**TASK 5**

Research Sample

A research sample is a group of individuals, things, or products chosen for evaluation from a wider population (Education Centre, 2006). To guarantee that the findings from the study sample can be applied to the entire population, the sample should be representative of the population.

For my research, I collected Satellite data from the Sentinel-2 mission, which was applied for its relatively high 10 m spatial resolution, and for having radiometry with three vegetation red edge bands (Abdi, 2019). These two characteristics make the Sentinel-2 data appealing for LCLU mapping. The said datasets were downloaded from the satellite imagery platform, the Copernicus Hub. These are cloudless images, spread over the span of 4 years: 2 years between 2015 to 2017, and 2 years between 2020 to 2022. Each set of years has 11 images, for a total of 22 images for both groups, which are roughly in alternate months. It was crucial for these images not to be affected by distortion, such as clouds. As a result, care has been taken in determining which images to obtain for the winter months when cloud-free images are scarce.

Another dataset gathered was from ARPA, regarding agriculture activity detected within a whole year from their on-the-spot land inspections. This dataset was split into two classes, being NMM i.e., no-minimum agricultural activity, and Arable. These classes were observed as “no Ploughing” i.e., the declaration is confirmed, and another as “bare Soil” i.e., the declaration is not correct. For every parcel, there were 1, 2 or 3 confirmed inspection dates. The total number of parcels within the study area includes 85 ‘arable’ parcels, 48 ‘non-cultivated’ parcels and 34 ‘other’ parcels. After running both algorithms, data extraction was done, and a custom dataset in Geo Tiff format was created. The files were utilized on QGIS to view and extract the outputted marked land listed under their respective classes (arable, non-cultivated or other) and to see the changes in the land from the first set of images from 2015 to 2017 to the second set from 2020 to 2022.

Ethics

* *Prior to beginning the study*

For my study, I first had to research how my paper’s objective would be beneficial and what was missing in the field of remote sensing. After deciding, I then submitted a research proposal form, also known as the Statement of Intent (SOI) for the Institute Research Sub-Committee (IRC) formally established at every institute to verify that the proposal sufficiently follows the examination bodies' standards as well as the ethical guidelines. It then got approved for me to carry out the research intended. Considering that my study is a quantitative one, which did not consist of any participants, no ethical measures were needed. This is scientific research on proof of concept with an environmental significance. Hence, no personal and confidential data will be breached and no identities to divulge. There is also no consequence of possible physical harm during the implementation of this research and no consequence of possible moral harm during the implementation of this research. This study is not tied to any commercial interest, hence there will be no consequences on businesses during the implementation of this research.

* *Beginning the study*

In Kuss et al’s study (2022), the participants were chosen after a prior study was completed. Those that were interested were encouraged to give their contact information at the end of the survey. All participants who provided contact information were contacted and invited to participate in an interview. As a thank you for their time, participants were given a GBP 20 voucher. From this, 20% expressed an interest in participating. Participants in both studies had to be able to play video games, converse in English, and identify as female. Informed consent was sought from all subjects involved in the study, along with the individual's preference for how they wanted to be interviewed and their availability. For convenience, participants could choose from a range of interview techniques, including in-person, Skype audio, or Skype written, which was offered in case they didn't have access to a microphone or sound equipment, or if they were apprehensive about speaking out loud. All interviews were conducted in private, with the consent of the participants to record the session. In-depth and personalized comments were encouraged in private interviews.

* *Collecting data*

The data collected for my study included an on-the-spot land inspection dataset from ARPA, in which land surveyors inspect which lands are being cultivated and which are not. Prior to this, I contacted them to ask me to supply this information where a meeting was set up to discuss further. The manager I spoke to later contacted his superior to get permission for them to supply this Land Parcel Identification System (LPIS) GIS data. He obtained approval on the 22nd of November 2021 and sent me a non-disclosure agreement (NDA) for me to sign before forwarding the data to me and asked for my ID card number. The NDA covers the data specifications, the intended use of data, confidentiality and publication, indemnity, and the sharing of data. The document was duly filled and signed by both parties. The data was later sent via SharePoint as shapefile layer to be transferred to QGIS.

In Yang and Cornelius’s study (2004), data collection was done by interviews, observations, and documents. Each participant in this study underwent two interviews: an unstructured interview and a structured interview. To ensure ethical considerations, the interviews took place wherever the participants were most comfortable, such as in their dorm room, office, or the school library. Following the participants' consent, the first observation was arranged to see how he/she worked for the online class. Two observations were conducted to collect more qualitative data, where each session lasted approximately 60 minutes. Participants were given the option of scheduling their observations at their leisure and in conducive environments, in the same setting where they completed their online coursework, including the participants' homes, dorm rooms, offices, and the campus library, among other places. Documents were requested from all parties which contained printouts and other data to substantiate the information acquired and witnessed during the observation and interview process.

* *Analysing data*

When it came to analysing data from my study, I went for an unbiased approach with testing two common algorithms used in remote sensing for land cover change, the random forest classification and K-means clustering. According to past papers I read for my literature review which utilized such algorithms, they all conclