover datasets, and so cannot suffer from low data quality, but would require comparatively much data manipulation and analysis, requiring more expertise with GIS tools and data science than the scope of this project might permit. Finally, Archeo-sites did not yield interesting combinations in this phase, as there was no discernable pattern of interest in site distributions.

	1. Geomorph	2. Hydromorph	3.Wildfire	4. Diff	5. Archeo
How interesting is the pattern?	++	+	++	+/++	-
Is the data combination novel?	++	+	+	+	+
Aesthetic potential	+/++	+	+/++	+/++	+

Data Quality	+	+	+	n/a	+/-
Feasibility	+	++	-	-	+
Uses outside data	no	no	no	yes	yes

Figure 5: Convergence table

As a result of the convergence step of ideation, the principle data combination to chosen is then “*Vegetation presence & type * Slope & Aspect & Elevation * Soil type & Depth”*. These data then are obtained, analysed for interesting patterns, and supplemented as needed.

Chapter 5 — Specification

Specification marks the start of the second half of the project, beginning with further concept validation of the chosen potential data story through unstructured in-person discussion with the client representative. Potential requirements were broadly covered, and the availability of the needed data and the means of obtaining these were confirmed. It was decided in this stage that an explorative approach would be taken, whereupon data would be fully processed and analysed in search of patterns. A data story would then result from whatever patterns were present in these data, an assumption previously validated as likely.

It was clear at this stage that the nature of this project required a slight deviation from the CreaTe design method, as it was estimated that a good portion of the work would be research oriented. Whereas this design method would usually require a clear potential target audience, the order of operations taken instead demanded that the right audience would be found for whatever actual patterns emerged from the research. For these reasons, specification focused on those requirements necessary to conduct the research, in the estimation that this would allow us to find a subset of the general population interested in the results. For these reasons, the needed data and tools to process and analyze these data were the principle focus at this stage, and would greatly inform the set of requirements generated.

Further, since stakeholder validation was to take place through regular meetings with the client representative, the shared vision of how the story might take shape was deemed sufficient at this stage not to warrant further envisioning. By not committing to any further specification to the data story itself beyond those already discussed, it was reasoned that research could be conducted without undue preconceptions. Only by not marking the end-point ahead of time, can explorative research of this kind allow a data story to be truely driven by the data, as opposed to the existing preferences of the researcher.

5.1 Tools

5.1.1 Geographic Information Systems

As noted in the review of the literature, GIS tools are extremely useful, bordering necessary, for studying geospatial data [1]. Since QGIS offers a free opensource platform with plentiful documentation and integration with libraries such as GDAL for the reading and processing of geospatial data, it was quickly chosen as the preferred tool for the job. The client representative indicated that members of his team who would aid in providing the needed data were familiar with ArcGIS, and it was found that these were capable of similar operations, and that data formats likely to be used were compatible between them. Since this choice was clear, it played no further part in specification.

5.1.2 Python

Similarly, the python programming language is available on a host of opernsource platforms, and has a wealth of libraries frequently useful to the processing, analysing, and visualizing of data. These data tasks likely to be encountered, such as joining tables or performing various statistical tests, are so frequently performed with Python that these functionalities are essentially available as boilerplate code, needing very little adjustment to perform any needed tasks. For these reasons Python was chosen as the second main tool in executing this project. The pandas library was selected for data manipulation and analysis, and seaborn as well as matplotlib for visualization for these same reasons.

5.2 Data

From the findings in ideation the data dimensions required to conduct the research necessary for this project were initially clear: “Vegetation presence & type * Slope & Aspect & Elevation * Soil type & Depth”. Since the client indicated at this stage that “Tree Classification” data marking the most predominant type of trees across Cyprus would be available, this specific data was chosen for “Vegetation presence & type”. In order to obtain the correct data from the client, it was necessary to understand how these data are typically formatted for GIS software, which was done via online research and confirming with the client representative.

It was found that two principle methods of representing geospatial data were most prominent and likely to be used: vector and raster data, via the “Shapefile” and “GeoTIFF” formats respectively. Since vector representations are not bound by resolution, but may equate and be converted to rasters, shapefiles were deemed preferable where possible. The geomorphological characteristics of “Slope, Aspect, Elevation” are frequently rasterized in the Digital Terrain Model (DTM) format as GeoTIFF files, so this was a viable alternative these data.

At this stage it was unclear what format the Tree Class data would come in, but discussions with the client representative made it clear that it would consist of a grid-like structure where each cell of the grid would be coded with the tree most common in that cell, and that this classification was carried out by means of an AI algorithm, which had predicted labels, and was corrected where these predictions were incorrect. Recalling that the granularity of data is crucial to representing the underlying reality, and to analysing it successfully, as high a resolution as possible was desired. Early discussions indicated that a resolution of $25m^2$ might be available.

GIS systems have to project data about the earth's surface, which is curved, onto a flat representation on screen. To this end, geospatial data formats frequently encode a Coordinate Reference System (CRS), ensuring proper representation of the data and allowing conversion between multiple projections or reference systems. The World Geodetic System (WGS 84) is the standard for QGIS, it would be preferable if all data conformed to this CRS, else conversion operations would be necessary to properly align the data later. These data formats also frequently include metadata alongside a CRS, which may include information such as when the data was obtained. Though these are not crucial, the inclusion of metadata is noted in the research as positive feature of data stories [9]. This section of properties that may come with the data represent non-essential elements which nonetheless are helpful.

Finally, the workstation available for this research had 16GB of RAM, 1TB of available storage, and a slow processor. The data would have to not exceed this size in either storage or in loading tables into RAM. Some testing indicated that tables represented as .csv files should not exceed ~20m rows * 10 columns of typical geospatial data, lest operations could not be performed.

5.3 Functional Requirements

Requirements may be split into those about the research project on which the data story is to be based, and the data story itself. Functional requirements pertain to the essential requirements that allow something to function, whereas non-functional requirements describe how something might be done: things which are nice to have but will not be missed in a minimum viable product.

5.3.1 FR: Data & Research

- Tree Classification data
- Available in a format which can be made to work in QGIS.
    - Vector representation
    - Raster representation
    - Tables
- Not exceeding 1TB in total dataset size.

- Not exceeding individual, unsplittable data sizes such that a computer with 16GB of working memory cannot compute simple data operations.
    - read, join, etc.
- The data is the best consistent and contingent representation of the requested data dimensions available at the time of research.
- The research conducted on these data shall correlate the available data dimensions in search of notable patterns, and will seek to validate these patterns to the extent that the researcher's time and abilities allow.
- All conclusions resulting from this research which are presented in the resulting data story or thesis documentation shall be a truthful representation of the researcher's interpretation of the data, so that these may be useful to readers.

5.3.2 FR: Data Story

- Describes the physical characteristics as captured by the selected data dimensions and their combinations.
- Presents results of research on these patterns to readers in a text and image based format compatible with publication on the web.

5.4 Non-Functional Requirements

5.4.1 Non-FR: Data & Research

- All other requested data dimensions
- A consistent CRS to skip alignment operations
- A consistent format within each data dimension to minimize preprocessing operations
- Lossless conversion and transfer of data where possible
- The highest resolution available not exceeding functional limits.
- Explicit or else implicit model accuracy data and/or test results.
- Availability of metadata where possible
- Research shall focus on finding patterns that are interesting, novel, or that have good aesthetic potential in a data story.
- Research shall focus on correlations between Tree Class data and the other data dimensions available, not on correlations between non-Tree Class data.

- Research interpretation shall be limited to the domain knowledge of the researcher, leaving complex potential implications to future researchers.

5.4.2 Non-FR: Data Story

- Presents results in such a way that they are clear to the average informed reader, and potentially useful or informative to the subset of readers with a specific interest in the data and characteristics covered.
- The visualizations chosen are easy to read and understand to those familiar with the mode of representation.
- The patterns presented are novel
- The patterns presented could lead to future research

5.5 Client Validation

These requirements were discussed and validated with the client representative before they were formalized, and before eventual further stakeholders could be identified. These capture the set of considerations that were informally present from the start of the realisation phase, and no major alterations or additions were deemed necessary in these phases. These specifications capture the researcher's understanding of the shared intention and vision as the project entered the research phase of realisation.

Chapter 6 — Realisation

In documenting a project using the CreaTe design process, the realisation chapter both demonstrates the resulting design artefact of the project, and documents the process by which it was produced. It seems producent to first show the Data Story in full, and to then after cover the steps and methods used in creating it. Following the Data Story, this chapter shall therefore continue where specification left off with the process of data acquisition, processing, research, analysis, and the selection process of the results that would form the basis of the data story. Finally, evaluation was conducted with respect to the requirements.

The data story is ultimately intended for publication on the web, unconstrained by the the margins of the A4 format, the included visualizations are therefore represented slightly smaller here than they likely will be in final publication. Figures are intended to be self-explanatory given the context, and figure numbering is here adjusted to follow the thesis format, included (in parentheses). References likewise are included separately after the conclusion in future rendition.

6.1 — The Data Story

Data Story:

Cyprus lies in the ‘Mediterranean Forests, Woodlands, and Scrub’ biome, hosting a variety of vegetation across its landscape. In this data story, we will explore how geospatial data may be used to better understand the relations between various flora and their enviornment. Our data shall come from the recently released ‘GAEA’ geospatial tool, developed by CYENS’s SuPerWorld Research Group as a digital twin of Cyprus [2]. They have developed an AI model for classifying trees from satellite images, letting us know where in the country each type of tree is most predominant. Let us consider the spatial distribution of Figs and Walnuts for an example:

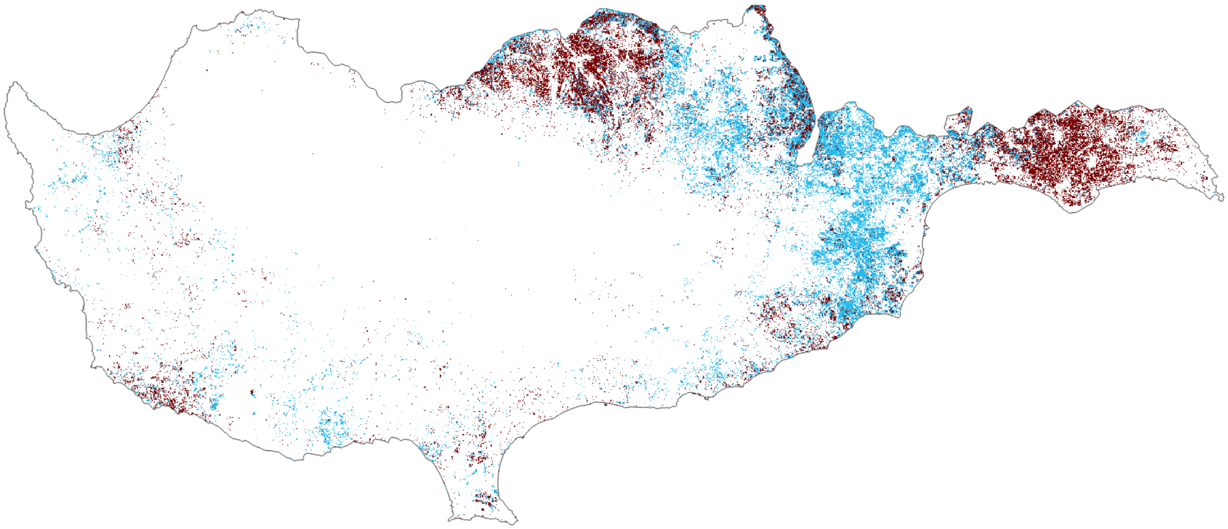


Figure 1 (6)

Your eyes may first be drawn to the northeast, where these trees appear to be quite common compared to their sparse presence around the coasts. Then, you may notice the blank void in the middle of the country. If we add an elevation map underneath we are likely to get a clear picture of the cause of this distribution:

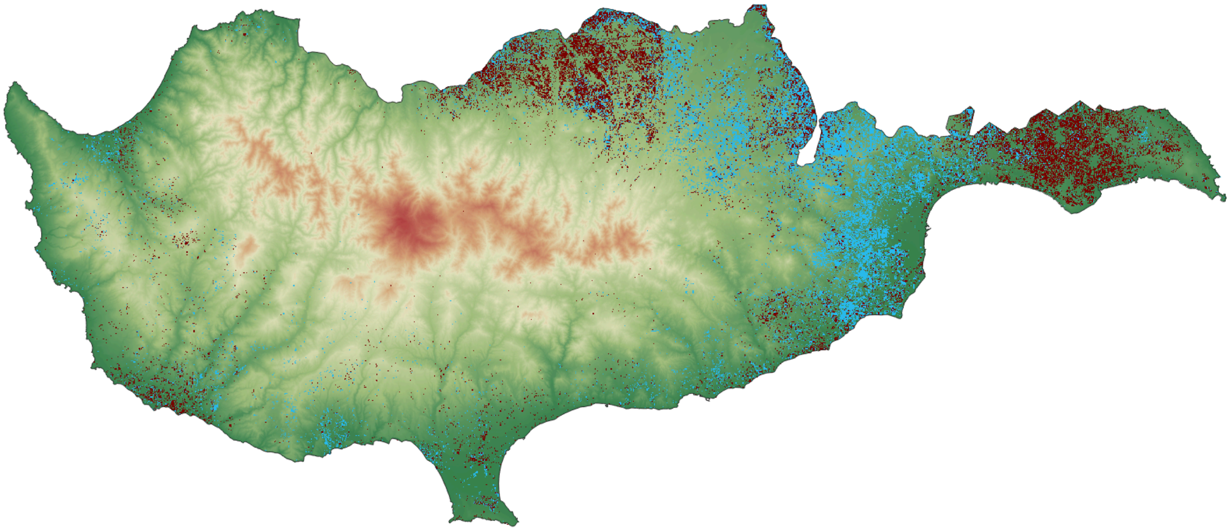


Figure 2 (7)

It would appear that these trees share in common that they prefer to grow at lower elevations. We cannot however be sure if what we are seeing holds unless we confirm by running the stats and seeing exactly how these trees are distributed across elevation. Our data covers 13 different classes of trees, each dot on the map representing a 45m^2 area where that class of tree is judged more common than the rest. We can count all these up and distribute them across elevation via a box plot:

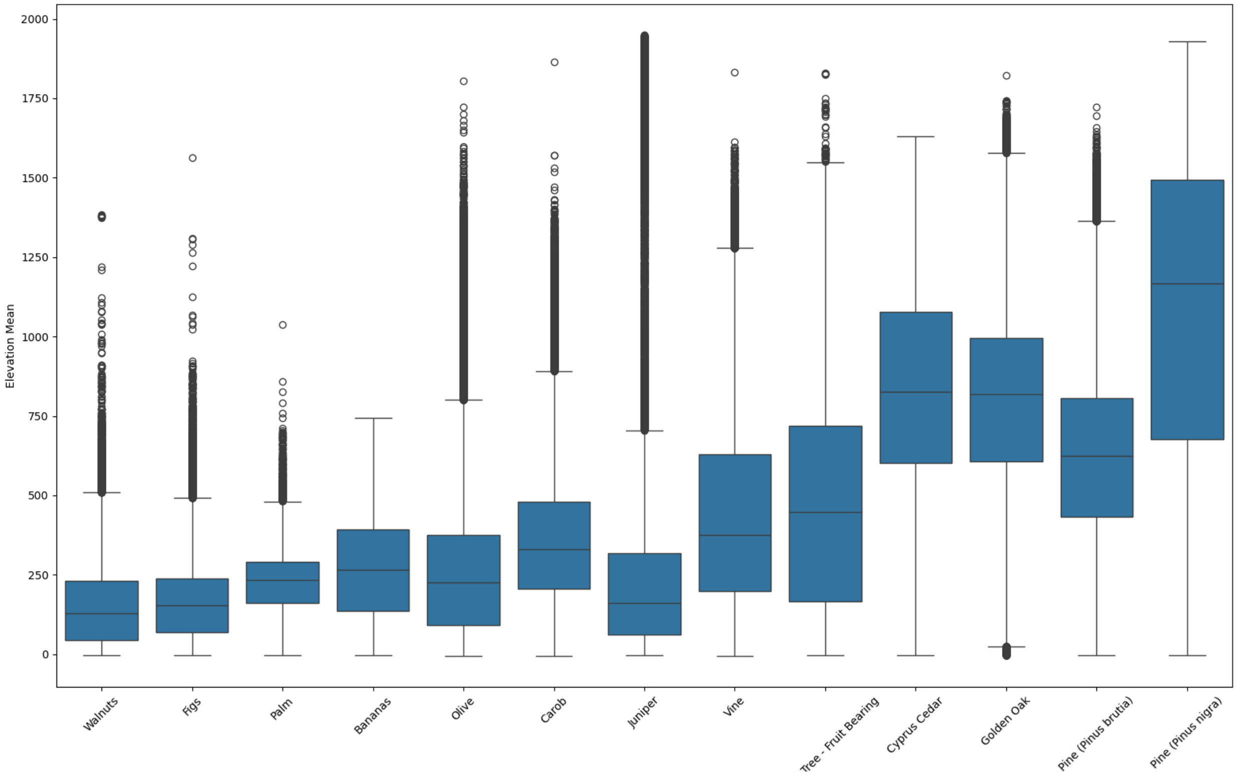


Figure 3 (8)

The center of each box marks the average elevation where each tree occurs, with whiskers and individual dots extending beyond each whisker to indicate the full range of each tree, even where it is less common, or where there are individual outliers. Walnuts and figs show a very similar pattern, much preferring lower elevations as compared with trees further up the plot. We might be able to identify further patterns by grouping tree classes by their apparent distribution across elevation:

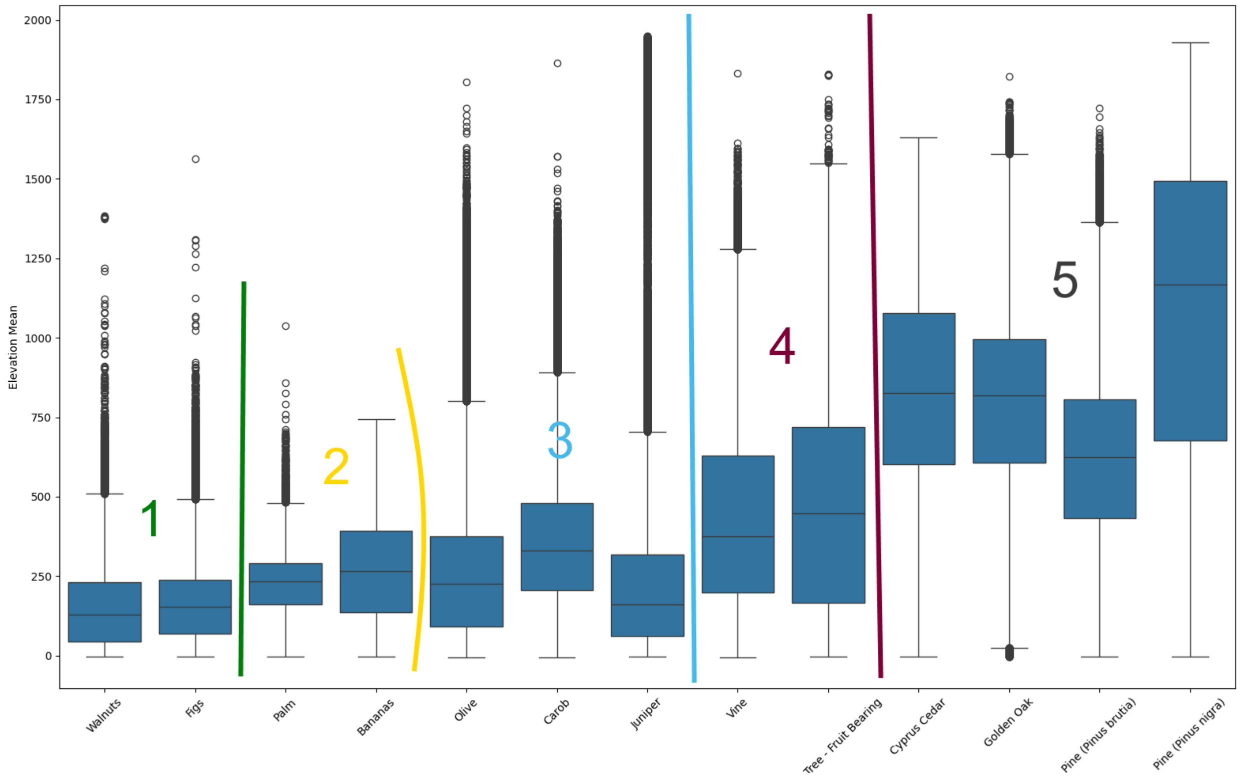


Figure 4 (9)

There are many ways to group these classes, but the 5th grouping here is perhaps the most striking, as these might correspond to the area in the center of the country where other trees are relatively rare. By looking only at this group of trees over elevation we get a clear picture of trees common to the Paphos Forest in the Troodos Mountains: Brutia Pine, Black Pine, Golden Oak, and Juniper.

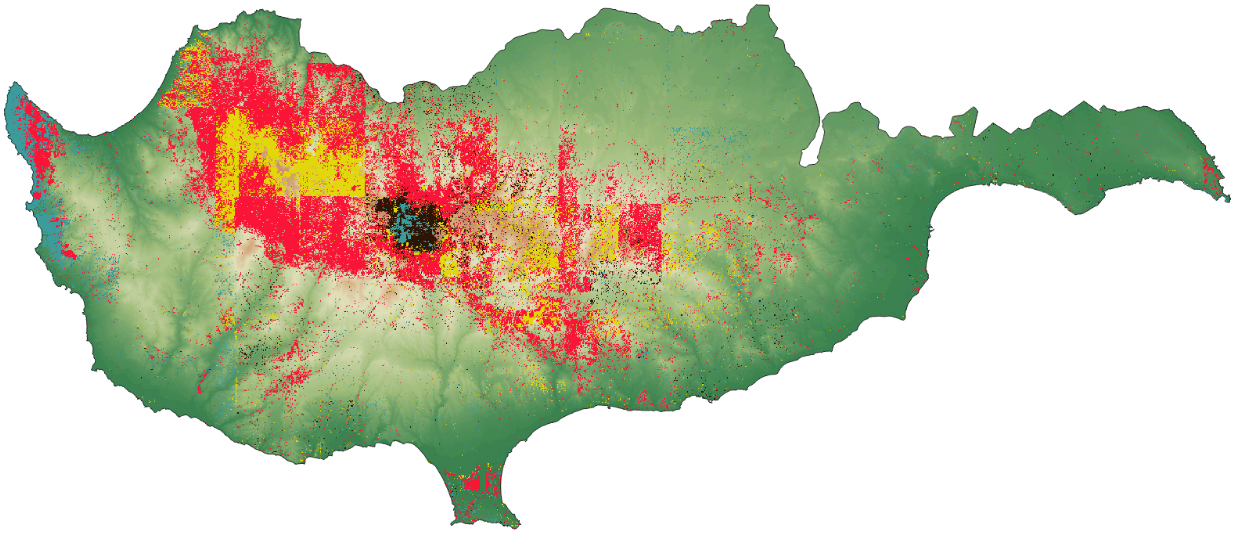


Figure 5 (10)

Juniper appears to be a special case, present in two large clusters to the far west and central heights of the mountains. To gain a deeper understanding of the patterns we are seeing it is necessary to consider other data dimensions, such as the Slope, Aspect Facing, Soil Type, and Soil Depth. As we zoom into our area of study we can overlay the Soil Type over the relief map:

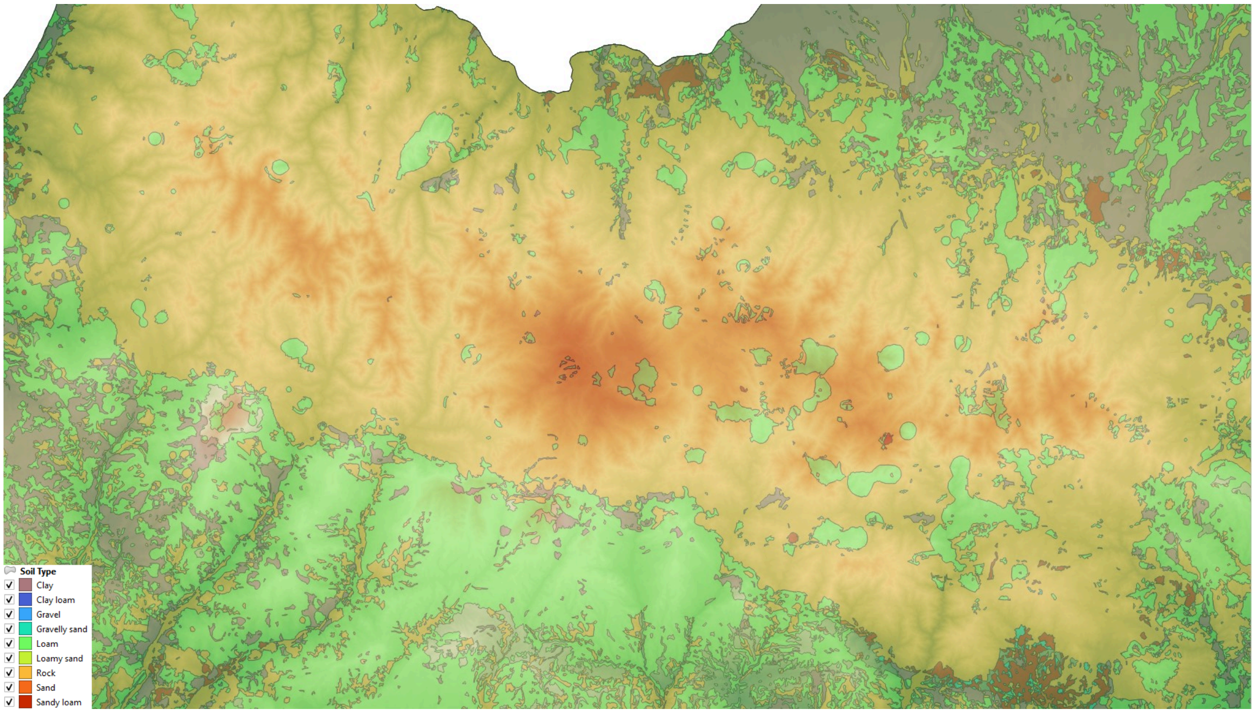


Figure 6 (11)

We find Rock is the most common soil in this region, with patches of Loam and Clay present also. By now overlaying our grouped tree data we may learn more:

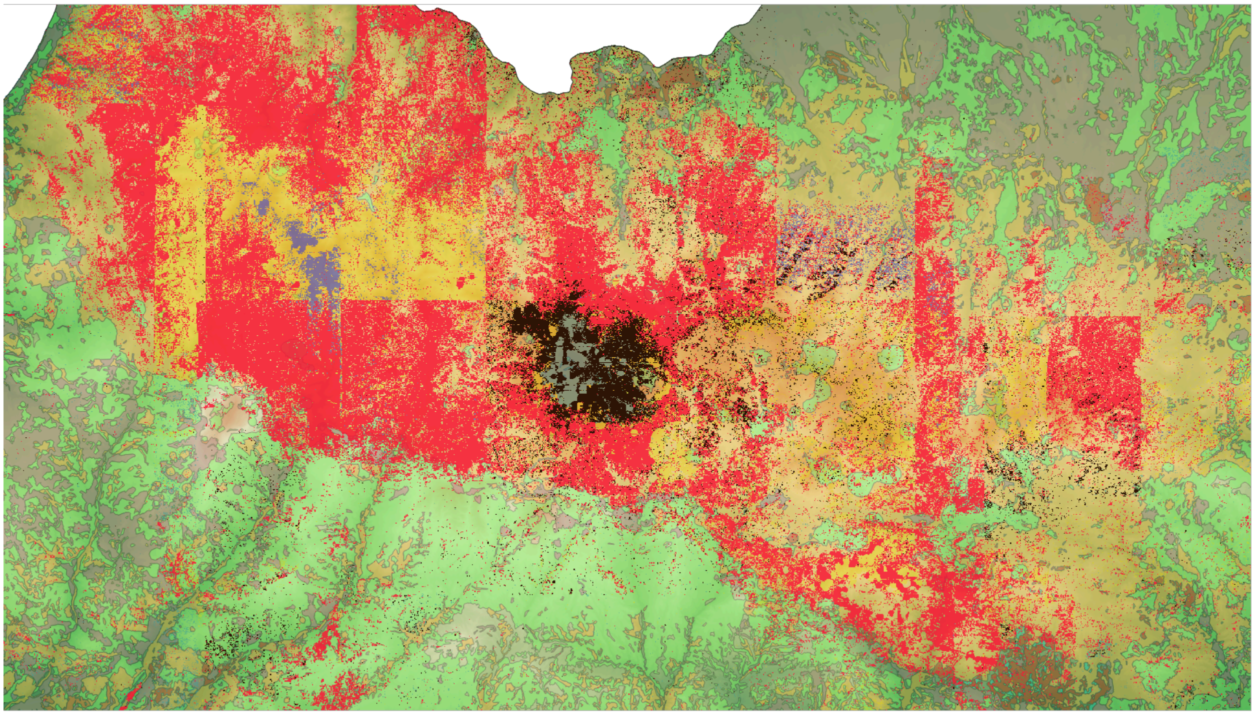
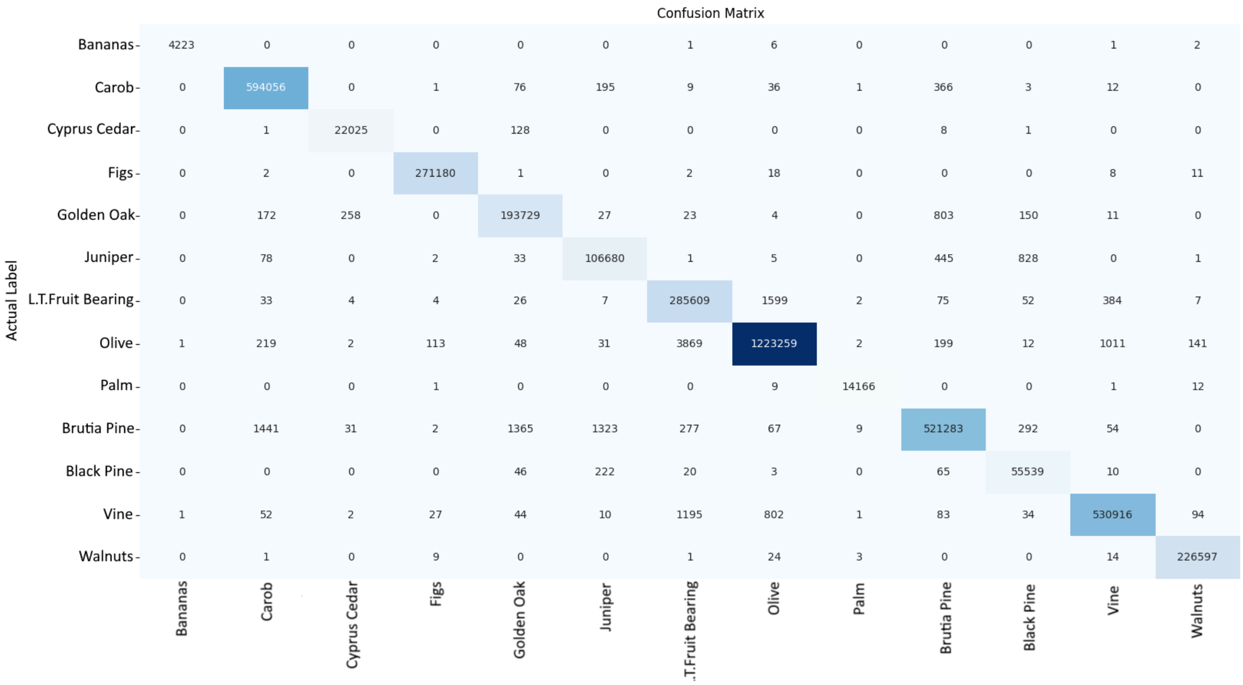


Figure 7 (12)

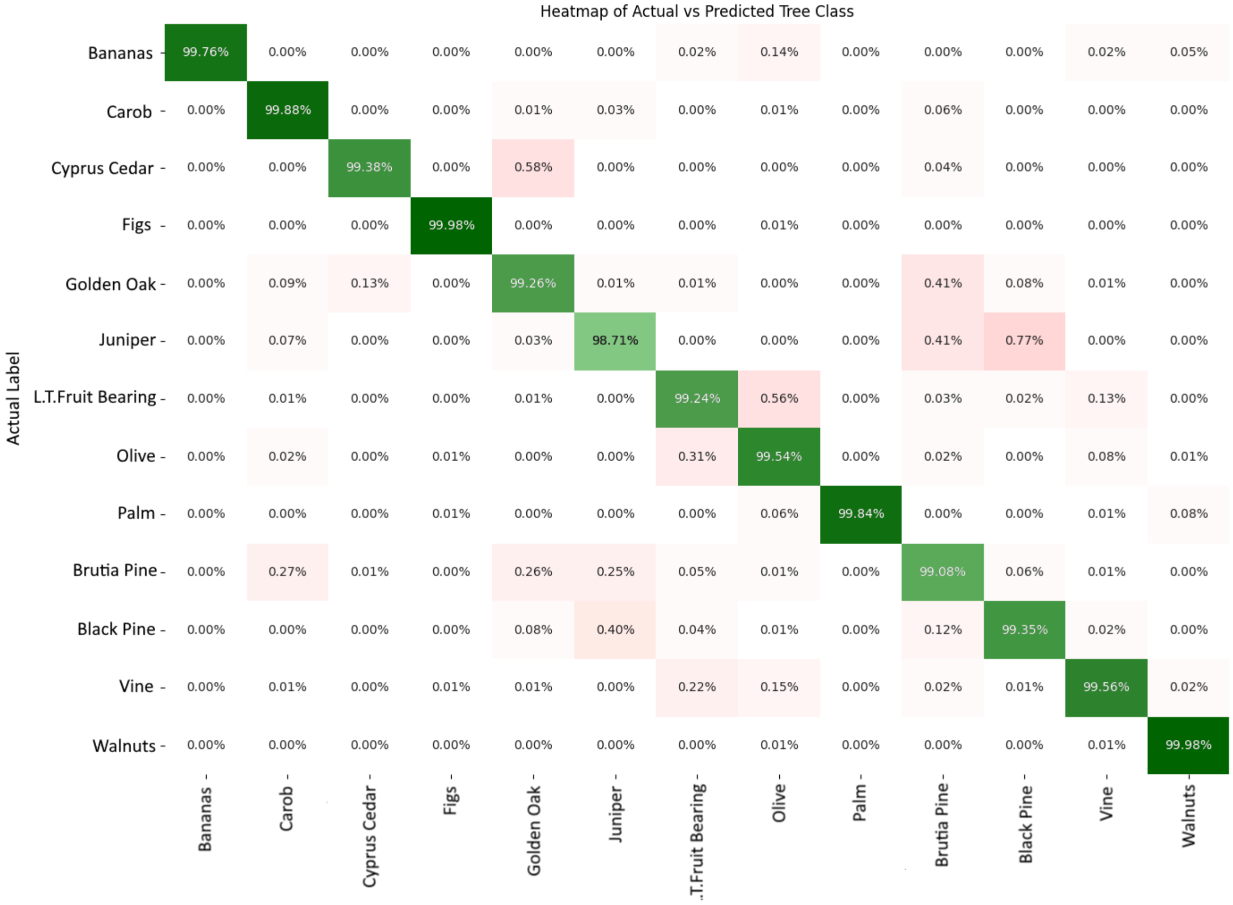
Rocky soil, which permits little depth for water, is fine for these trees [15], though they appear scattered also across Loam. Further, we can see that Fig clusters together with Black Pine around the highest area of the forested mountains. Some hard edges are visible in the tree class data. Since a patch-wise algorithm has to determine which of the tree classes is more common than the rest, when both are present slight variations across patches can produce these artefacts, which average out across the data. To study these relationships in more depth, and to make confident claims about the patterns we are observing, we can turn to statistical software to analyse further.

Firstly, we can confirm that the AI algorithm correctly assigned tree classes by checking in what percentage of cases the initial label had to be corrected. Juniper had the lowest initial label accuracy at a strong 98.71%, whereas Brutia and Black Pine scored 99.08% and 99.35% respectively:



Confusion Matrix

	Bananas	Carob	Cyprus Cedar	Figs	Golden Oak	Juniper	.T.Fruit Bearing	Olive	Palm	Brutia Pine	Black Pine	Vine	Walnuts
Bananas	4223	0	0	0	0	0	1	6	0	0	0	1	2
Carob	0	594056	0	1	76	195	9	36	1	366	3	12	0
Cyprus Cedar	0	1	22025	0	128	0	0	0	0	8	1	0	0
Figs	0	2	0	271180	1	0	2	18	0	0	0	8	11
Golden Oak	0	172	258	0	193729	27	23	4	0	803	150	11	0
Juniper	0	78	0	2	33	106680	1	5	0	445	828	0	1
L.T.Fruit Bearing	0	33	4	4	26	7	285609	1599	2	75	52	384	7
Olive	1	219	2	113	48	31	3869	1223259	2	199	12	1011	141
Palm	0	0	0	1	0	0	0	9	14166	0	0	1	12
Brutia Pine	0	1441	31	2	1365	1323	277	67	9	521283	292	54	0
Black Pine	0	0	0	0	46	222	20	3	0	65	55539	10	0
Vine	1	52	2	27	44	10	1195	802	1	83	34	530916	94
Walnuts	0	1	0	9	0	0	1	24	3	0	0	14	226597



Heatmap of Actual vs Predicted Tree Class

	Bananas	Carob	Cyprus Cedar	Figs	Golden Oak	Juniper	.T.Fruit Bearing	Olive	Palm	Brutia Pine	Black Pine	Vine	Walnuts
Bananas	99.76%	0.00%	0.00%	0.00%	0.00%	0.00%	0.02%	0.14%	0.00%	0.00%	0.00%	0.02%	0.05%
Carob	0.00%	99.88%	0.00%	0.00%	0.01%	0.03%	0.00%	0.01%	0.00%	0.06%	0.00%	0.00%	0.00%
Cyprus Cedar	0.00%	0.00%	99.38%	0.00%	0.58%	0.00%	0.00%	0.00%	0.00%	0.04%	0.00%	0.00%	0.00%
Figs	0.00%	0.00%	0.00%	99.98%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%
Golden Oak	0.00%	0.09%	0.13%	0.00%	99.26%	0.01%	0.01%	0.00%	0.00%	0.41%	0.08%	0.01%	0.00%
Juniper	0.00%	0.07%	0.00%	0.00%	0.03%	98.71%	0.00%	0.00%	0.00%	0.41%	0.77%	0.00%	0.00%
L.T.Fruit Bearing	0.00%	0.01%	0.00%	0.00%	0.01%	0.00%	99.24%	0.56%	0.00%	0.03%	0.02%	0.13%	0.00%
Olive	0.00%	0.02%	0.00%	0.01%	0.00%	0.00%	0.31%	99.54%	0.00%	0.02%	0.00%	0.08%	0.01%
Palm	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%	0.06%	99.84%	0.00%	0.00%	0.01%	0.08%
Brutia Pine	0.00%	0.27%	0.01%	0.00%	0.26%	0.25%	0.05%	0.01%	0.00%	99.08%	0.06%	0.01%	0.00%
Black Pine	0.00%	0.00%	0.00%	0.00%	0.08%	0.40%	0.04%	0.01%	0.00%	0.12%	99.35%	0.02%	0.00%
Vine	0.00%	0.01%	0.00%	0.01%	0.01%	0.00%	0.22%	0.15%	0.00%	0.02%	0.01%	99.56%	0.02%
Walnuts	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%	0.01%	99.98%

Figures 8, 9 (13, 14)

By referencing the confusion tables we can see which combinations caused trouble with assignment, namely those that grow very close together in densely forested areas. Our grouping from the elevation data shows up again, the Pines are initially relatively commonly confused for nearby trees. By replacing these known misclassified labels with known actual labels early on the validity of future analysis is greatly improved.

The two types of pines common to Cyprus may have a complicated relationship. In a recent study, Petrou et al. showed that in the elevation areas Black Pine prefers, they grow adjacent to each other in “mixed stands” [15]. Further, it was discovered that Brutia Pine promotes biodiversity especially well in areas which were not of “bad productivity”, low soil depth:

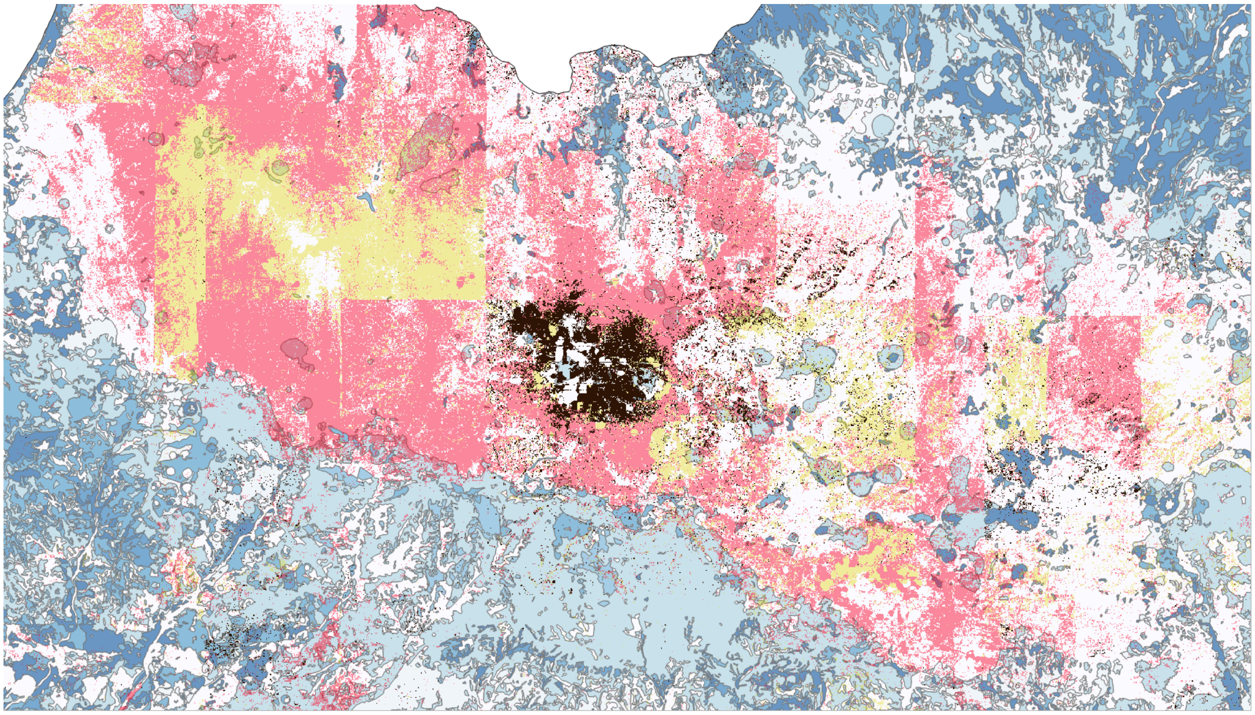


Figure 10 (15)

Recalling soil type, the lowest soil depth corresponds with bedrock, whereas deeper soil depths are found in loam, especially clay loam and clay proper. Through correlating these Tree Classes with soil, we can find that Brutia Pine is found predominantly on rocky soil, at 83.97% of cases, and secondly in loam, at 12.64%. It shows to some extent in all soil types, confirming older research claims that this species can adapt to any soil type present on Cyprus, as cited by Petrou et al. [15].

Interestingly, Black Pine only grows on bedrock in 78.54% of cases, being more common in loam than Brutia Pine (16.91% v. 12.64%), and especially relatively common in Clay (3.27% v. 1.72%). It is relatively less common in sand (0.38% v. 0.69%), yet more common in sandy loam (0.59% v. 0.48%). Surely then Black pine must also be more common in Clay Loam? Not so: (0.27% v. 0.40%). This is a surprising result, the number of Black Pines is low enough in general that their presence may be underrepresented in classification data at this resolution, or there may be some unexplored interaction in this particular soil.

We can also consider slope and aspect with respect to our trees. There is little difference with regard to preference for slope, Brutia averages a few degrees higher than black pine at 23.98 degrees, only beaten by the mean slope of the Cyprus Cedar and Golden Oak at ~26-27.

Aspect was considered as the nearest cardinal direction a slope faces, due north, east, south, west. Here we find some significant patterns:

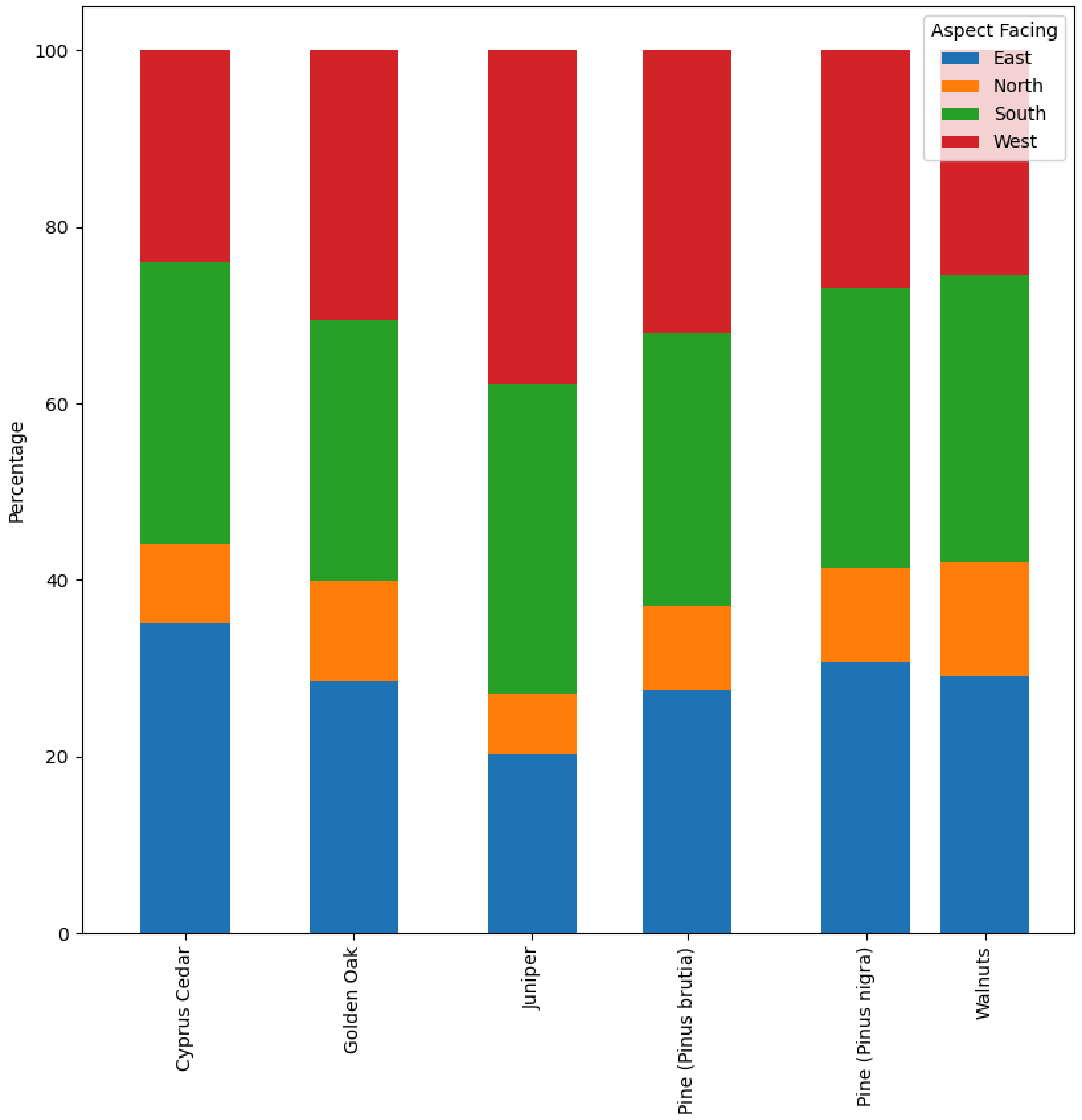


Figure 11 (16)

The great outlier in this subset is Juniper, caused by its presence on cliffs near the north-western shore. Differences here may seem minor, but note that Brutia Pine prefers the west facing slope at 32.04% of all cases v. 26.84% in Black Pine. We can observe a strong preference for south over north among all tree classes, understandable given where the sun shines in the northern hemisphere. Random as these distributions may seem, a Chi^2 test can find out if this is the case, or if the distribution of each tree reveals an underlying preference. For each direction, the average preference across all classes is taken and compared with the observed variance.

Since this test uses the total counts of each rather than percentages, it is stronger when we have a lot of data, which we do for most trees. Out of all 52 combinations, 51 were highly significant. Tree Classes v. Aspect Facing resulted in: (Chi^2: 44577, df = 36, p < .00001), confirming a very strong correlation. Tree Classes v. Soil Type was similarly tested: (Chi^2:710203, p < .00001, df = 96), we again find there is a very strong pattern between the type of tree and soil of preference.

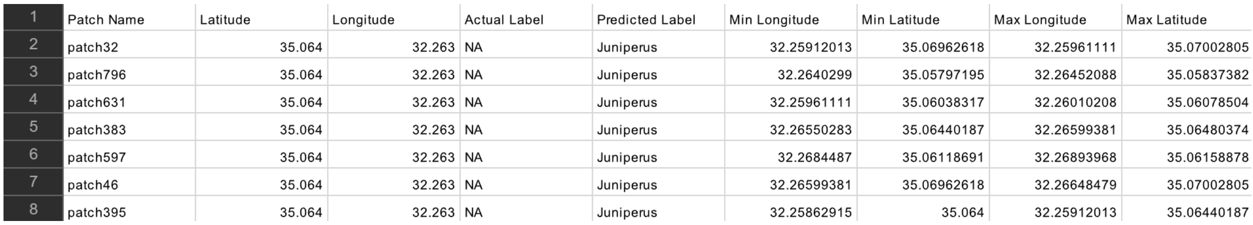
Conclusion:

By exploring these geospatial data we have gained significant insight into what geomorphological characteristics are relevant to the flourishing of several types of trees important to Cyprus. The pines of Cyprus especially have a relationship too complicated to pierce from the outside, but these data and analysis may show that digital twins such as GAEA have an important role to play in monitoring large environments. It is hoped that these results are already of some use for purposes of forestry, or even just in growing one’s understanding and appreciation of the many preferences of trees, or of those who plant them. In either case, as more data becomes increasingly available at increasing resolutions, and as data of this kind becomes available across time, the general method herein may prove vital in monitoring an environment changing with the climate.

6.2 — Data Aquisition

6.2.1 Initial Tree Data

Following specification, with the data requirements in hand it was possible to request the needed data from the client. These data met all functional requirements, and were provided at native resolution with metadata including various CRSs. Tree Class data was provided as a set of 4890 .csv tables, which each covered rectangular patches of area across Cyprus. Tree Class data was given as two columns: ‘Predicted Label’, and ‘Actual Label’, the latter being not applicable if the predicted label was initially correct. As was ultimately discovered, the Latitude and Longitude columns represent the center coordinate of each patch, whereas the min/max long/lat columns represent the bounding box where the Tree Class data of that row applies. These were assumed to represent an area of roughly 25m^2. The first column is an ID non-unique across tables, and so had no further use and would have to be cut. Figure 17 gives a snippet of this initial format:



1	Patch Name	Latitude	Longitude	Actual Label	Predicted Label	Min Longitude	Min Latitude	Max Longitude	Max Latitude
2	patch32	35.064	32.263	NA	Juniperus	32.25912013	35.06962618	32.25961111	35.07002805
3	patch796	35.064	32.263	NA	Juniperus	32.2640299	35.05797195	32.26452088	35.05837382
4	patch631	35.064	32.263	NA	Juniperus	32.25961111	35.06038317	32.26010208	35.06078504
5	patch383	35.064	32.263	NA	Juniperus	32.26550283	35.06440187	32.26599381	35.06480374
6	patch597	35.064	32.263	NA	Juniperus	32.2684487	35.06118691	32.26893968	35.06158878
7	patch46	35.064	32.263	NA	Juniperus	32.26599381	35.06962618	32.26648479	35.07002805
8	patch395	35.064	32.263	NA	Juniperus	32.25862915	35.064	32.25912013	35.06440187

Figure 17: example Tree Class table section (IMG0001.csv)

6.2.2 Geomorphological Data

The geomorphological data of Aspect, Slope, and Elevation was given in raster format, as GeoTIFF packages. A raster is ultimately an image comprised of pixels, each of these data was provided as 4 such packages, which had to be combined for further analysis. On import into

QGIS they become represented as layers, and appear as seen in Figures 18-20:

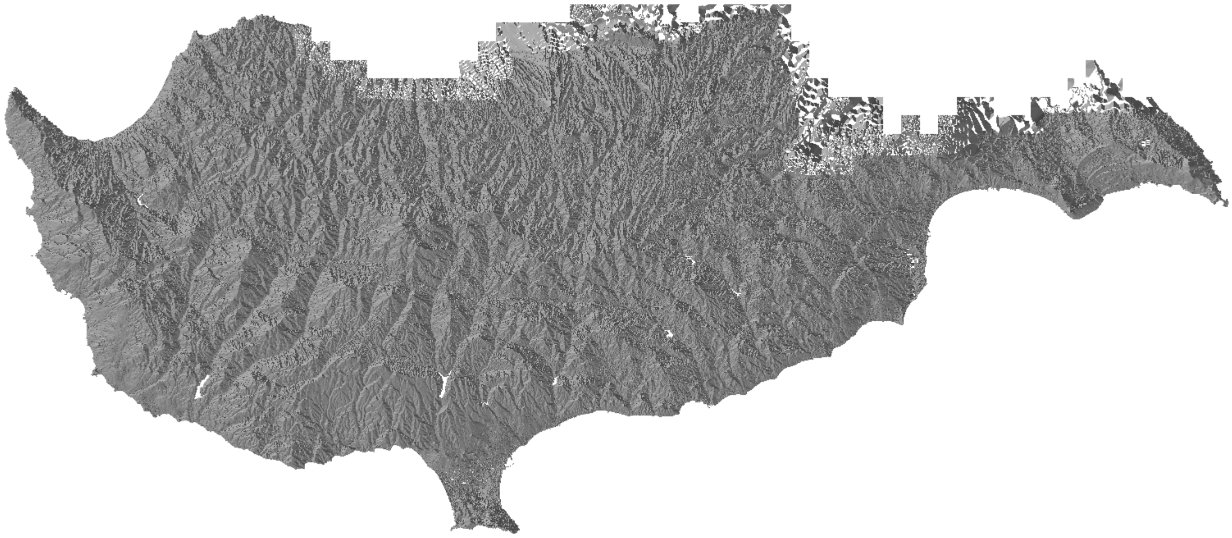


Figure 18: Raw Aspect layers

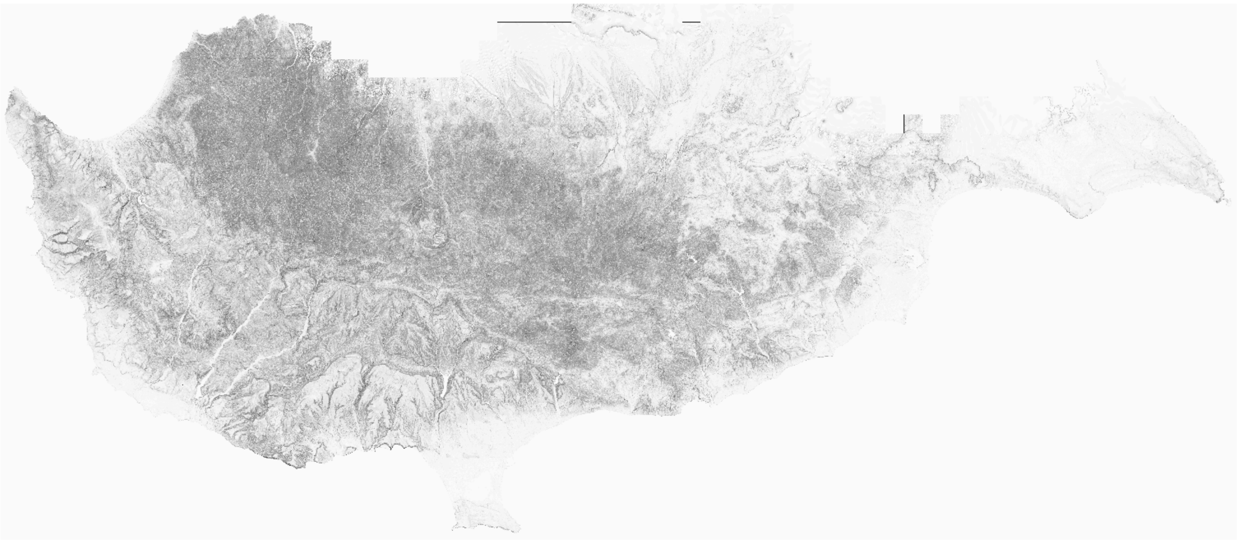


Figure 19: Raw Slope layers (colour inverted)

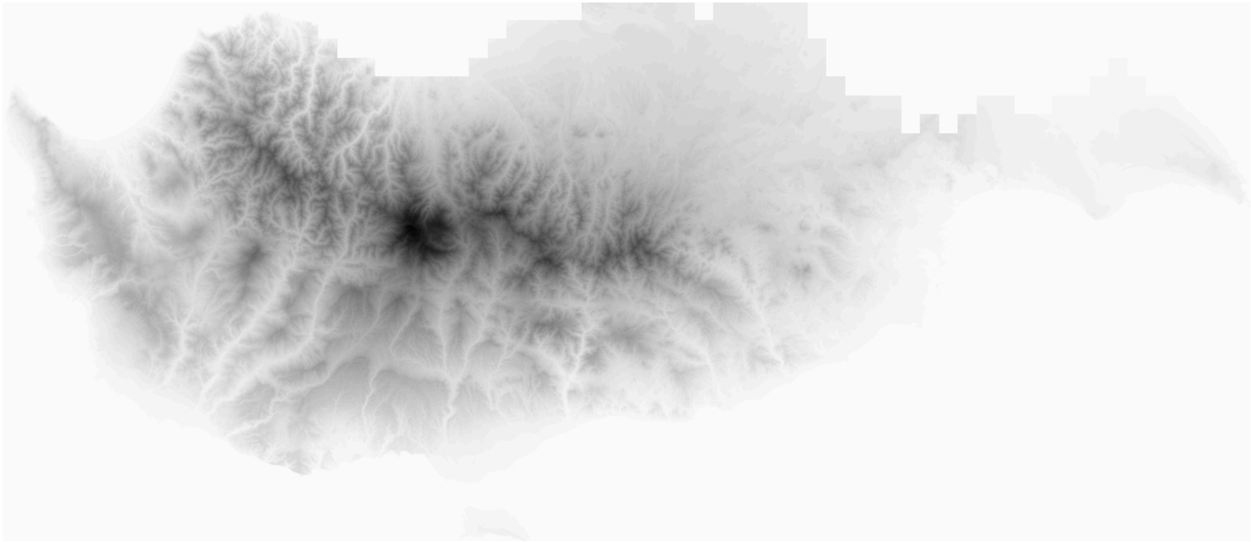


Figure 20: Raw Elevation layers (colour inverted)

From the metadata included with these packages, it was clear that each came at a resolution of 5m^2, with data ranging from 360 degrees for aspect, 87 degrees for slope, and up to 1934 meters for elevation, the latter being a continuous variable. If the Tree Class data could be aligned with these layers it would permit accurate analysis, as each Tree Class cell could contain exactly the same subset of geomorphological data.

As a consequence of the border situation in Cyprus, data accuracy to the north-eastern end of the country appeared limited, as evident from the regular patterns in figure 18. Further, in flat areas aspect data also showed regular patterns. Fortunately, since these correspond to areas with no slope, this could be accounted for. Metadata indicates that all these layers are known to represent the underlying data at extreme accuracy. The CRS and custom projection used for these data is the same as is used in the GAEA tool, a variation on “ETRS89 / UTM zone 36N (N-E)”.

6.2.3 Soil Data

Soil type and depth were provided as vector layers in the Shapefile format. Soil type is classified by the relative concentration of clay, silt, sand, and gravel present, and covers 9 classes distributed across non-intersecting irregular polygons to comprise the dataset. Figure 21 shows this classification and the extent of the data.

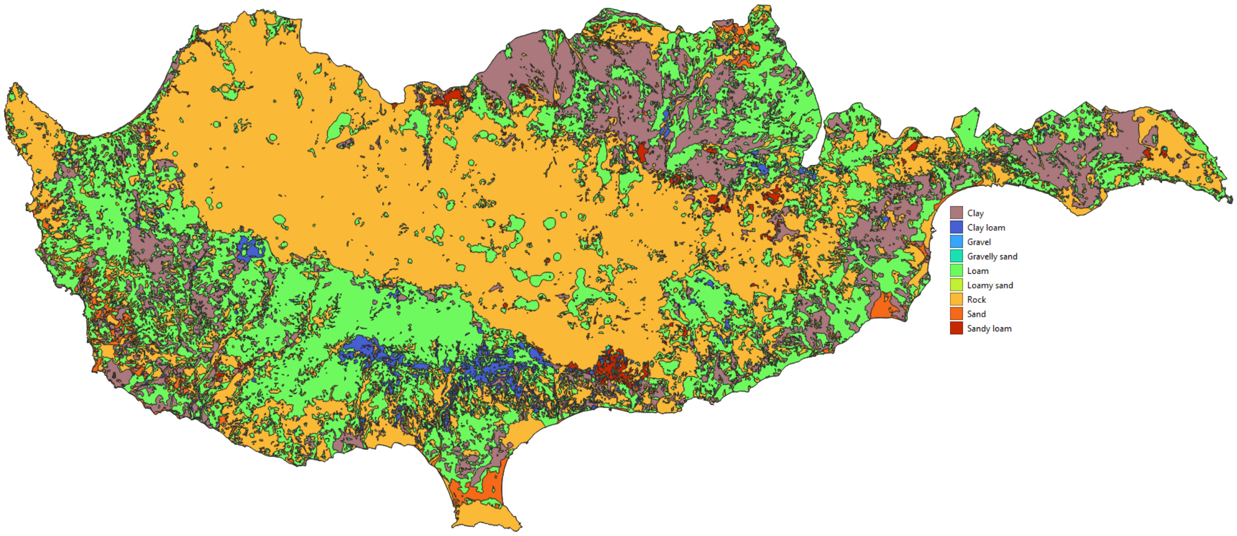


Figure 21: Soil Type class map, including legend

Soil depth refers to how deep into the soil water and nutrients are available for vegetation. Soil depth was classified into bins of 5, 17, 30, 37, 50, 62, 87, and 120cm, coded in figure 22 from light to dark blue. These data were provided in the "WGS-84" CRS format. Both sets of soil data originate from government surveys, and the extent is thereby limited.

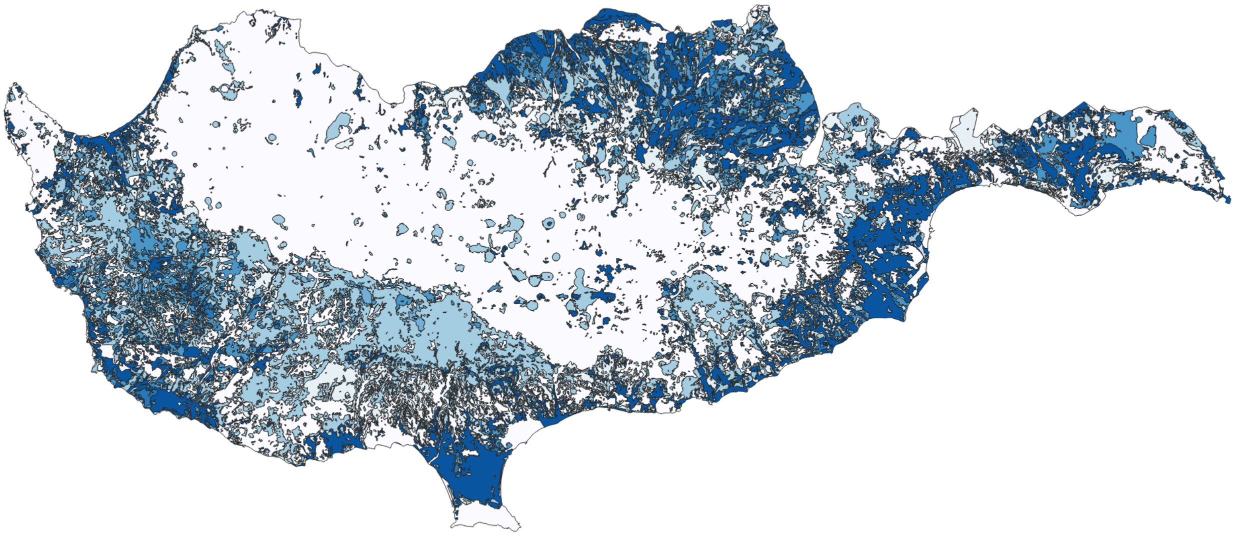


Figure 22: Soil Depth class map

6.3 Data Processing

6.3.1 Processing the Tree Class data

With all data imported, an ultimately lengthy task of processing, reconciling and aligning the various datasets could begin. First, the tree data had to be combined into a single table, comprising nearly 5 million rows, while culling patchID data and combining "Actual/Predicted" Tree Class data. Initially, such as a data set is imported as a "Delimited Text Layer", which must be asigned a CRS to be represented as a point layer, wherein the coordinates contained within must be assigned to some geometry. Without providing instructions on how to convert these data into polygons, it would not be possible to produce a vector layer, which is the condition for rasterizing such that we can ultimately compare our data.

By overlaying the initial point layer over the aspect raster as illustrated in figure 23, it became evident that there existed a significant amount of overlap at the edges of the various patches of data. After this layer could be converted into vector, the overlap would have to be resolved in such a way that the integrity of the data is maintained.

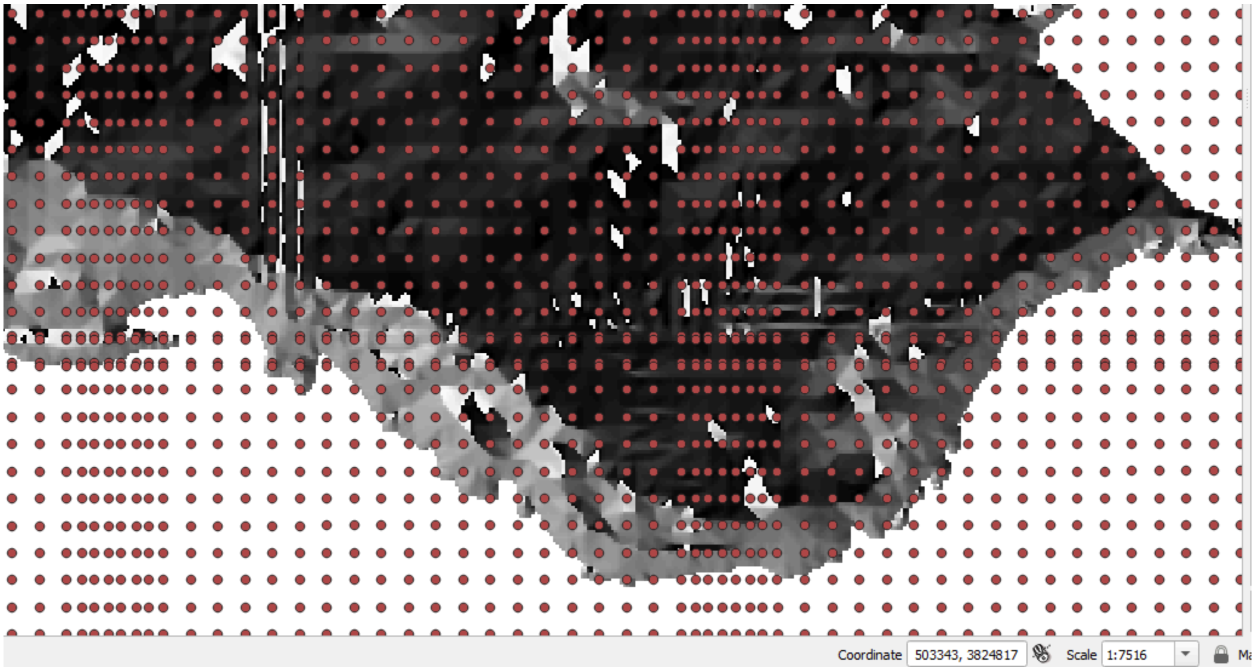
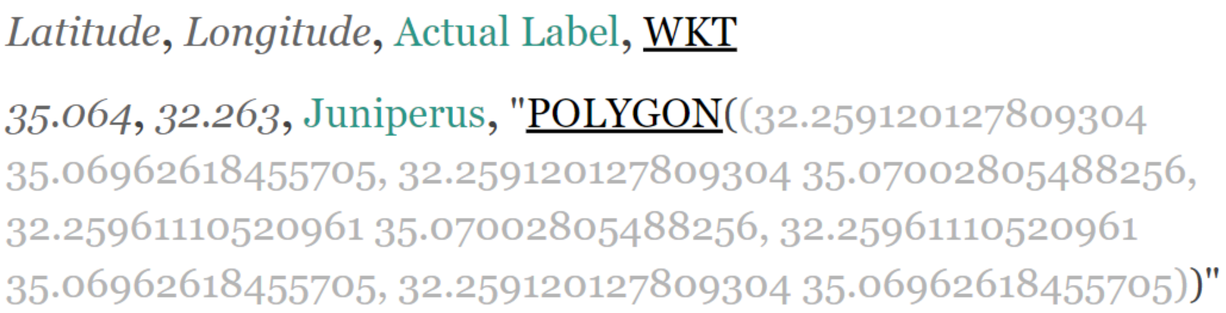


Figure 23: Tree Class data as dot layer over Aspect raster

In order to vectorize the data, it was found that the bounding box coordinates had to be converted into the Well-Known Text (WKT) format. Figure 26 depicts this conversion:



Latitude, *Longitude*, Actual Label, WKT

35.064, *32.263*, Juniperus, "POLYGON((32.259120127809304 35.06962618455705, 32.259120127809304 35.07002805488256, 32.25961110520961 35.07002805488256, 32.25961110520961 35.06962618455705, 32.259120127809304 35.06962618455705))"

Figure 24: Polygon encoding via WKT format

After vectorizing the Tree Class data, a sample patch was further rasterized to to observe the resulting pattern, so that with the aid of the client these issues could be resolved. Figure 25 depicts the overlay of one such patch over the Tree Class vector layer, where it becomes clear that it is well aligned, but overlapped. Much time was spent in this stage resolving these issues, until the combination of a consistent method for reprojecting to a shared CRS and rasterizing using the greatest overlap over each cell allowed this issue to be resolved.

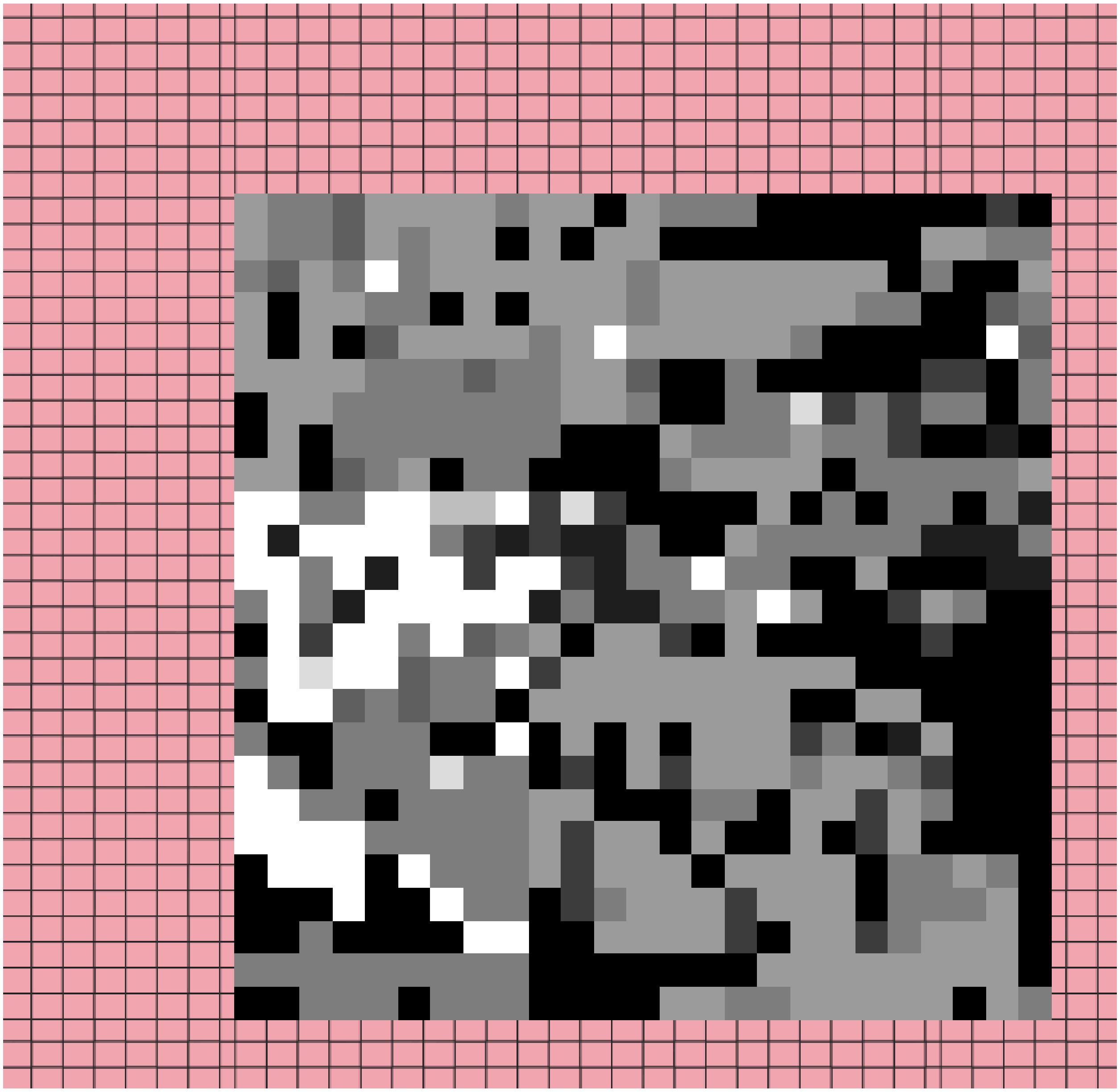


Figure 25: rasterized patch over combined Tree Class vector

Next, alignment with the underlying geomorphological raster data layers was sought. In this process, it was discovered that the Tree Class data has a resolution of roughly ~44.6-44.8m^2, explaining why the process was stuck. Fortunately, by creating a 45m^2 vector grid aligned with our underlying data, the Tree Class data could be filtered into this grid, introducing but a small margin of error, which in consulting with the client representative deemed unlikely to significantly affect analysis. This grid layer now containing the Tree Class data was aligned, and since 45m is a multiple of 5m it ensures that 81 data points of underlying raster layers become associated with that Tree Class data. Figure 26 illustrates this property:

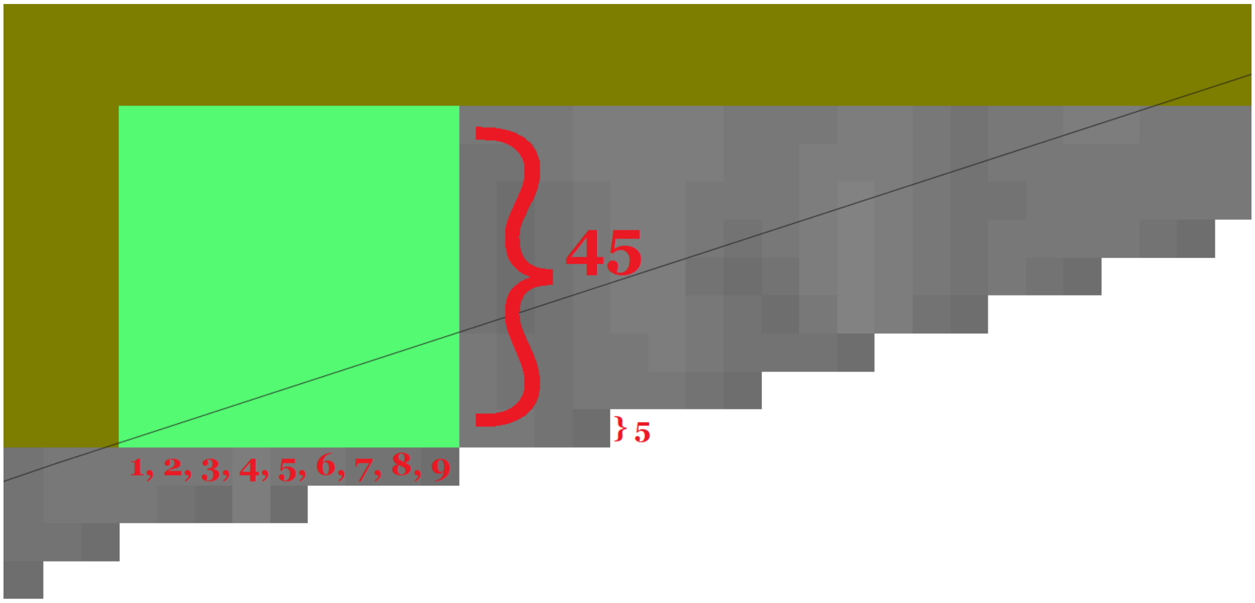


Figure 26: perfect raster alignment at 9:1 scale

6.3.2 Aligning the Data

In order to perform meaningful correlational analysis, the area of study should be limited to one where there is full overlap of all layers studied. Since the soil layers are of the smallest extent, they were used as the cut-off for the Tree Data, which would itself only be compared to the data underneath it, which was also clipped by this same process. Figures 26-29 demonstrate this process visually, which was achieved by clipping. Fortuitously, this process culls only unwanted data, such as tree classifications in the sea, or the aforementioned northeastern aspect data, corresponding to the urban area of Nicosia.

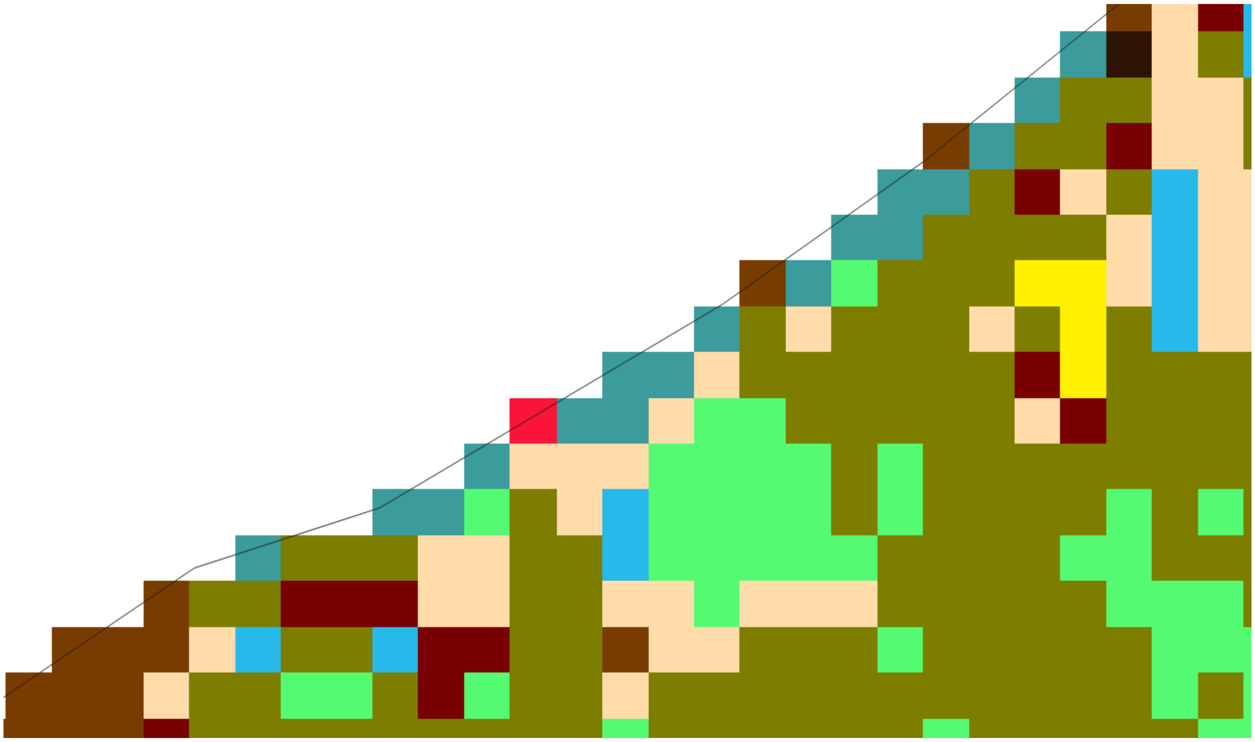


Figure 27: Clipped Tree Class raster

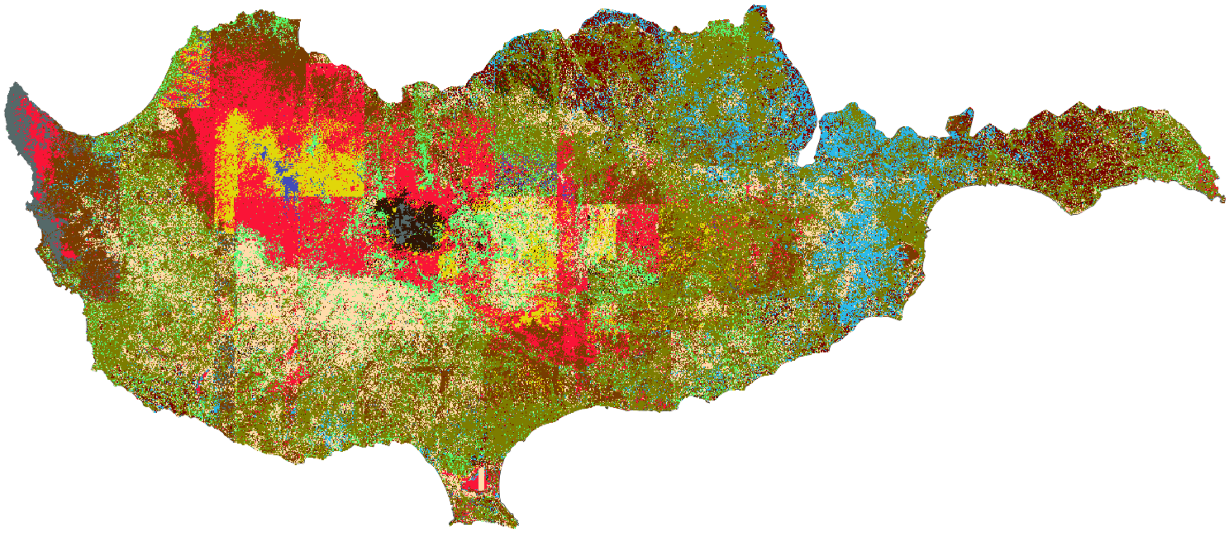


Figure 28: Clipped Tree Class Layer (All Classes)

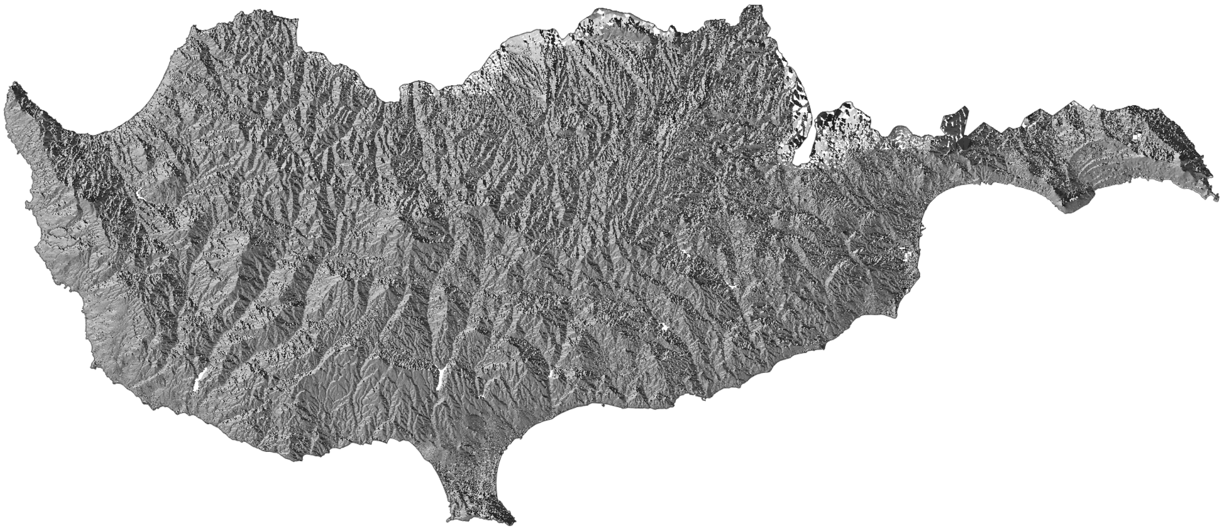
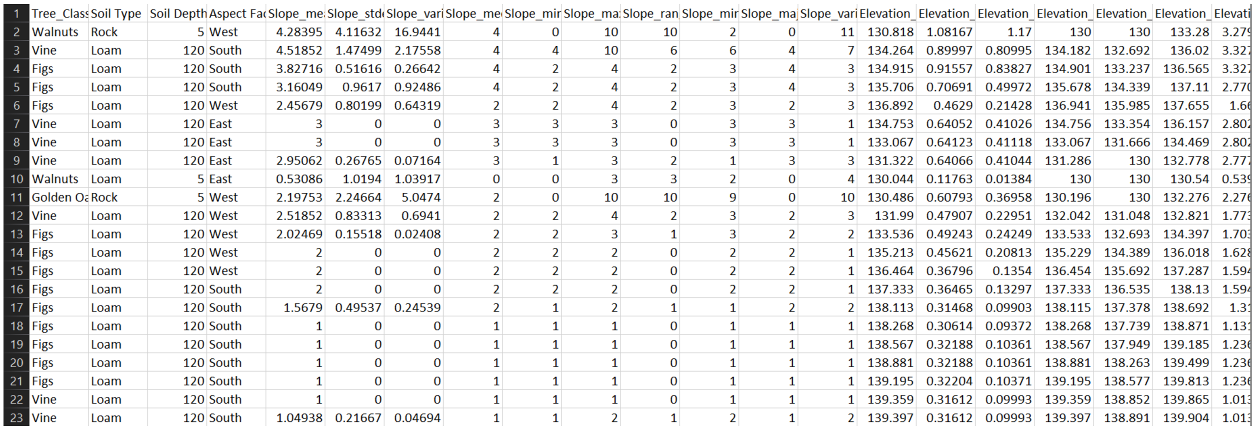


Figure 29: Clipped Aspect Layer

6.4 — Data Joining

In order to ready the data for statistical analysis, the soil layers were similarly filtered into the overlaying vector grid, and each of the five types of geodata were correlated based on their coordinate position, and had zonal statistics performed, providing summary statistics. These data could then all be assigned to an extensive table, such that correlative and other statistical analysis could be performed. Aside from the zonal statistics operation in QGIS, all these further joins were carried out via simple Python scripts utilizing Pandas. Data columns were also rearranged, renamed, and culled as needed, as ID or coordinate data was no longer necessary for further analysis outside QGIS. Figure 30 shows the largest of the combined data table, including for each row of known Tree Class the soil type, soil depth, aspect facing, and for slope, elevation, aspect, the {mean, std deviation, variance, median, min, max, range, minority, majorit, variety}. Most analyses conducted took a subset of this table.



1	Tree_Class	Soil Type	Soil Depth	Aspect Fa	Slope_me	Slope_std	Slope_vari	Slope_me	Slope_mir	Slope_ma	Slope_ran	Slope_mir	Slope_ma	Slope_vari	Elevation_	Elevation_	Elevation_	Elevation_	Elevation_	Elevation_	Elevati
2	Walnuts	Rock	5	West	4.28395	4.11632	16.9441	4	0	10	10	2	0	11	130.818	1.08167	1.17	130	130	133.28	3.279
3	Vine	Loam	120	South	4.51852	1.47499	2.17558	4	4	10	6	6	4	7	134.264	0.89997	0.80995	134.182	132.692	136.02	3.327
4	Figs	Loam	120	South	3.82716	0.51616	0.26642	4	2	4	2	3	4	3	134.915	0.91557	0.83827	134.901	133.237	136.565	3.327
5	Figs	Loam	120	South	3.16049	0.9617	0.92486	4	2	4	2	3	4	3	135.706	0.70691	0.49972	135.678	134.339	137.11	2.770
6	Figs	Loam	120	West	2.45679	0.80199	0.64319	2	2	4	2	3	2	3	136.892	0.4629	0.21428	136.941	135.985	137.655	1.66
7	Vine	Loam	120	East	3	0	0	3	3	3	0	3	3	1	134.753	0.64052	0.41026	134.756	133.354	136.157	2.802
8	Vine	Loam	120	East	3	0	0	3	3	3	0	3	3	1	133.067	0.64123	0.41118	133.067	131.666	134.469	2.802
9	Vine	Loam	120	East	2.95062	0.26765	0.07164	3	1	3	2	1	3	3	131.322	0.64066	0.41044	131.286	130	132.778	2.777
10	Walnuts	Loam	5	East	0.53086	1.0194	1.03917	0	0	3	3	2	0	4	130.044	0.11763	0.01384	130	130	130.54	0.539
11	Golden Oa	Rock	5	West	2.19753	2.24664	5.0474	2	0	10	10	9	0	10	130.486	0.60793	0.36958	130.196	130	132.276	2.276
12	Vine	Loam	120	West	2.51852	0.83313	0.6941	2	2	4	2	3	2	3	131.99	0.47907	0.22951	132.042	131.048	132.821	1.773
13	Figs	Loam	120	West	2.02469	0.15518	0.02408	2	2	3	1	3	2	2	133.536	0.49243	0.24249	133.533	132.693	134.397	1.703
14	Figs	Loam	120	West	2	0	0	2	2	2	0	2	2	1	135.213	0.45621	0.20813	135.229	134.389	136.018	1.628
15	Figs	Loam	120	West	2	0	0	2	2	2	0	2	2	1	136.464	0.36796	0.1354	136.454	135.692	137.287	1.594
16	Figs	Loam	120	South	2	0	0	2	2	2	0	2	2	1	137.333	0.36465	0.13297	137.333	136.535	138.13	1.594
17	Figs	Loam	120	South	1.5679	0.49537	0.24539	2	1	2	1	1	2	2	138.113	0.31468	0.09903	138.115	137.378	138.692	1.31
18	Figs	Loam	120	South	1	0	0	1	1	1	0	1	1	1	138.268	0.30614	0.09372	138.268	137.739	138.871	1.131
19	Figs	Loam	120	South	1	0	0	1	1	1	0	1	1	1	138.567	0.32188	0.10361	138.567	137.949	139.185	1.236
20	Figs	Loam	120	South	1	0	0	1	1	1	0	1	1	1	138.881	0.32188	0.10361	138.881	138.263	139.499	1.236
21	Figs	Loam	120	South	1	0	0	1	1	1	0	1	1	1	139.195	0.32204	0.10371	139.195	138.577	139.813	1.236
22	Vine	Loam	120	South	1	0	0	1	1	1	0	1	1	1	139.359	0.31612	0.09993	139.359	138.852	139.865	1.013
23	Vine	Loam	120	South	1.04938	0.21667	0.04694	1	1	2	1	2	1	2	139.397	0.31612	0.09993	139.397	138.891	139.904	1.013

Figure 30: full results tabel first rows (All_data_for_stats.csv)

6.5 — Statistical Analysis Results

Various analyses were conducted throughout the course of this project; especially correlative analysis between Tree Class and the other five data dimensions and their zonal statistics. This section shall cover those results which were of potential relevance to the data story. As discussed earlier, stock standard functions for these tests were employed from the Pandas library of Python, as the format of these data well fit dataframes.

It is to be kept in mind that a value such as the 'Slope mean' of an individual row represents the average of the 81 slope instances integrated into that cell, now row of data. By taking the average of those averages can we represent the entire data set, it is essentially equivalent to binning the original slope data, and averaging over those bins. 2,936,160 such rows exist in the full table.

6.5.1 Descriptive Statistics

Tree Class data was first correlated with each of the geomorphological characteristics, as captured in the boxplots of figures 31-33. These follow the standard of a box ranging from Q1-Q3 with Q2 at the median and whiskers at 1.5*IQR, with outliers noted separately. data story figures (8, 9) display this same data rearranged by grouping.

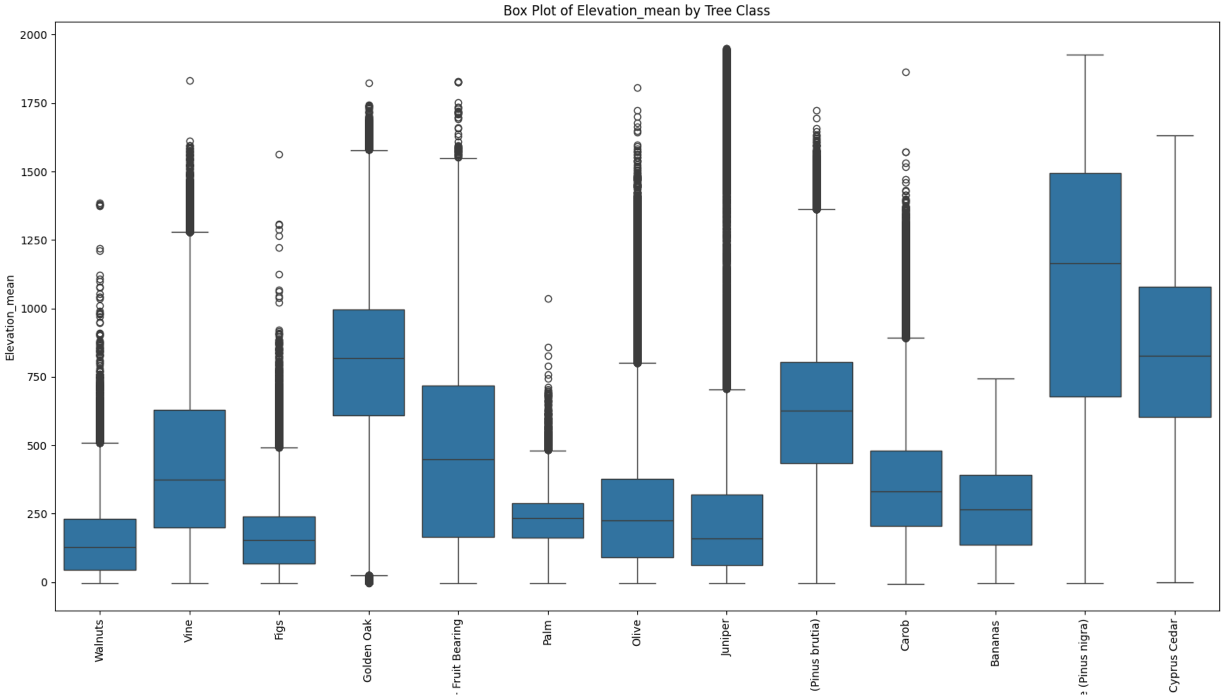


Figure 31: mean Elevation by Tree Class

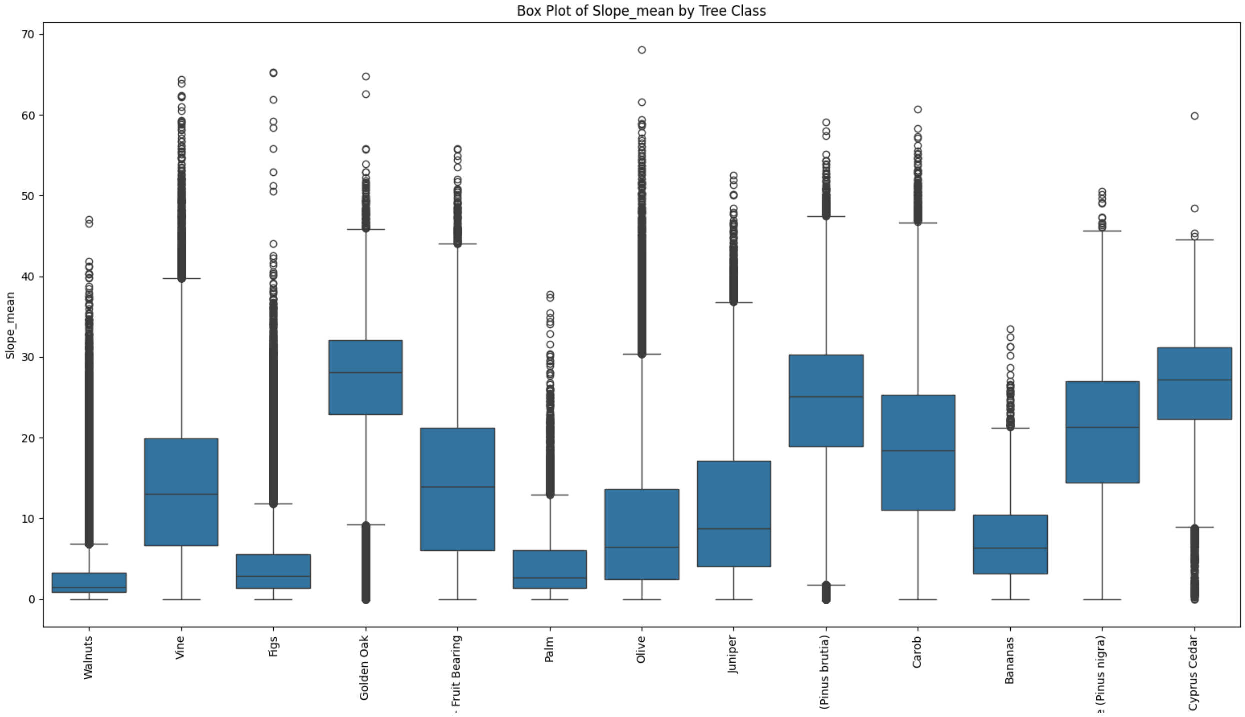


Figure 32: mean Slope by Tree Class

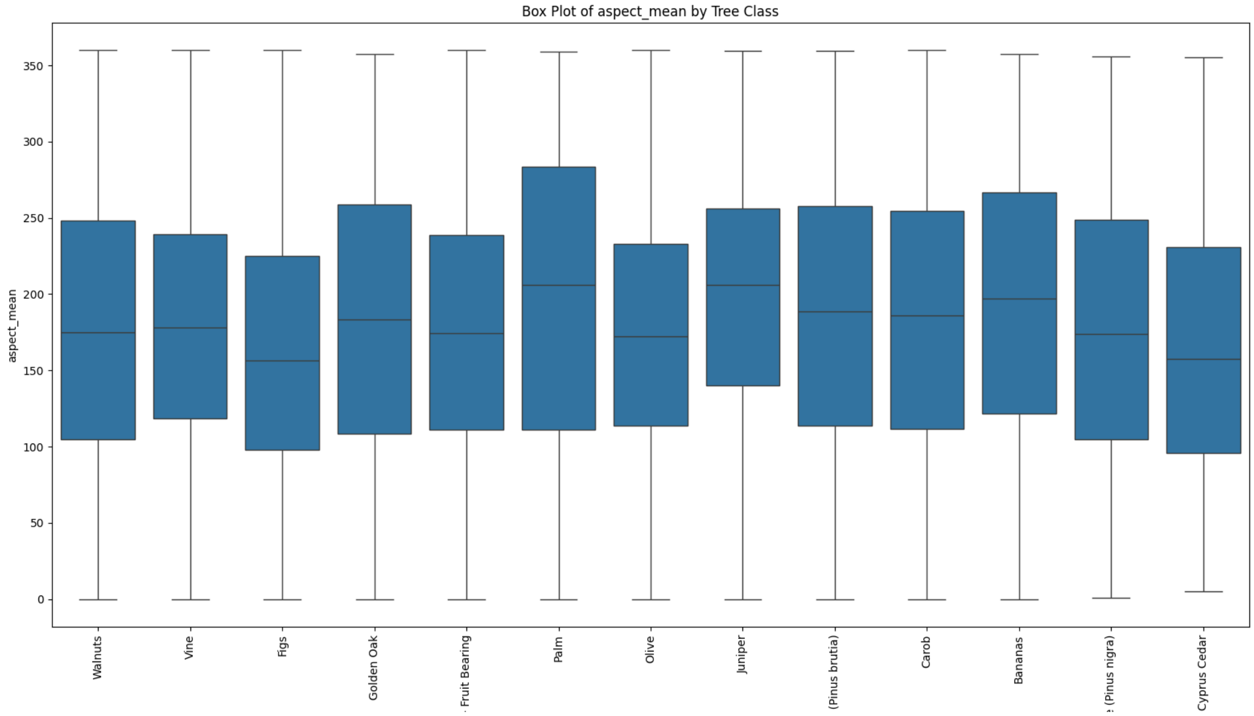


Figure 33: mean Aspect by Tree Class

6.5.2 Aspect Facings

The 360 degrees of aspect data may be converted to Aspect Facings corresponding to the cardinal directions by a simple formula: if (aspect >= 0 and aspect <= 45) or (aspect >= 315 and aspect <= 360): return 'North'. Else if aspect > 45 and aspect <= 135: return 'East', etc.

This categorization creates four 4 bins of high count for aspect data of each Tree Class to fit in, see: [Appendix.II.1, pg. 67]. This data may then be normalized into percentages for visual representation, such as in data story figure (16), or its full version in figure 34 below.

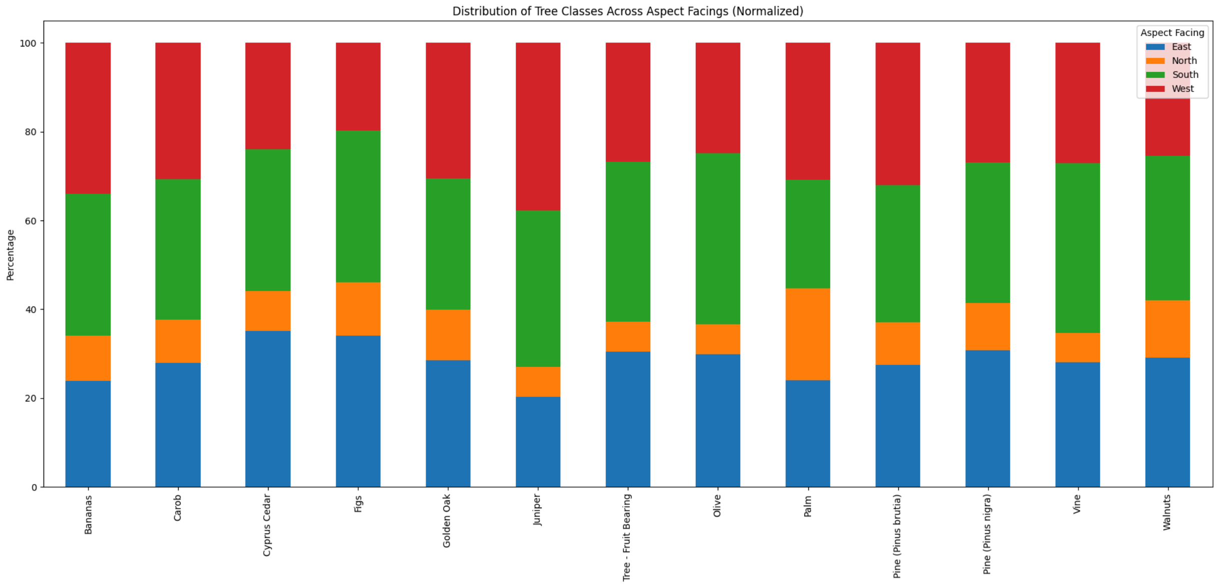


Figure 34: Distribution of Tree Class Across Aspect Facings (Normalized)

As detailed in the data story, Chi² tests were conducted to determine if the correlation between Aspect Facings and Tree Classes observed were strongly correlated, and they were. This test should be possible because the data consists of two categorical variables with high counts relating them. See [Appendix.II.2, pg. 67-69] for the table of individual results.

6.5.3 Soil Type

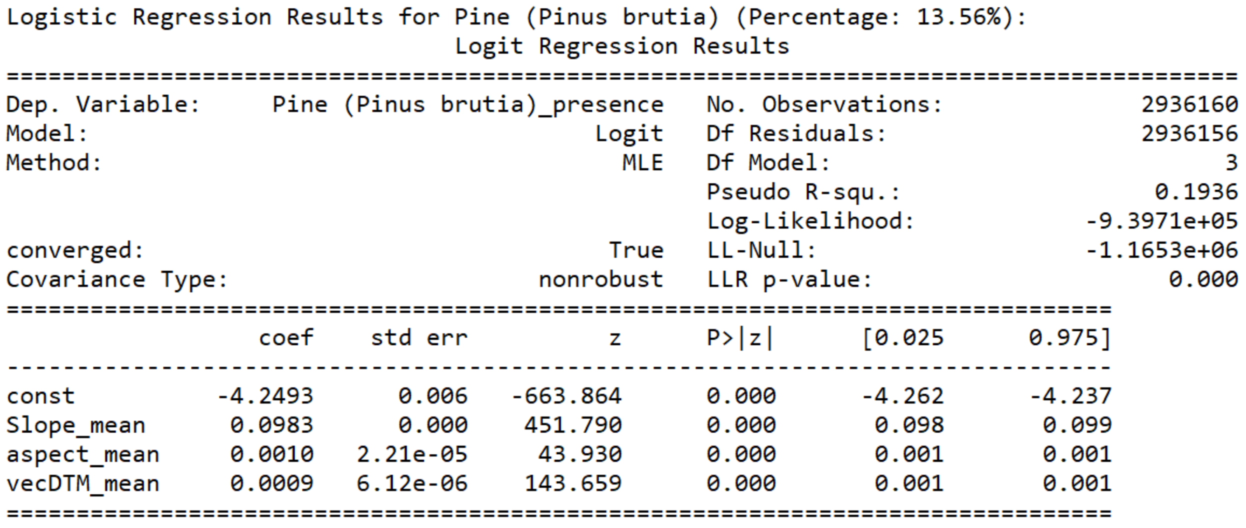
Soil type data is also straightforwardly categorical, and can be similarly tabulated; see: [Appendix.II.3, pg. 69-70]. Chi² test again yields extreme significance: (Chi² :710203, df = 96, p < .00001).

6.5.4 Confusion Matrices

Via the merged Tree Class data before it was further preprocessed for QGIS, it is possible to plot the "Predicted Label" vs the "Actual Label" in a confusion matrix as described in data story figures (13, 14), via matplotlib using colour remapping to ensure contrast. These results were not saved separately, as they were already available to the client.

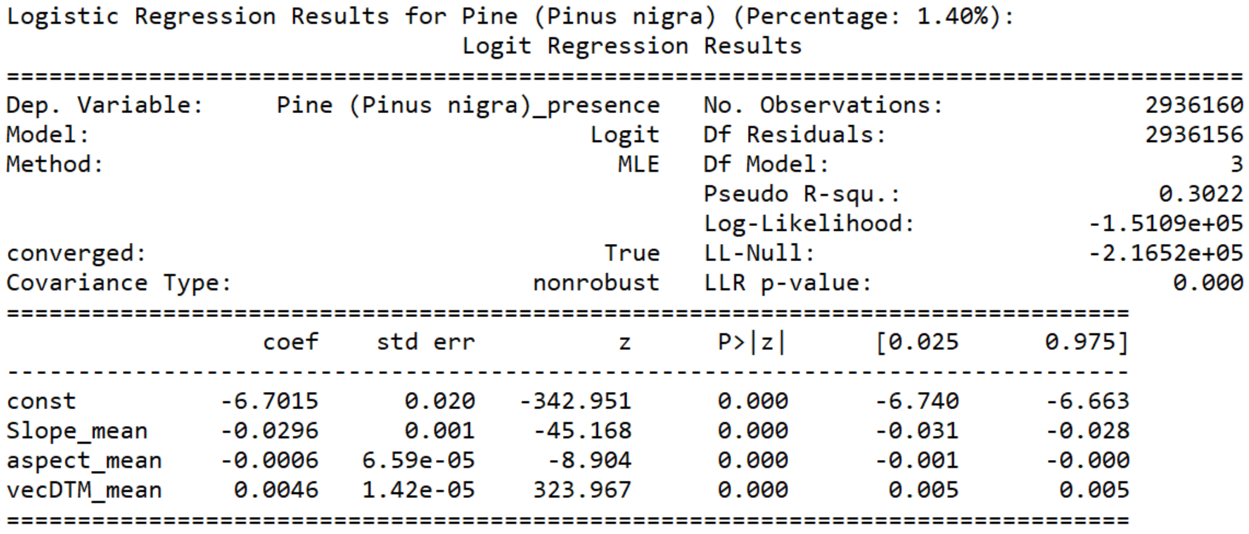
6.5.5 Logistic Regression

If Tree Class is treated as a categorical dependent variable and Slope, Aspect, Elevation means as continuous independent variables, the potential direction of their relationships can be tested and it may be estimated if these relationships are significant. Using the statsmodels.api.Logit function, logistical regression analysis was performed on these combinations. As figures 35, 36 indicate below, the results were highly significant. Only Bananas v. Elevation did not render significant results. For all other combinations of Tree Class and underlying geomorphological data the relationships plotted in figures 31-33 are quantified, as the resulting coefficients give indication of the strength of each relationship. Misc results are available in [Appendix II.4, pg. 70-75].



```
Logistic Regression Results for Pine (Pinus brutia) (Percentage: 13.56%):
                           Logit Regression Results
==============================================================================
Dep. Variable:     Pine (Pinus brutia)_presence   No. Observations:            2936160
Model:                                    Logit   Df Residuals:                2936156
Method:                                     MLE   Df Model:                          3
                                                  Pseudo R-squ.:              0.1936
                                                  Log-Likelihood:         -9.3971e+05
converged:                                 True   LL-Null:                -1.1653e+06
Covariance Type:                      nonrobust   LLR p-value:                  0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -4.2493      0.006   -663.864      0.000      -4.262      -4.237
Slope_mean     0.0983      0.000    451.790      0.000       0.098       0.099
aspect_mean    0.0010   2.21e-05     43.930      0.000       0.001       0.001
vecDTM_mean    0.0009   6.12e-06    143.659      0.000       0.001       0.001
==============================================================================
```

Figure 35: Logit results for Brutia Pine v. {Slope, Aspect, Elevation}_mean



```
Logistic Regression Results for Pine (Pinus nigra) (Percentage: 1.40%):
                           Logit Regression Results
==============================================================================
Dep. Variable:      Pine (Pinus nigra)_presence   No. Observations:            2936160
Model:                                    Logit   Df Residuals:                2936156
Method:                                     MLE   Df Model:                          3
                                                  Pseudo R-squ.:              0.3022
                                                  Log-Likelihood:         -1.5109e+05
converged:                                 True   LL-Null:                -2.1652e+05
Covariance Type:                      nonrobust   LLR p-value:                  0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -6.7015      0.020   -342.951      0.000      -6.740      -6.663
Slope_mean    -0.0296      0.001    -45.168      0.000      -0.031      -0.028
aspect_mean   -0.0006   6.59e-05     -8.904      0.000      -0.001      -0.000
vecDTM_mean    0.0046   1.42e-05    323.967      0.000       0.005       0.005
==============================================================================
```

Figure 36: Logit results for Black Pine v. {Slope, Aspect, Elevation}_mean

6.6 — Data Visualization: Maps

Aside from the analyses ran outside of QGIS, the tool itself provides ample means for visualization in the form of maps, as evident from data story figures (6, 7, 10-12, 15). By excluding specific parts of a data set, such as specific Tree Classes, and by adjusting the opacity of remaining data, it is possible to layer multiple datasets on top of each other in a manner that implies correlation. This implication is powerful by itself, but becomes informative only when matched with further analysis of the data presented.

The number of combinations and variations of these data possible to fit on a map is endless, while the space in a story, and even a thesis, is ultimately limited. 60 maps were exported and considered for potential inclusion as the story took shape, of which only those that fit the emergent story best were ultimately retained. It is nonetheless possible that some of these maps contain hints for future research, and so a subset of these are included in [Appendix I, pg. 62-66].

6.7 — Determining the Story

As results come in it became evident that there were somewhat distinct groupings of Tree Classes present in the data set, as illustrated by data story figure (9). Though it was attempted to head into analysis with as few preconceived notions as possible, the client representative had mentioned that there was likely to be some interaction between the two types of pines, as the Cyprus Department of Forests had expressed preliminary concerns that black pine could be encroaching on the habitat of brutia pine. These two Tree Classes were both present in the most interesting grouping, and thereby became the main focus of the data story. Since the statistical results pertaining to these trees were significant it was possible to maintain this course in the latter part of the realisation phase.

With further feedback from the client representative it was determined best if the data story would be of interest both to an informed citizen looking to better understand the environment of Cyprus, as well as potentially to policy makers, as the results available could be relevant to preservation efforts.

6.8 — Evaluation

When a CreaTe design artefact is meant to meet specific end-user needs beyond what the client representative can verify, user evaluations may be carried out to determine if the product is meeting intended goals. In this case, it was determined in concert with the client representative that this step was of diminished relevance, and should only be carried out if an abundance of time was available, which ended up not being the case. The nature of this project has provided ample opportunity for the types of analysis one would carry out on survey results.

When considering if the functional and non-functional requirements outlined in the specification phase were met, it seems evident that those requirements relating to the data and research were. The non-functional requirements for the data story require further examination to determine if the present story meets these needs:

- The patterns presented are novel
- The patterns presented could lead to future research

This set of Tree Class data is exceptionally new, having to the writer's present understanding been based in part on recent lessons learned together with S. Karatsiolis et al. in automating annotation of landsat data [16], and having only been become available with the specification phase of this project. This indicates that significant patterns found are likely novel, potentially enabling future research. The remaining set of non-functional requirements are harder to judge, as they are ultimately subjective in a way impervious to the writer:

- Presents results in such a way that they are clear to the average informed reader, and potentially useful or informative to the subset of readers with a specific interest in the data and characteristics covered.
- The visualizations chosen are easy to read and understand to those familiar with the mode of representation.

It is here that a survey seeking to measure the reader's knowledge of the topics covered before and after reading the data story would have been helpful, together with any persistent feedback regarding the way the data is presented. Further, had the data story been completed earlier the present rendition could have been discussed with the client representative, which would undoubtedly have resulted in actional feedback. The way each person processes and presents information may be particular to them, a data story must be iterated on if the writer deviates significantly from the expected norm.

Chapter 7 — Discussion

7.1 — Limitations

In order to properly contextualize the results and findings presented herein it is important to first consider any applicable limitations. In dividing this project into the needs of data analysis and data story telling, the first half of the project focused partially on data story telling and data selection, while much time was spent on background research not relied upon much later on. A great amount of time was spent on enabling and performing analysis in the latter half of the project, preventing additional feedback and validations steps on the final story. In retrospect, ensuring a great familiarity with QGIS and a rapid formulation of data required would have freed up a lot of research time, enabling further iteration. Though this emphasis on the data is not without result, these results could likely have been better communicated given this further iteration. The large size and complexity of the data also frequently resulted in required operations taking many hours to compute, evading nearly every attempt at optimization. By choosing to minimize loss of quality in the data in processing a great many hours of extra computing time were necessary.

Whilst much was learnt in the conduct of this project, in order to fully match these findings with existing knowledge on the complex interactions between the Tree Classes studied and the terrain in which they are embedded, further domain expertise and embodied knowledge about these ecosystems is required.

The Tree Classification data studied and correlated in this project is derived from an algorithm still very much in development, of which the exact specifications were unknown. Though it is clear that it assigned classifications based on whichever class of tree best fit its understanding, and that these were properly assigned to specific sets of coordinates, since it is required to assign something in all cases it may produce misleading classifications on a local level when made to classify a patch of sea or a densely urban area. Further, though the known prediction accuracy is impressive, it is not known how many false classifications remained unspotted. By concentrating research on those species predominant in the mountainous forested area in the center of the country, these issues were somewhat mitigated. Given the large number of cells classified and the high resolution of the underlying data layers, if classification errors are somewhat uniformly distributed then by the law of large numbers the impact on the validity of results may be minimized.

This same argument may hold for the filtration process whereby the raw Tree Class data was matched to a regular $45m^2$ grid, a minor uniform shift is likely evened out across space. This data resolution is limiting however, as smaller patches of trees surrounded by another species may not be spotted. Straight line border artefacting across the Tree Class data indicate that there is some form of variance in classification logic between patches, causing some patches to see no classification of a certain type of tree. This is surprising as the distribution surrounding these areas would make it seem likely that some pixels would be classified with the missing tree type. See [Appendix I, pg. 66] for examples.

The dimension of time is also still missing, as without being able to consider this dataset as it develops over the years we cannot yet get a full picture of how the relative distribution of each species is progressing. Likewise, since the the available resolution is presently bound by what landsat data is available to classify, the effectiveness of classification itself is limited by the quality of images available to classify. Finally, though this dataset hosts a great quantity of data, which may mitigate some of the effects discussed here, the Tree Classes of Palm and Banana especially had low sample sizes relative to the rest, and so no general claims about these data should be taken to apply to these with the same certainty as might be ascribed elsewhere.

7.2 — Interpreting Results

Both the data story produced as well as the analysis from which it was derived may be considered as the results of this project. Though the ultimate objective has always been to produce this data story, the central question of *how* these data could lead to an interesting and insightful data story has produced a set of results which should not be considered as only secondary to this main goal. If the limitations discussed do not invalidate these data, both the research findings and any insights derived from the data story may prove to be of future utility.

Firstly, the descriptive statistics derived from the combined dataset implied that there may be significant variance in the “preferences” of each Tree Class with respect to the elevation, slope, aspect, aspect facing, soil type, and soil depth that they grow in with respect to other Tree Classes. Logistic regression analysis together with Chi^2 tests appear to confirm this strongly, with exceptionally significant results showing these data correlate, showing that for each pairing of Tree Class to any other single data class there is some firm non-random relationship waiting for a causal explanation.

The use of corrected labels where available, together with strong implied initial accuracy of the provided labels, further underscores the likely scenario that these results will hold true. The confusion matrix from this result is derived incidentally also indicates where classification is most difficult.

By focusing on the Black and Brutia Pines in the data story, not only was the relevance of these findings underscored, but an appearantly new pattern was found in their distinctive soil preferences. As interpretation of the factors involved in causing these relationships was regarded as best left to others, the data story also focused on relaying correlations, only referring to outside research to point out a straightforward confirmation of existing knowledge.

The data story itself is a microcosm of the explorative method employed throughout this project, aiming to be approachable to a wide range of readers, leading them through increasingly complex results to convey the most relevant findings accurately, without too many readers giving up part of the way.

7.3 — Implications

Given that the tests conducted hold, and that the efforts taken to mitigate any limitations where possible were successful, there are a wide range of possible implications from this research. Someone who lives within the area studied may now be more informed about the conditions of their natural environment, may experiment when planting these trees with conditions where those trees are plentifully found.

By considering what the data story relays about the relationship between the two pines, a policy maker interested in the conservation of either species, or in the halting of its spread to unwanted habitats, can consider the differences in conditions preferred by each tree, and note for example the outsized preference Black Pine seems to have for clay and loam.

Even by just considering the spatial distributions of these tree classes, rare occurrences of particular trees may be identified, and culled if deemed undesirable. By looking at common trees such as Vine, any encroachment of agricultural flora on protected areas may be studied.

When projects such as this are able to validate correlations in data incorporated into an Environmental Digital Twin such as GAEA, the purpose of these tools as representations of the underlying physical reality is reinforced. The methods and processes described within this paper can be used as general guidelines for those who seek to undertake similar projects.

It is prudent to be conservative when estimating the direct implications of this type of work, as almost any consequence one can think of will soon imply future work.

7.4 — Future Work

It would be very interesting to further improve the data story by processing more feedback and running surveys. By continuing to iterate data stories on these data more of these patterns can be clearly communicated to those who would be interested in knowing about them. An effective transfer of information that is engaging and non-discouraging, with a low attrition rate this data story and others like it can be made more informative by simply reaching more readers, and retaining those that begin reading better.

Different data stories could be produced from the same data, focusing on different combinations of trees for different audiences. The great numbers of Olive and Vine for example indicate that a group of readers who cultivate these is likely out there, and could be well served by a focused story which better incorporates existing literature on optimizing growth conditions for these trees.

Specific significant results, such as the strong specific preference for particular Aspect Facings provide claims which can be tested in future research, and which may support existing hypotheses not considered within the scope of this work.

If new, higher resolution landsat images become available, and these are further classified, many of the limitations discussed will be mitigated. By enabling the dimension of time those parties interested in the preservation and management of the trees of Cyprus will suddenly have an extremely useful set of data. Any future confirmation of the claims and data laid out in this project also increases the value of the existing data.

The datasets produced in this work may serve as a jumping off point for further research. Tree Class data could be correlated with multiple other data dimensions, whereupon unique combinations of these data could be found and studied on the ground.

This project focused on analysing correlations between Tree Classes and other data, but correlations between these other data classes can also be studied, and their correlation may reveal more about the influence of each of these data dimensions on the overal picture. [Appendix II.5, pg. 76] contains a correlation matrix which was not pursued further in the scope of this project, but which may be a starting point to this end.

Chapter 8 — Conclusion

This thesis has sought to demonstrate how the data from an Environmental Digital Twin can be selected, processed, and analysed in order to produce data stories relevant to readers. An explorative technique to data story creation was employed, and the many data dimensions available via GAEA were considered based on factors resulting from review. The essential tools for this, consisting of QGIS and Python were identified, and requirements for both the story and the data from which it would be derived were generated. By exchanging ideas with the client representative early and frequently a vision could be established toward which research could be conducted. By carefully processing data and using common statistical techniques, various significant results came out of this research, which could form the foundation for the data story. By leading the reader through a miniature version of this process, it is hoped that the conclusions presented in the final data story will stick with them, as they did for the researcher.

By showing a data story can be derived in this way from the first country-wide Environmental Digital Twin, it is demonstrated that so long as the data are accurate and made readily available in a digestible format, many insights can be gained by simply exploring the data. This combination of bountiful data and an effective means of digesting it shows imminent potential for lowering the barriers of entry, such that an interested citizen or policy maker can more much easily become informed about the true state of their environment.

References

"In the writing of this document, I used no generative artificial intelligence tools."

[1] L. S. Macarringue, É. L. Bolfe, and P. R. M. Pereira, "Developments in Land Use and Land Cover Classification Techniques in Remote Sensing: A Review," *Journal of Geographic Information System*, vol. 14, no. 01, pp. 1–28, 2022, doi:

[2] A. Jamil, Chirag Padubidri, Savvas Karatsiolis, I. Kalita, Aytac Guley, and A. Kamilaris, "GAEA: A Country-Scale Geospatial Environmental Modelling Tool: Towards a Digital Twin for Real Estate," Progress in IS, pp. 177–199, Jan. 2024, doi: https://doi.org/10.1007/978-3-031-46902-2_10.

[3] P. Palaiologou, K. Kalabokidis, M. A. Day, A. A. Ager, S. Galatsidas, and L. Papalampros, "Modelling Fire Behavior to Assess Community Exposure in Europe: Combining Open Data and Geospatial Analysis," ISPRS International Journal of Geo-Information, vol. 11, no. 3, p. 198, Mar. 2022, doi: https://doi.org/10.3390/ijgi11030198.

[4] C. E. Morales et al., "Earth Map: A Novel Tool for Fast Performance of Advanced Land Monitoring and Climate Assessment," Journal of remote sensing, vol. 3, Jan. 2023, doi: https://doi.org/10.34133/remotesensing.0003.

[5] X. Zhang, C. Ma, and G. Yang, "City appearance environment management system based on WebGIS," Third International Conference on Computer Science and Communication Technology (ICCSCT 2022), Dec. 2022, doi: https://doi.org/10.1117/12.2662219.

[6] L. Yang, J. Driscol, S. Sarigai, Q. Wu, H. Chen, and C. D. Lippitt, "Google Earth Engine and Artificial Intelligence (AI): A Comprehensive Review," Remote Sensing, vol. 14, no. 14, p. 3253, Jul. 2022, doi: https://doi.org/10.3390/rs14143253.

[7] S. Verma et al., "GeoEngine: A Platform for Production-Ready Geospatial Research," openaccess.thecvf.com, 2022. https://openaccess.thecvf.com/content/CVPR2022/html/Verma_GeoEngine_A_Platform_for_Production-Ready_Geospatial_Research_CVPR_2022_paper.html (accessed Apr. 15, 2024).

[8] M. Ignatius, N. H. Wong, M. Martin, and S. Chen, "Virtual Singapore integration with energy simulation and canopy modelling for climate assessment," IOP conference series. Earth and environmental science, vol. 294, no. 1, pp. 012018–012018, Jul. 2019, doi: https://doi.org/10.1088/1755-1315/294/1/012018.

[9] W. Weber, M. Engebretsen, and H. Kennedy, “Data stories. Rethinking journalistic storytelling in the context of data journalism,” Studies in Communication Sciences, vol. 18, no. 1, Nov. 2018, doi: https://doi.org/10.24434/j.scoms.2018.01.013.

[10] E. Segel and J. Heer, “Narrative Visualization: Telling Stories with Data,” IEEE Transactions on Visualization and Computer Graphics, vol. 16, no. 6, pp. 1139–1148, Nov. 2010, doi: https://doi.org/10.1109/tvcg.2010.179.

[11] A. Eftychiou, “CYENS SuPerWorld Research Group,” superworld.cyens.org.cy. https://superworld.cyens.org.cy/demo4.html.

[12] R. Castro-Salazar, M. Sacande, D. Maniatis, D. Mollicone. Ecosystem restoration as an immunization for humanitarian crisis: The case of Lake Chad. Georgetown Journal of International Affairs. 7 Aug 2020. Available: https://gjia.georgetown.edu/2020/08/07/ecosystem-restoration-as-an-immunization-for-humanitarian-crisis-the-case-of-lake-chad/

[13] A. H. Mader and Wouter Eggink, “A Design Process for Creative Technology,” University of Twente Research Information, pp. 568–573, 2014, Available: https://research.utwente.nl/en/publications/a-design-process-for-creative-technology

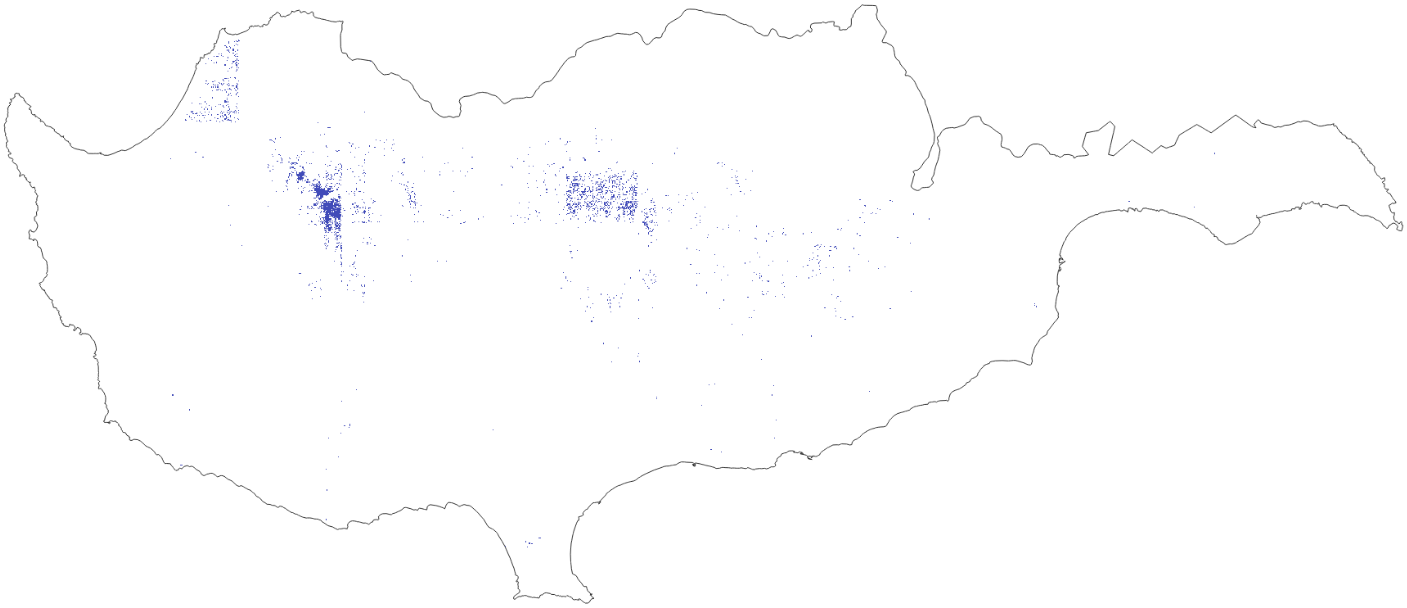
[14] K. A. Crawford and M.-A. Vella, “Cyprus Dataset: Settlements from 11000 BCE to 1878 CE,” Journal of Open Archaeology Data, vol. 10, 2022, doi: https://doi.org/10.5334/joad.96.

[15] P. Petrou et al., “Comparison of the Stand Structure Diversity of Open Pinus brutia Ten. Forests in Areas of Different Productivity in Central Cyprus,” Forests, vol. 14, no. 11, p. 2200, Nov. 2023, doi: https://doi.org/10.3390/f14112200.

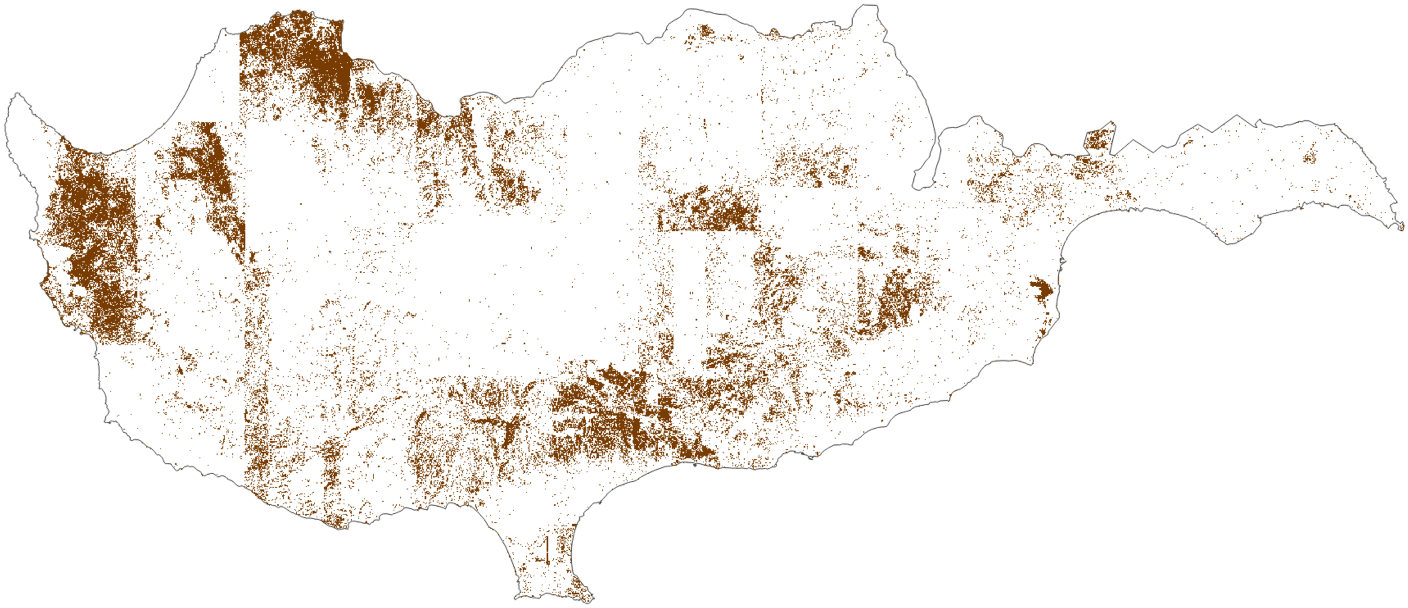
[16] S. Karatsiolis, C. Padubidri, and A. Kamilaris, “Scalable Retrieval of Similar Landscapes in Optical Satellite Imagery Using Unsupervised Representation Learning,” Remote sensing, vol. 16, no. 1, pp. 142–142, Dec. 2023, doi https://doi.org/10.3390/rs16010142.

Appendixes

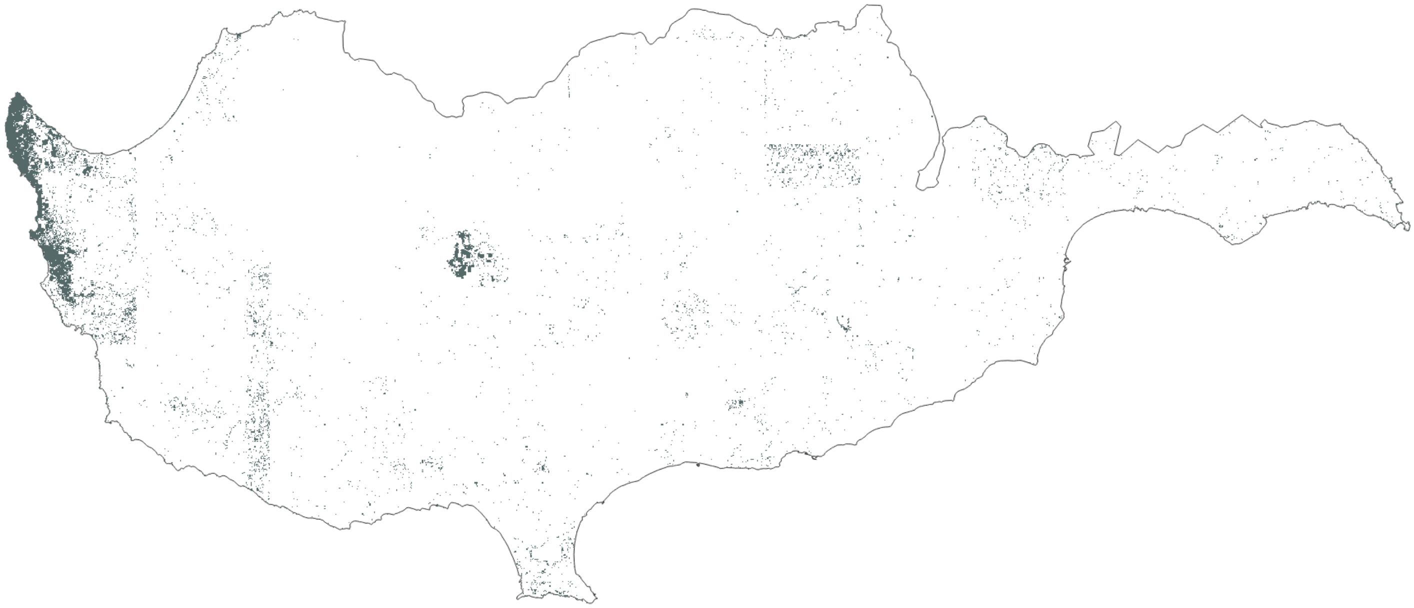
Appendix I — Misc Maps



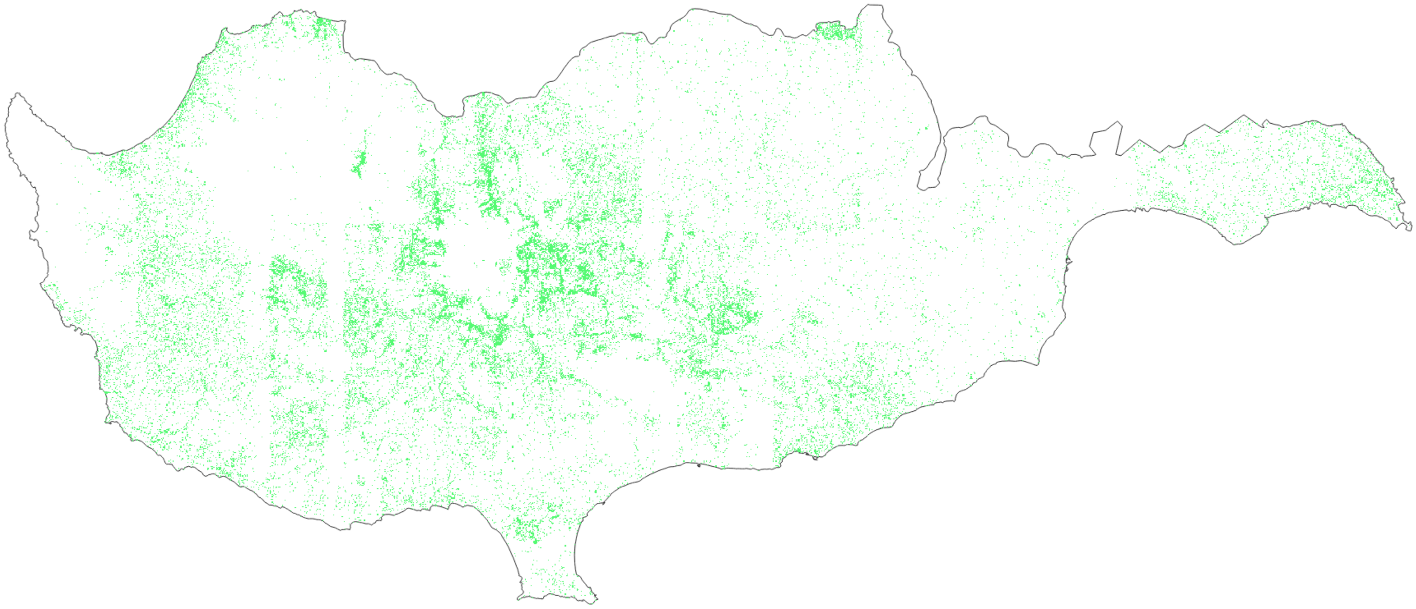
Appendix I.1: Cyprus Cedar Extent



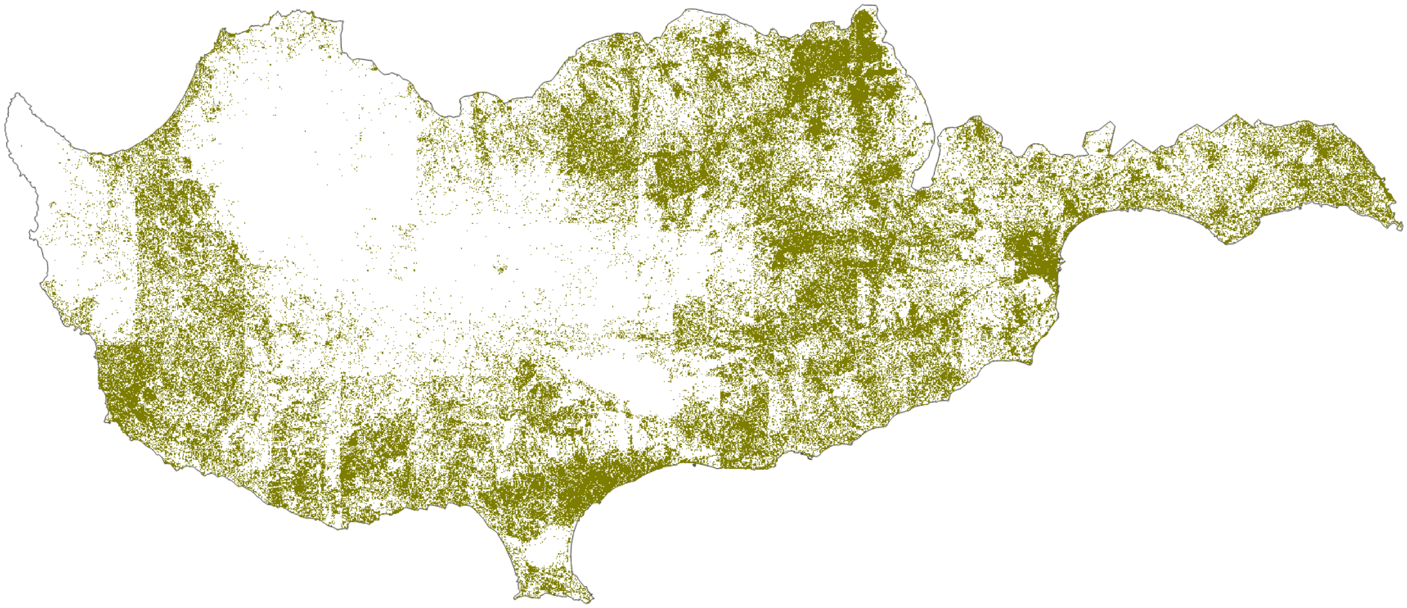
Appendix I.2: Carob Extent



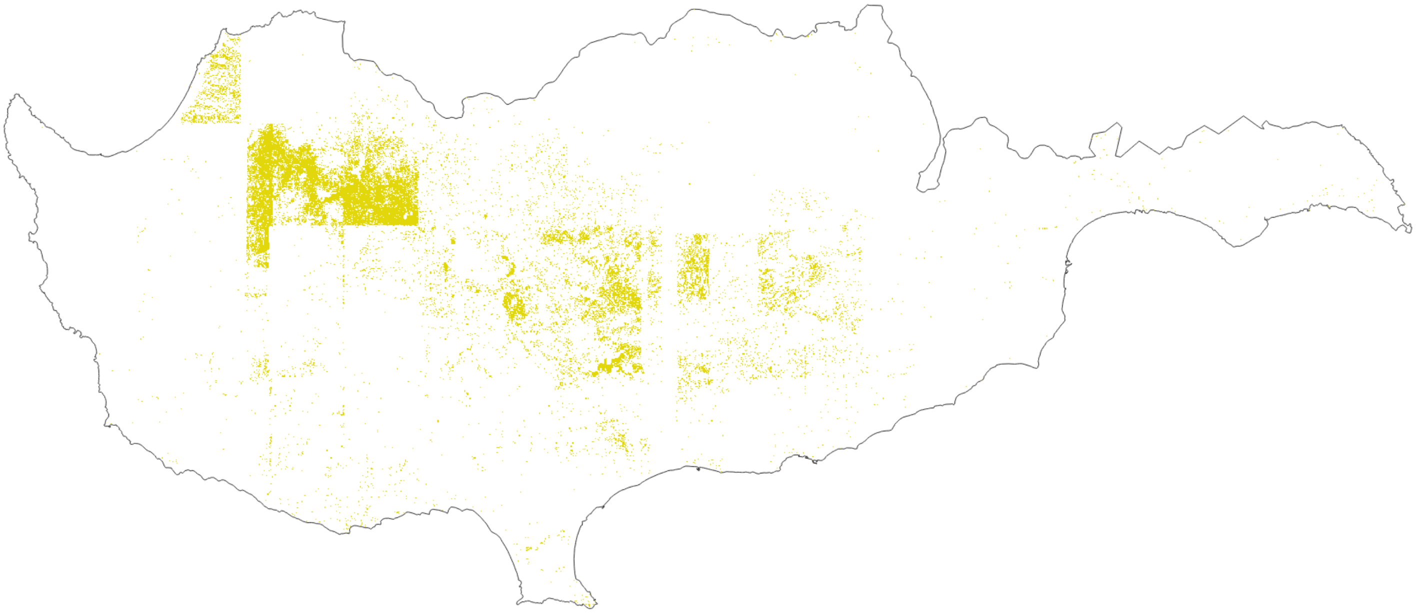
Appendix I.3: Juniper Extent



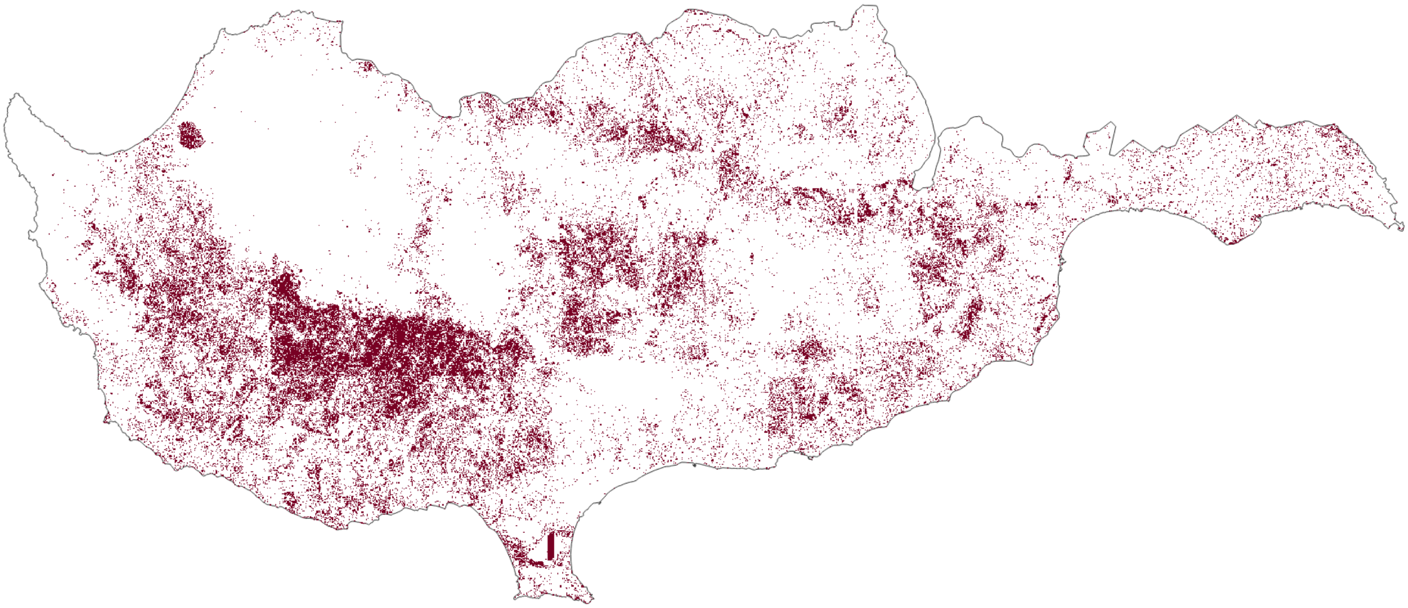
Appendix I.4: Leafy-Fruitbearing Extent



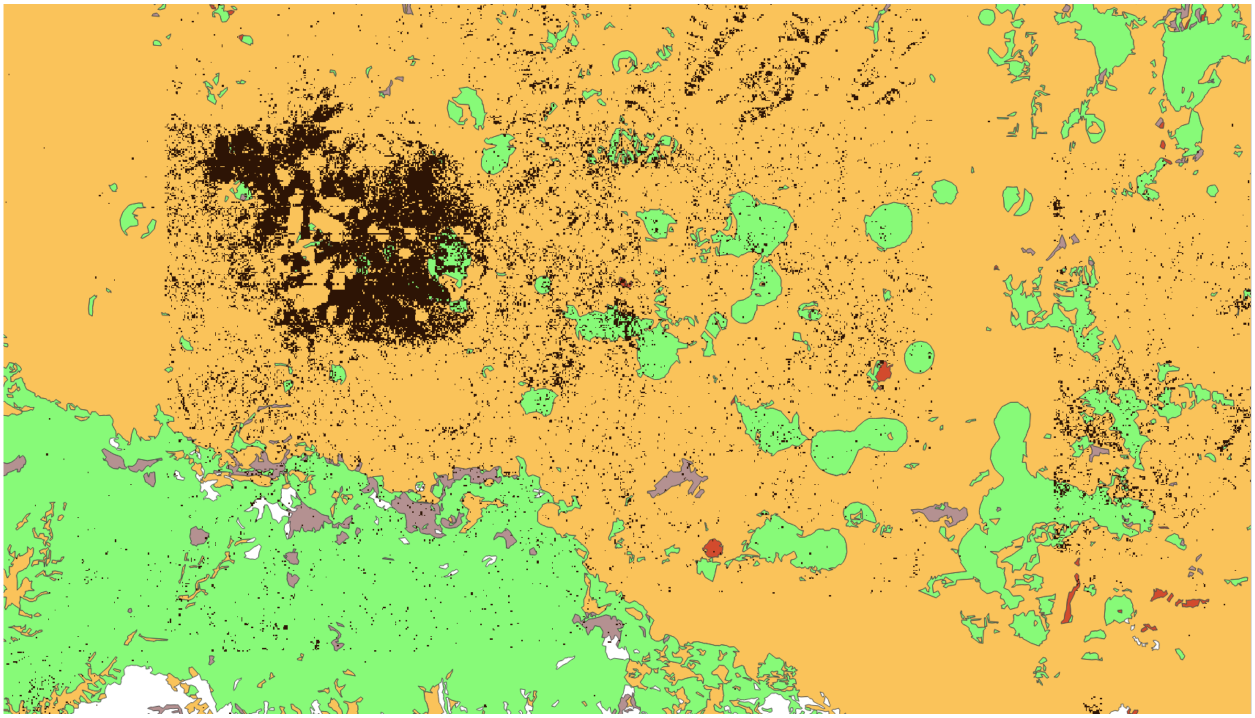
Appendix I.5: Olive Extent



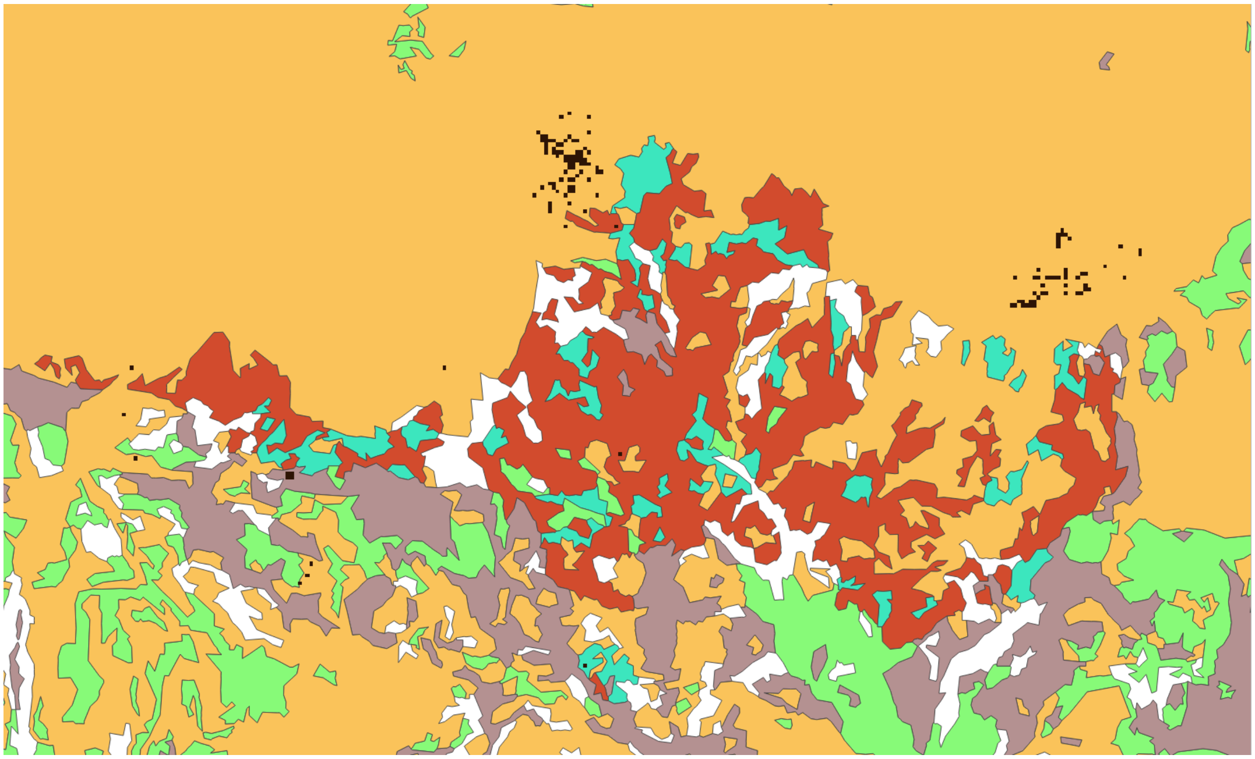
Appendix I.6: Golden Oak Extent



Appendix I.7: Vine Extent

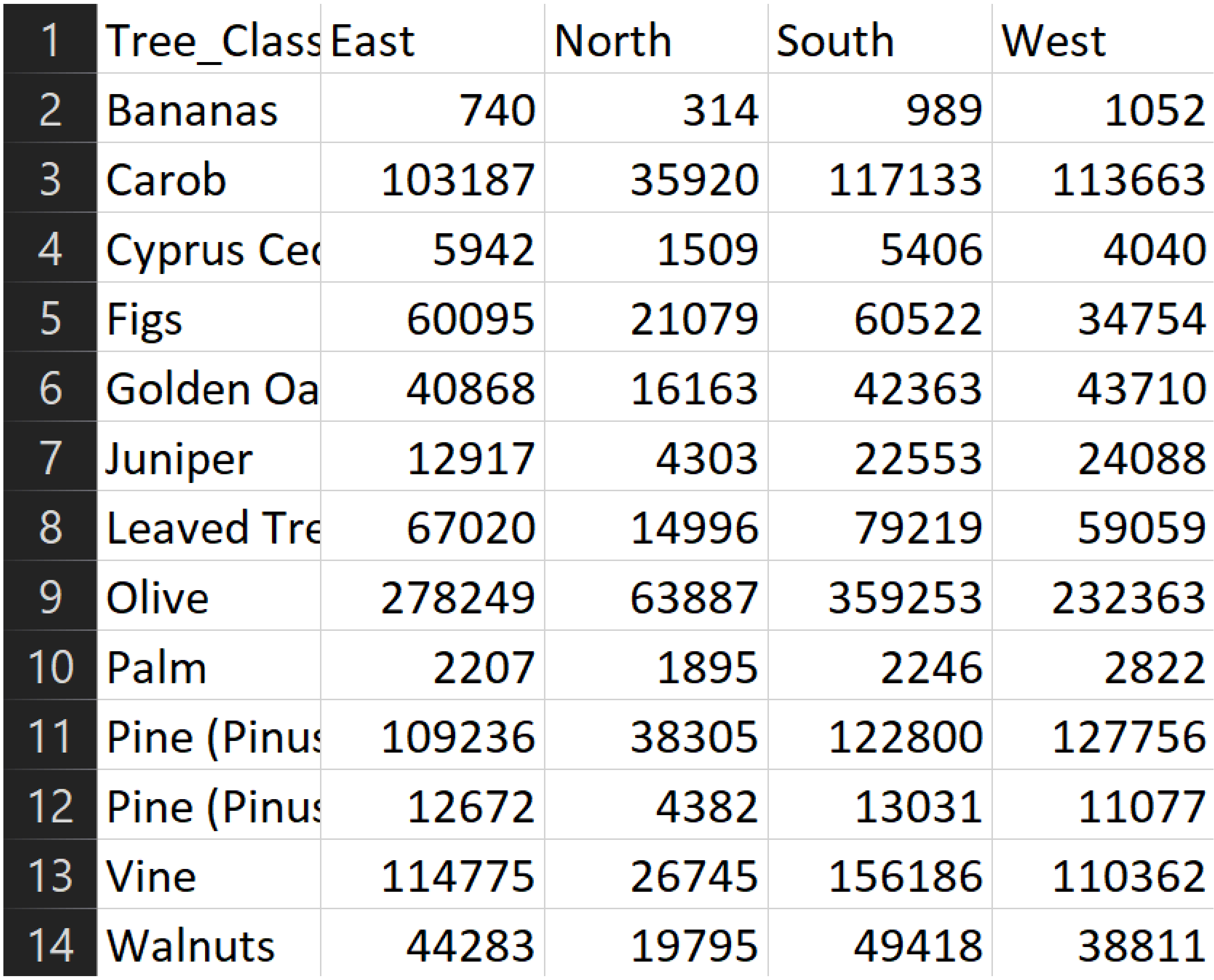


Appendix I.8: Black Pine over Clay Loam Rock Sandy Loam



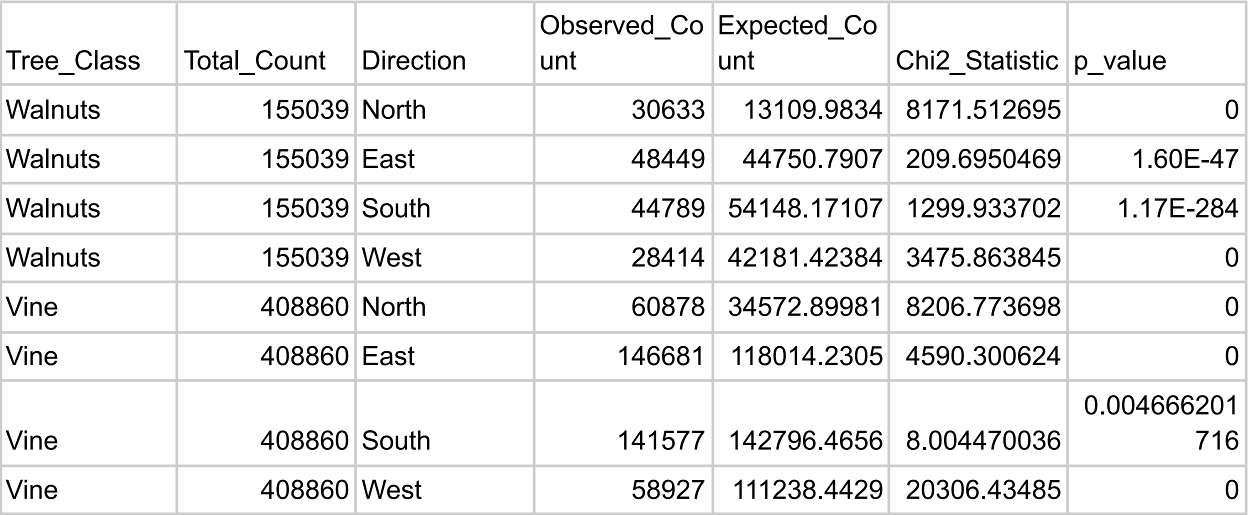
Appendix I.9: Black Pine over Clay Loam Rock Sandy Loam Gravely sand outlier

Appendix II — Raw Stats



1	Tree_Class	East	North	South	West
2	Bananas	740	314	989	1052
3	Carob	103187	35920	117133	113663
4	Cyprus Ce	5942	1509	5406	4040
5	Figs	60095	21079	60522	34754
6	Golden Oa	40868	16163	42363	43710
7	Juniper	12917	4303	22553	24088
8	Leaved Tre	67020	14996	79219	59059
9	Olive	278249	63887	359253	232363
10	Palm	2207	1895	2246	2822
11	Pine (Pinus	109236	38305	122800	127756
12	Pine (Pinus	12672	4382	13031	11077
13	Vine	114775	26745	156186	110362
14	Walnuts	44283	19795	49418	38811

Appendix.II.1: Tree Class v. Aspect Facing crosstab



Tree_Class	Total_Count	Direction	Observed_Count	Expected_Count	Chi2_Statistic	p_value
Walnuts	155039	North	30633	13109.9834	8171.512695	0
Walnuts	155039	East	48449	44750.7907	209.6950469	1.60E-47
Walnuts	155039	South	44789	54148.17107	1299.933702	1.17E-284
Walnuts	155039	West	28414	42181.42384	3475.863845	0
Vine	408860	North	60878	34572.89981	8206.773698	0
Vine	408860	East	146681	118014.2305	4590.300624	0
Vine	408860	South	141577	142796.4656	8.004470036	0.004666201716
Vine	408860	West	58927	111238.4429	20306.43485	0

Figs	181780	North	38376	15371.18262	11553.69323	0
Figs	181780	East	66902	52469.3705	2597.653491	0
Figs	181780	South	46051	63487.60336	3972.030341	0
Figs	181780	West	25096	49456.84135	10012.72309	0
Golden Oak	143199	North	26645	12108.80174	6304.772637	0
Golden Oak	143199	East	43307	41333.26761	65.26782112	6.54E-16
Golden Oak	143199	South	43398	50012.98995	694.9778925	3.70E-153
Golden Oak	143199	West	29754	38960.11786	1622.394399	0
Leaved Tree - Fruit Bearing	220660	North	37630	18658.84672	7327.858398	0
Leaved Tree - Fruit Bearing	220660	East	77682	63691.7774	2036.706795	0
Leaved Tree - Fruit Bearing	220660	South	73309	77066.64406	142.352883	8.14E-33
Leaved Tree - Fruit Bearing	220660	West	31671	60034.9137	11073.09515	0
Palm	9313	North	1824	787.5004058	477.5468605	7.30E-106
Palm	9313	East	2119	2688.124367	90.49995558	1.85E-21
Palm	9313	South	2518	3252.613324	135.128483	3.09E-31
Palm	9313	West	2709	2533.785694	8.056839078	0.004533220341
Olive	937221	North	152652	79250.71597	26512.20734	0
Olive	937221	East	346858	270521.4869	14073.95616	0
Olive	937221	South	312085	327329.2722	551.5303093	5.85E-122
Olive	937221	West	122120	254989.4945	58604.41235	0
Juniper	64363	North	8095	5442.487772	580.3717928	3.11E-128
Juniper	64363	East	16863	18577.87487	114.368826	1.08E-26
Juniper	64363	South	27723	22479.10999	897.6023691	3.26E-197
Juniper	64363	West	11180	17511.2261	1797.233957	0
Pine (Pinus brutia)	398680	North	68105	33712.08652	13317.49457	0
Pine (Pinus brutia)	398680	East	118567	115075.8534	73.74422856	8.89E-18
Pine (Pinus brutia)	398680	South	130796	139241.048	399.2600033	7.98E-89
Pine (Pinus brutia)	398680	West	80629	108468.7727	5372.51152	0

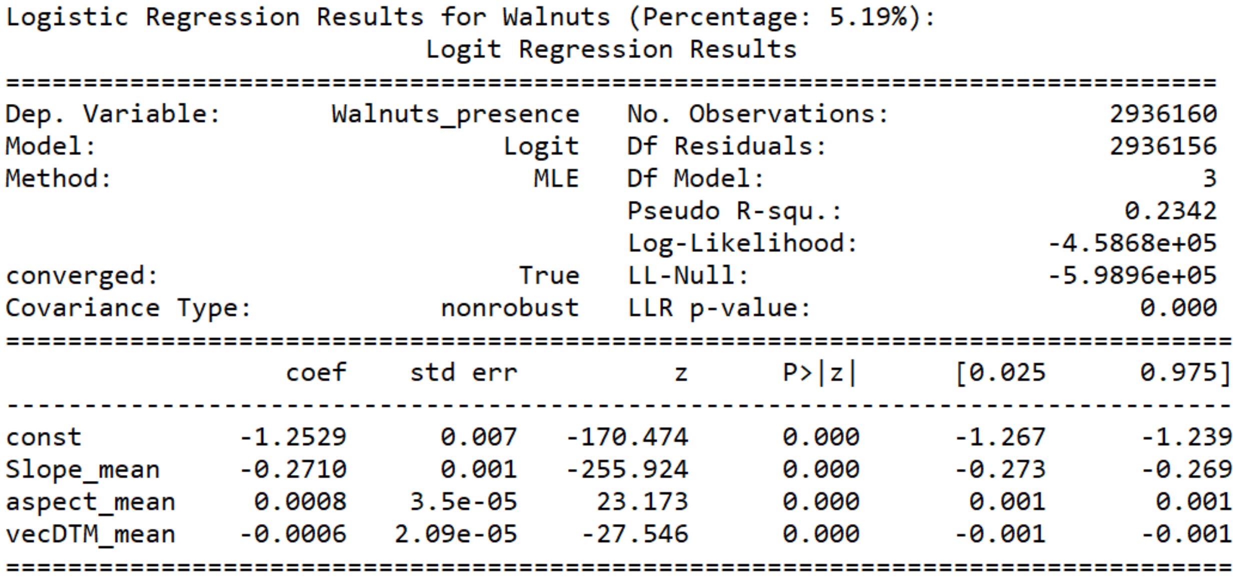
Carob	371932	North	65206	31450.29539	13548.41031	0
Carob	371932	East	111656	107355.2531	119.6397287	7.59E-28
Carob	371932	South	121245	129899.1709	450.1123853	6.82E-100
Carob	371932	West	71794	101191.4508	6509.236259	0
Bananas	3109	North	507	262.8947451	87.61179249	7.96E-21
Bananas	3109	East	847	897.3884525	1.943569524	0.1632814844
Bananas	3109	South	1003	1085.834299	4.827865331	0.02800328241
Bananas	3109	West	738	845.8648903	9.674621496	0.00186830756
Pine (Pinus nigra)	41264	North	7958	3489.253382	2024.534179	0
Pine (Pinus nigra)	41264	East	13567	11910.52979	155.606422	1.03E-35
Pine (Pinus nigra)	41264	South	12081	14411.66501	301.7176055	1.39E-67
Pine (Pinus nigra)	41264	West	7556	11226.68666	928.2251121	7.18E-204
Cyprus Cedar	16900	North	3784	1429.051526	1256.755483	2.82E-275
Cyprus Cedar	16900	East	6016	4878.052379	175.0894153	5.72E-40
Cyprus Cedar	16900	South	4808	5902.412239	163.4038164	2.04E-37
Cyprus Cedar	16900	West	2289	4597.978979	971.3800735	2.99E-213

Appendix.II.2: Tree Class v. Aspect Facing Chi2 test raw results

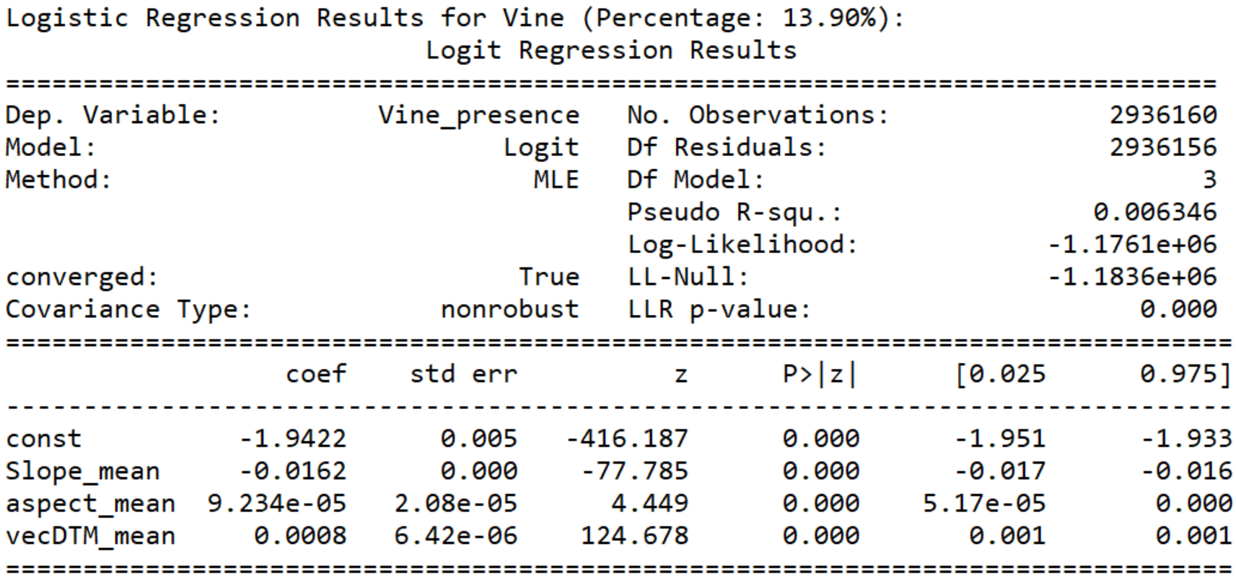
Tree_Class	Clay	Clay loam	Gravel	Gravelly sand	Loam	Loamy sand	Rock	Sand	Sandy loam
Bananas	1164	35	0	7	1352	3	409	88	37
Carob	26678	10330	25	619	105290	203	216703	6549	3506
Cyprus Cedar	172	6	0	1	592	0	16082	3	41
Figs	59020	1589	0	52	77852	322	31906	4333	1376
Golden Oak	1702	512	0	23	15233	14	125032	90	498
Juniper	6253	303	14	70	17854	44	35604	3383	336
Leaved Tree - Fruit	28322	5980	5	268	74644	226	104852	3762	2235

Bearing									
Olive	191243	27562	12	1630	350442	1772	321570	25027	14494
Palm	5474	47	0	0	2147	13	1123	204	162
Pine (Pinus brutia)	6876	1577	0	367	50304	39	334287	2731	1916
Pine (Pinus nigra)	1347	113	0	1	6963	10	32331	156	241
Vine	64526	13822	4	173	182488	234	135925	7225	3671
Walnuts	70004	1102	0	92	54092	154	21892	3490	1481

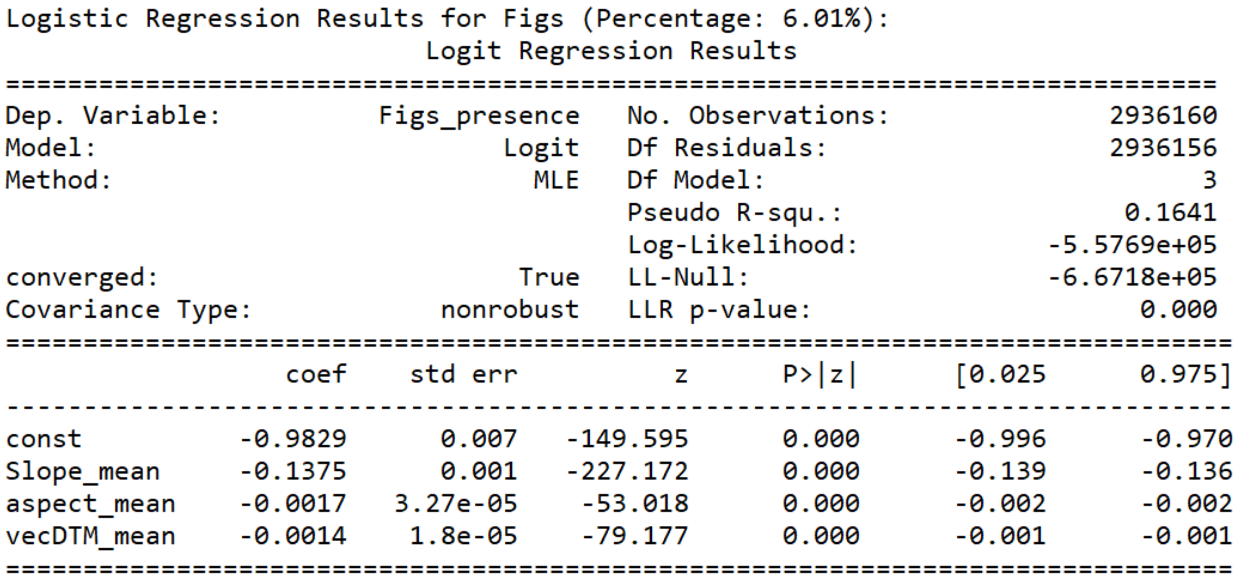
Appendix.II.3: Tree Class v. Soil Type table



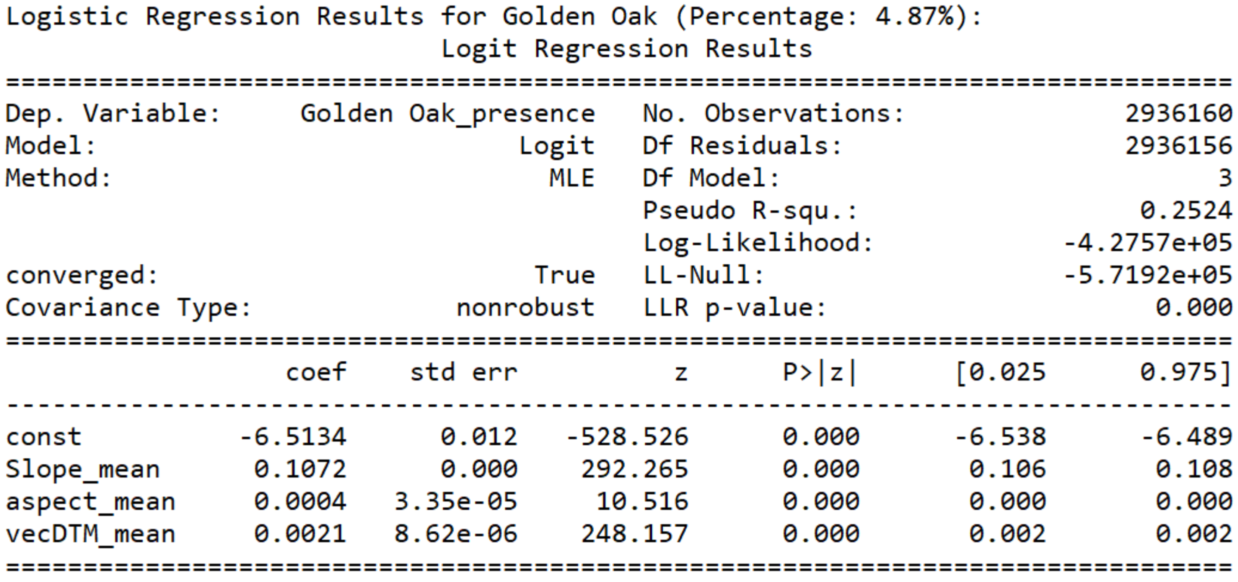
```
Logistic Regression Results for Walnuts (Percentage: 5.19%):
                           Logit Regression Results
==============================================================================
Dep. Variable:       Walnuts_presence   No. Observations:              2936160
Model:                          Logit   Df Residuals:                  2936156
Method:                           MLE   Df Model:                            3
                                        Pseudo R-squ.:                  0.2342
                                        Log-Likelihood:            -4.5868e+05
converged:                       True   LL-Null:                   -5.9896e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
===============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
const         -1.2529      0.007   -170.474      0.000      -1.267      -1.239
Slope_mean    -0.2710      0.001   -255.924      0.000      -0.273      -0.269
aspect_mean    0.0008    3.5e-05     23.173      0.000       0.001       0.001
vecDTM_mean   -0.0006   2.09e-05    -27.546      0.000      -0.001      -0.001
===============================================================================
```



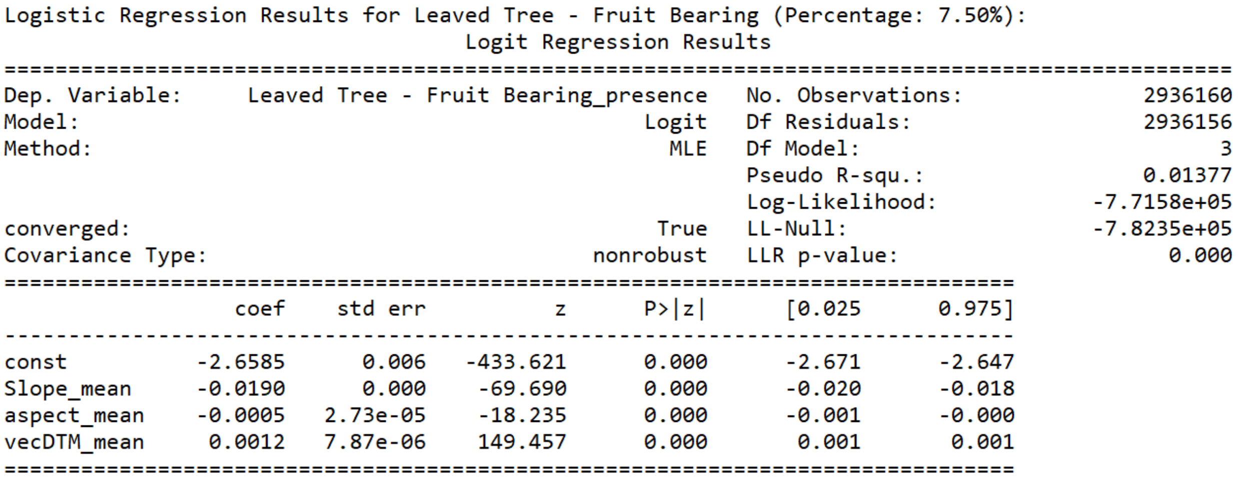
```
Logistic Regression Results for Vine (Percentage: 13.90%):
                           Logit Regression Results                           
==============================================================================
Dep. Variable:          Vine_presence   No. Observations:              2936160
Model:                          Logit   Df Residuals:                  2936156
Method:                           MLE   Df Model:                            3
                                        Pseudo R-squ.:                0.006346
                                        Log-Likelihood:            -1.1761e+06
converged:                       True   LL-Null:                   -1.1836e+06
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const         -1.9422      0.005   -416.187      0.000      -1.951      -1.933
Slope_mean    -0.0162      0.000    -77.785      0.000      -0.017      -0.016
aspect_mean 9.234e-05   2.08e-05      4.449      0.000    5.17e-05       0.000
vecDTM_mean    0.0008   6.42e-06    124.678      0.000       0.001       0.001
================================================================================
```



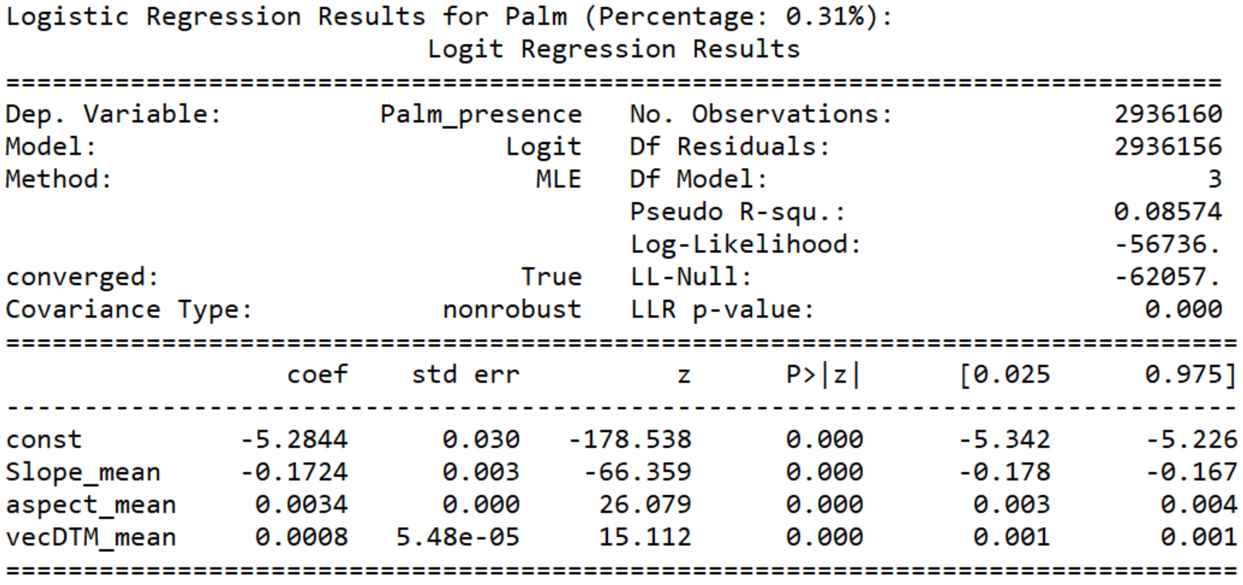
```
Logistic Regression Results for Figs (Percentage: 6.01%):
                           Logit Regression Results                           
==============================================================================
Dep. Variable:          Figs_presence   No. Observations:              2936160
Model:                          Logit   Df Residuals:                  2936156
Method:                           MLE   Df Model:                            3
                                        Pseudo R-squ.:                  0.1641
                                        Log-Likelihood:            -5.5769e+05
converged:                       True   LL-Null:                   -6.6718e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const         -0.9829      0.007   -149.595      0.000      -0.996      -0.970
Slope_mean    -0.1375      0.001   -227.172      0.000      -0.139      -0.136
aspect_mean   -0.0017   3.27e-05    -53.018      0.000      -0.002      -0.002
vecDTM_mean   -0.0014    1.8e-05    -79.177      0.000      -0.001      -0.001
================================================================================
```



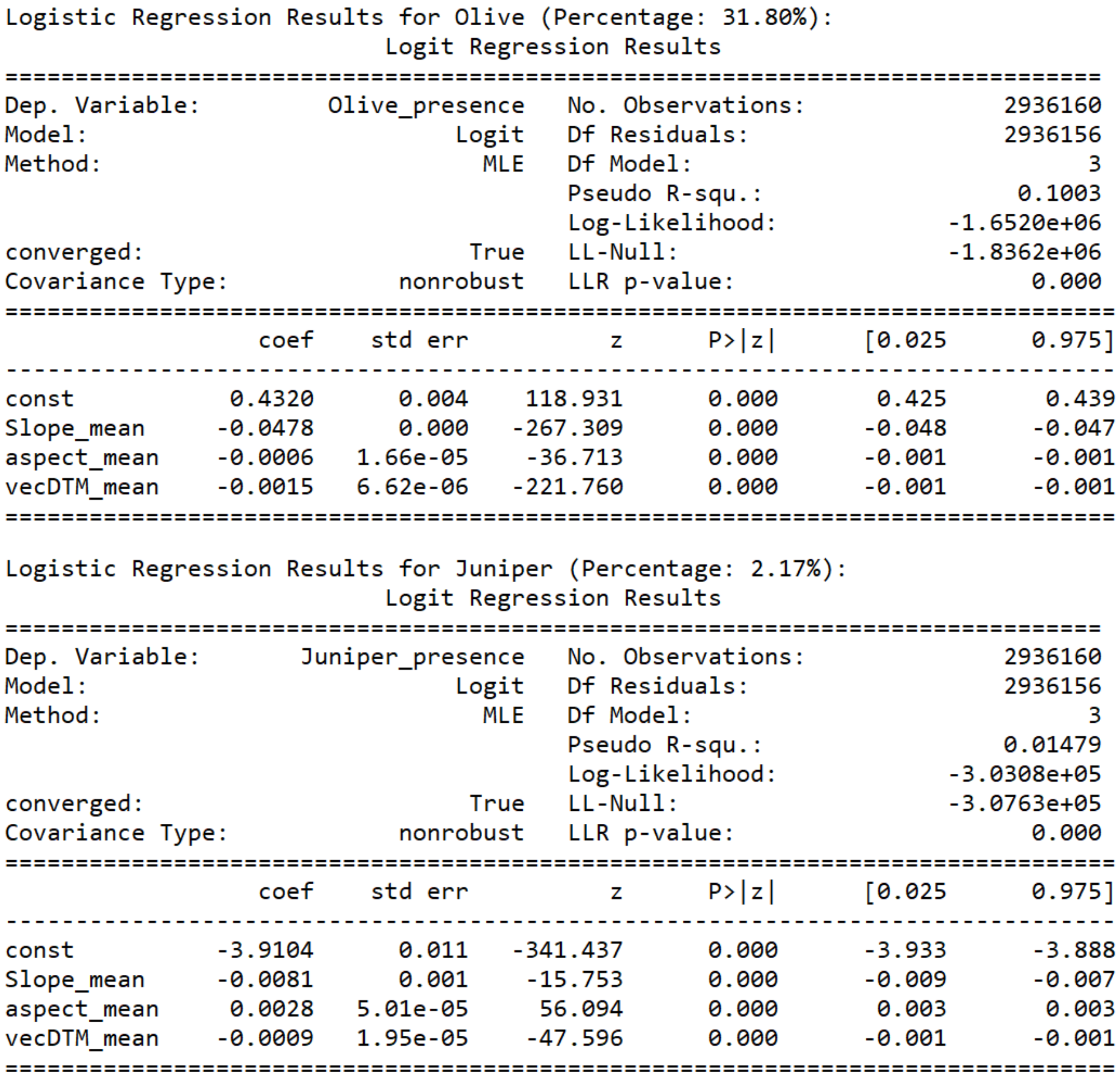
```
Logistic Regression Results for Golden Oak (Percentage: 4.87%):
                           Logit Regression Results
==============================================================================
Dep. Variable:     Golden Oak_presence   No. Observations:              2936160
Model:                           Logit   Df Residuals:                  2936156
Method:                            MLE   Df Model:                            3
                                         Pseudo R-squ.:                  0.2524
                                         Log-Likelihood:            -4.2757e+05
converged:                        True   LL-Null:                   -5.7192e+05
Covariance Type:             nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -6.5134      0.012   -528.526      0.000      -6.538      -6.489
Slope_mean     0.1072      0.000    292.265      0.000       0.106       0.108
aspect_mean    0.0004   3.35e-05     10.516      0.000       0.000       0.000
vecDTM_mean    0.0021   8.62e-06    248.157      0.000       0.002       0.002
==============================================================================
```



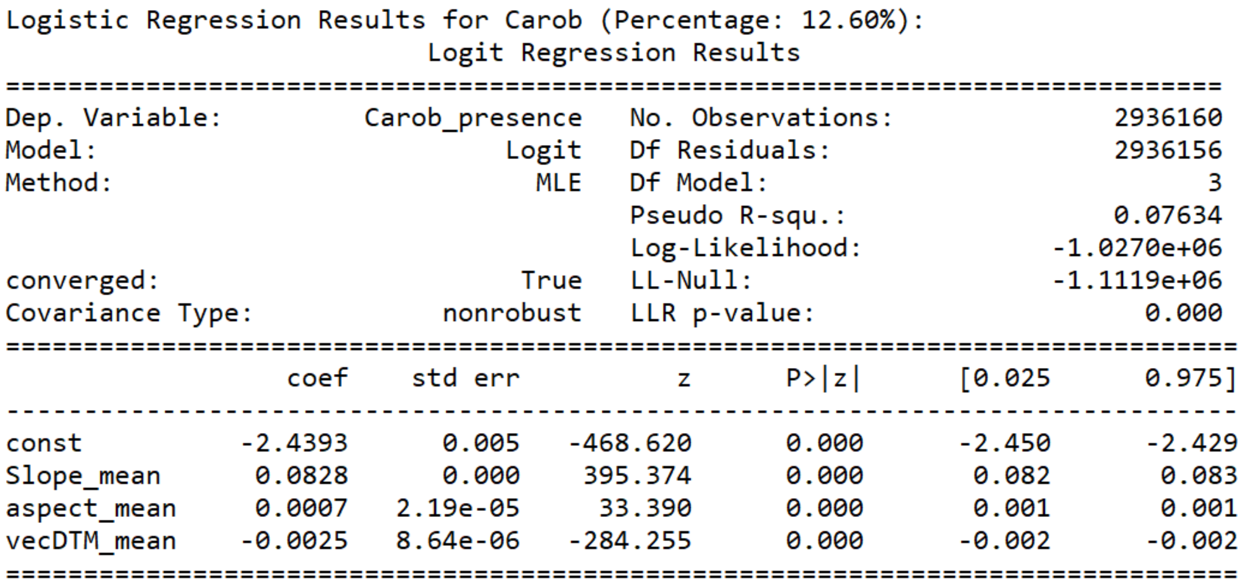
```
Logistic Regression Results for Leaved Tree - Fruit Bearing (Percentage: 7.50%):
                                Logit Regression Results
============================================================================================
Dep. Variable:     Leaved Tree - Fruit Bearing_presence   No. Observations:          2936160
Model:                                            Logit   Df Residuals:              2936156
Method:                                             MLE   Df Model:                        3
                                                          Pseudo R-squ.:             0.01377
                                                          Log-Likelihood:        -7.7158e+05
converged:                                         True   LL-Null:               -7.8235e+05
Covariance Type:                              nonrobust   LLR p-value:                 0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.6585      0.006   -433.621      0.000      -2.671      -2.647
Slope_mean    -0.0190      0.000    -69.690      0.000      -0.020      -0.018
aspect_mean   -0.0005   2.73e-05    -18.235      0.000      -0.001      -0.000
vecDTM_mean    0.0012   7.87e-06    149.457      0.000       0.001       0.001
==============================================================================
```



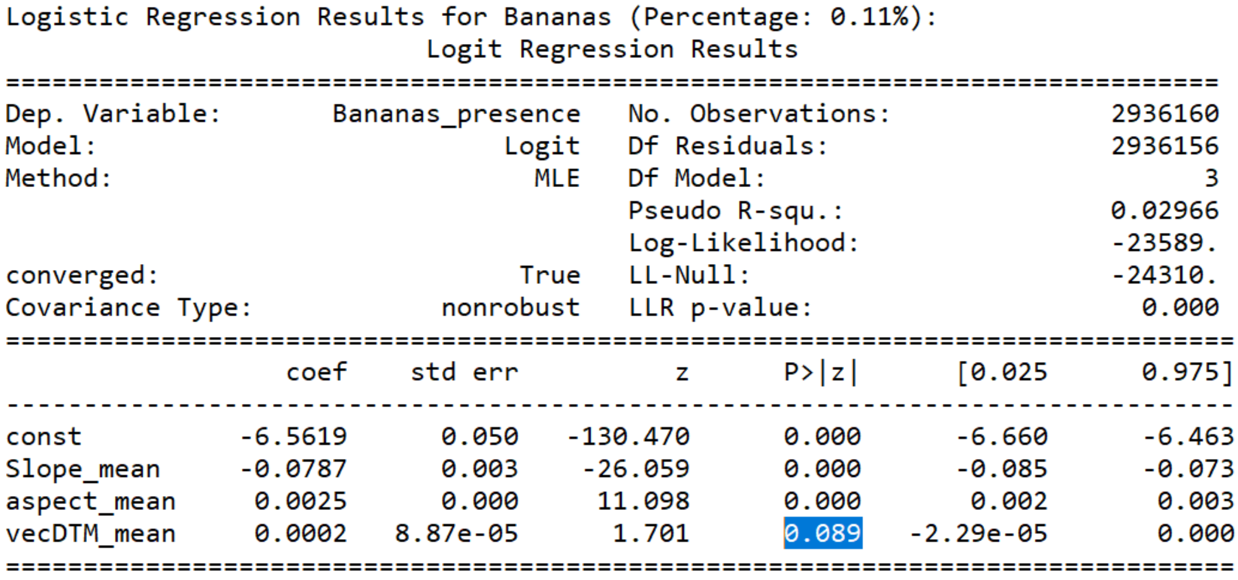
```
Logistic Regression Results for Palm (Percentage: 0.31%):
                           Logit Regression Results
==============================================================================
Dep. Variable:           Palm_presence   No. Observations:              2936160
Model:                           Logit   Df Residuals:                  2936156
Method:                            MLE   Df Model:                            3
                                         Pseudo R-squ.:                 0.08574
                                         Log-Likelihood:                -56736.
converged:                        True   LL-Null:                       -62057.
Covariance Type:             nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -5.2844      0.030   -178.538      0.000      -5.342      -5.226
Slope_mean    -0.1724      0.003    -66.359      0.000      -0.178      -0.167
aspect_mean    0.0034      0.000     26.079      0.000       0.003       0.004
vecDTM_mean    0.0008   5.48e-05     15.112      0.000       0.001       0.001
==============================================================================
```



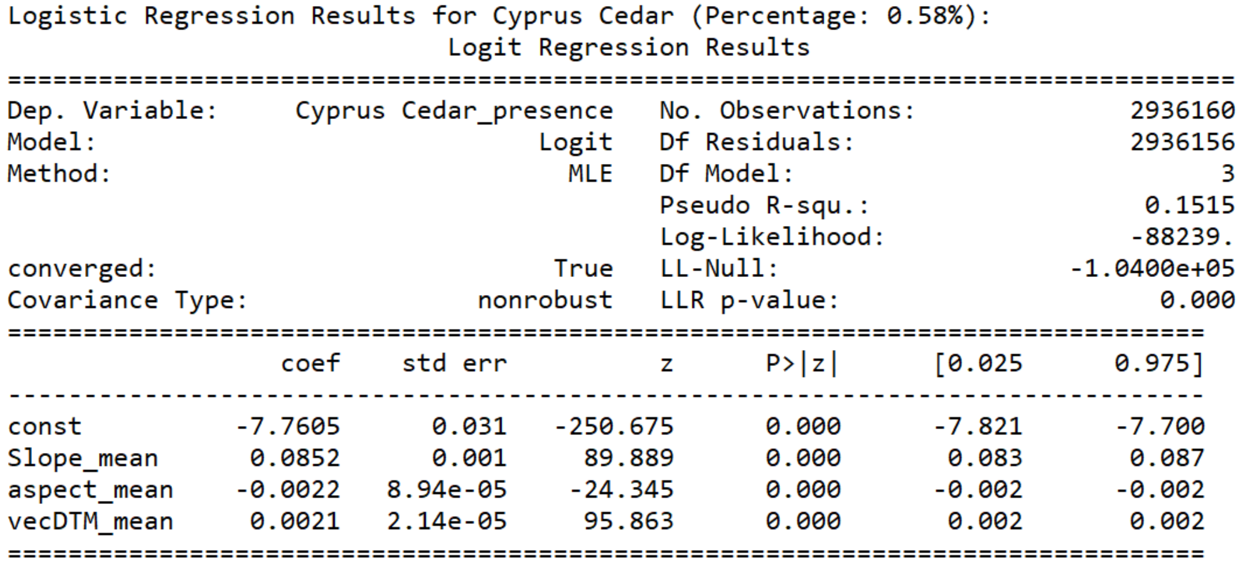
```
Logistic Regression Results for Olive (Percentage: 31.80%):
                           Logit Regression Results
==============================================================================
Dep. Variable:          Olive_presence   No. Observations:              2936160
Model:                           Logit   Df Residuals:                  2936156
Method:                            MLE   Df Model:                            3
                                         Pseudo R-squ.:                  0.1003
                                         Log-Likelihood:            -1.6520e+06
converged:                        True   LL-Null:                   -1.8362e+06
Covariance Type:             nonrobust   LLR p-value:                     0.000
================================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const           0.4320      0.004    118.931      0.000       0.425       0.439
Slope_mean     -0.0478      0.000   -267.309      0.000      -0.048      -0.047
aspect_mean    -0.0006   1.66e-05    -36.713      0.000      -0.001      -0.001
vecDTM_mean    -0.0015   6.62e-06   -221.760      0.000      -0.001      -0.001
================================================================================

Logistic Regression Results for Juniper (Percentage: 2.17%):
                           Logit Regression Results
==============================================================================
Dep. Variable:        Juniper_presence   No. Observations:              2936160
Model:                           Logit   Df Residuals:                  2936156
Method:                            MLE   Df Model:                            3
                                         Pseudo R-squ.:                 0.01479
                                         Log-Likelihood:            -3.0308e+05
converged:                        True   LL-Null:                   -3.0763e+05
Covariance Type:             nonrobust   LLR p-value:                     0.000
================================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const          -3.9104      0.011   -341.437      0.000      -3.933      -3.888
Slope_mean     -0.0081      0.001    -15.753      0.000      -0.009      -0.007
aspect_mean     0.0028   5.01e-05     56.094      0.000       0.003       0.003
vecDTM_mean    -0.0009   1.95e-05    -47.596      0.000      -0.001      -0.001
================================================================================
```



```
Logistic Regression Results for Carob (Percentage: 12.60%):
                           Logit Regression Results                           
==============================================================================
Dep. Variable:         Carob_presence   No. Observations:              2936160
Model:                          Logit   Df Residuals:                  2936156
Method:                           MLE   Df Model:                            3
                                        Pseudo R-squ.:                 0.07634
                                        Log-Likelihood:            -1.0270e+06
converged:                       True   LL-Null:                   -1.1119e+06
Covariance Type:            nonrobust   LLR p-value:                     0.000
===============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
const          -2.4393      0.005   -468.620      0.000      -2.450      -2.429
Slope_mean      0.0828      0.000    395.374      0.000       0.082       0.083
aspect_mean     0.0007   2.19e-05     33.390      0.000       0.001       0.001
vecDTM_mean    -0.0025   8.64e-06   -284.255      0.000      -0.002      -0.002
===============================================================================
```

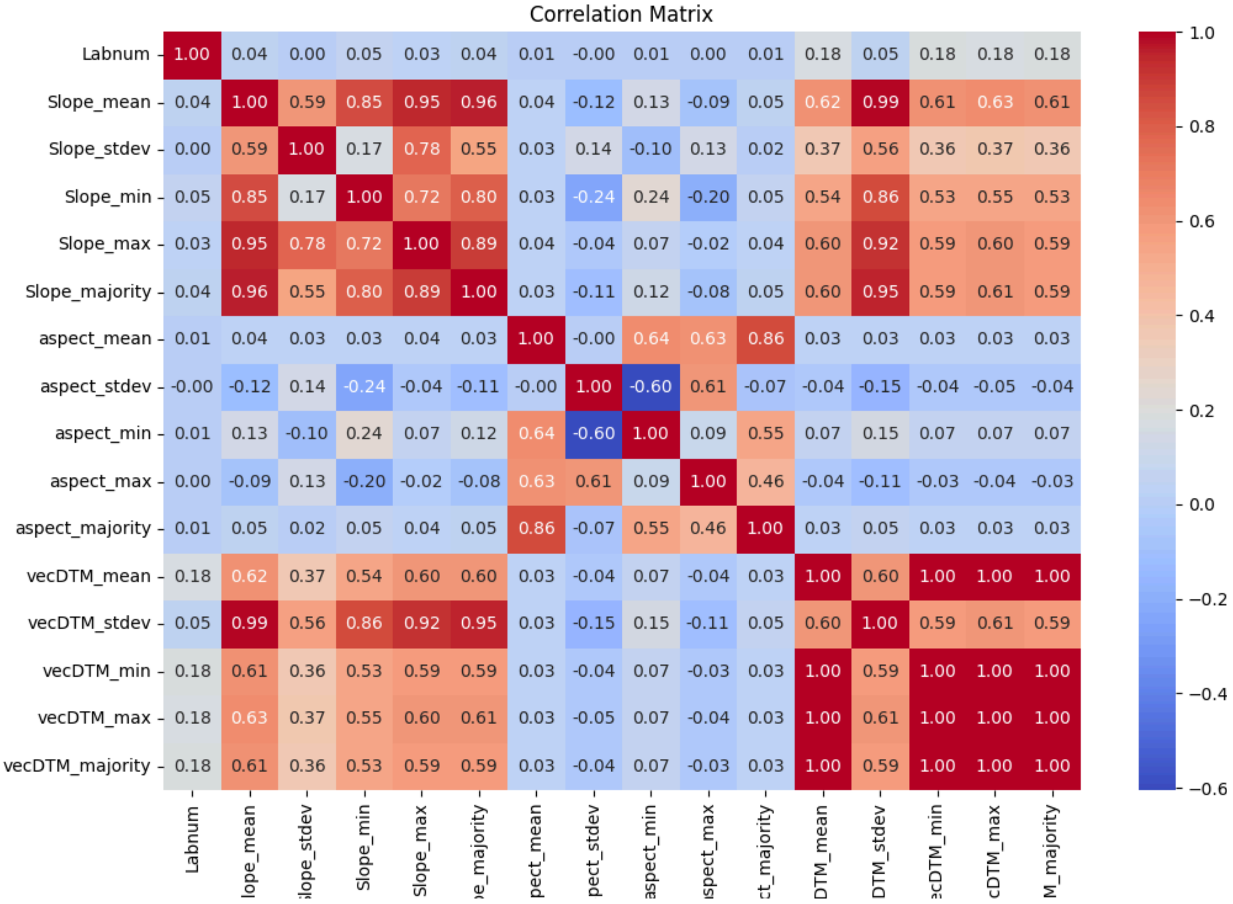


```
Logistic Regression Results for Bananas (Percentage: 0.11%):
                           Logit Regression Results
==============================================================================
Dep. Variable:        Bananas_presence   No. Observations:              2936160
Model:                           Logit   Df Residuals:                  2936156
Method:                            MLE   Df Model:                            3
                                         Pseudo R-squ.:                 0.02966
                                         Log-Likelihood:                -23589.
converged:                        True   LL-Null:                       -24310.
Covariance Type:             nonrobust   LLR p-value:                     0.000
===============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
const          -6.5619      0.050   -130.470      0.000      -6.660      -6.463
Slope_mean     -0.0787      0.003    -26.059      0.000      -0.085      -0.073
aspect_mean     0.0025      0.000     11.098      0.000       0.002       0.003
vecDTM_mean     0.0002   8.87e-05      1.701      0.089   -2.29e-05       0.000
===============================================================================
```



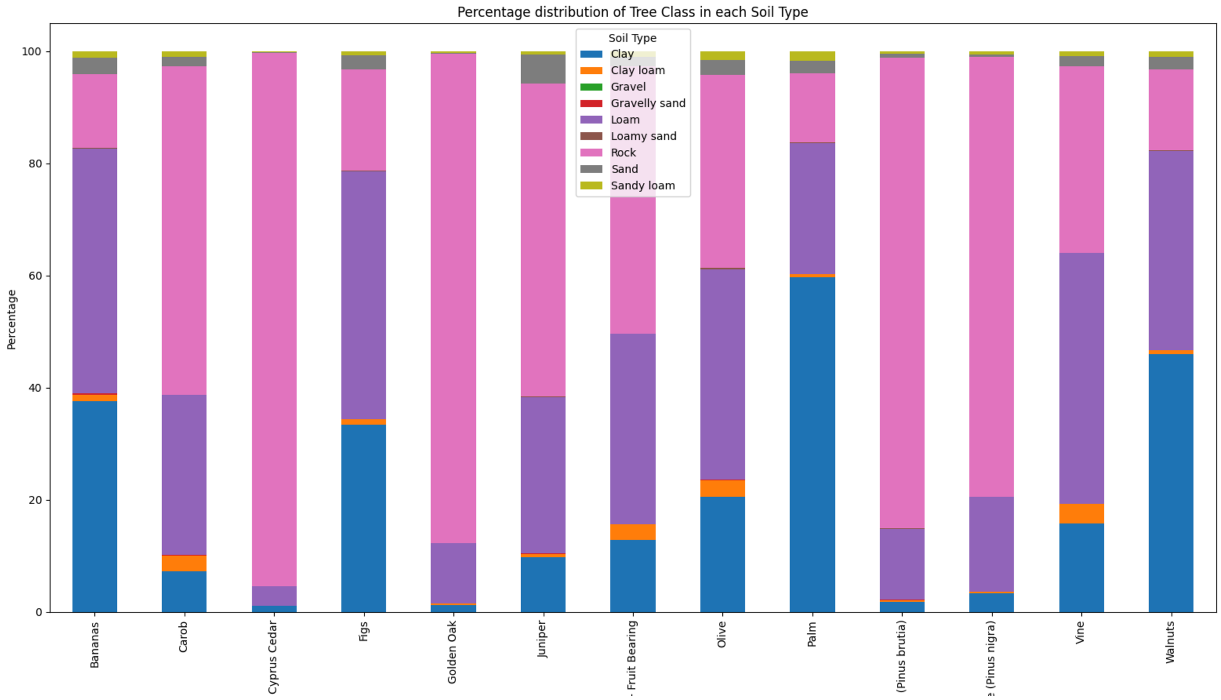
```
Logistic Regression Results for Cyprus Cedar (Percentage: 0.58%):
                            Logit Regression Results
================================================================================
Dep. Variable:     Cyprus Cedar_presence   No. Observations:             2936160
Model:                             Logit   Df Residuals:                 2936156
Method:                              MLE   Df Model:                           3
                                           Pseudo R-squ.:               0.1515
                                           Log-Likelihood:              -88239.
converged:                          True   LL-Null:                 -1.0400e+05
Covariance Type:               nonrobust   LLR p-value:                    0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -7.7605      0.031   -250.675      0.000      -7.821      -7.700
Slope_mean     0.0852      0.001     89.889      0.000       0.083       0.087
aspect_mean   -0.0022   8.94e-05    -24.345      0.000      -0.002      -0.002
vecDTM_mean    0.0021   2.14e-05     95.863      0.000       0.002       0.002
==============================================================================
```

Appendix.II.4: Logit Results



	Labnum	lope_mean	ılope_stdev	Slope_min	Slope_max	ɔe_majority	pect_mean	pect_stdev	aspect_min	ıspect_max	ct_majority	DTM_mean	DTM_stdev	ɛcDTM_min	cDTM_max	M_majority
Labnum	1.00	0.04	0.00	0.05	0.03	0.04	0.01	-0.00	0.01	0.00	0.01	0.18	0.05	0.18	0.18	0.18
Slope_mean	0.04	1.00	0.59	0.85	0.95	0.96	0.04	-0.12	0.13	-0.09	0.05	0.62	0.99	0.61	0.63	0.61
Slope_stdev	0.00	0.59	1.00	0.17	0.78	0.55	0.03	0.14	-0.10	0.13	0.02	0.37	0.56	0.36	0.37	0.36
Slope_min	0.05	0.85	0.17	1.00	0.72	0.80	0.03	-0.24	0.24	-0.20	0.05	0.54	0.86	0.53	0.55	0.53
Slope_max	0.03	0.95	0.78	0.72	1.00	0.89	0.04	-0.04	0.07	-0.02	0.04	0.60	0.92	0.59	0.60	0.59
Slope_majority	0.04	0.96	0.55	0.80	0.89	1.00	0.03	-0.11	0.12	-0.08	0.05	0.60	0.95	0.59	0.61	0.59
aspect_mean	0.01	0.04	0.03	0.03	0.04	0.03	1.00	-0.00	0.64	0.63	0.86	0.03	0.03	0.03	0.03	0.03
aspect_stdev	-0.00	-0.12	0.14	-0.24	-0.04	-0.11	-0.00	1.00	-0.60	0.61	-0.07	-0.04	-0.15	-0.04	-0.05	-0.04
aspect_min	0.01	0.13	-0.10	0.24	0.07	0.12	0.64	-0.60	1.00	0.09	0.55	0.07	0.15	0.07	0.07	0.07
aspect_max	0.00	-0.09	0.13	-0.20	-0.02	-0.08	0.63	0.61	0.09	1.00	0.46	-0.04	-0.11	-0.03	-0.04	-0.03
aspect_majority	0.01	0.05	0.02	0.05	0.04	0.05	0.86	-0.07	0.55	0.46	1.00	0.03	0.05	0.03	0.03	0.03
vecDTM_mean	0.18	0.62	0.37	0.54	0.60	0.60	0.03	-0.04	0.07	-0.04	0.03	1.00	0.60	1.00	1.00	1.00
vecDTM_stdev	0.05	0.99	0.56	0.86	0.92	0.95	0.03	-0.15	0.15	-0.11	0.05	0.60	1.00	0.59	0.61	0.59
vecDTM_min	0.18	0.61	0.36	0.53	0.59	0.59	0.03	-0.04	0.07	-0.03	0.03	1.00	0.59	1.00	1.00	1.00
vecDTM_max	0.18	0.63	0.37	0.55	0.60	0.61	0.03	-0.05	0.07	-0.04	0.03	1.00	0.61	1.00	1.00	1.00
vecDTM_majority	0.18	0.61	0.36	0.53	0.59	0.59	0.03	-0.04	0.07	-0.03	0.03	1.00	0.59	1.00	1.00	1.00

Appendix.II.5: Correlation Matrix



Appendix II.6: Percentage distribution of Tree Class v. Soil Type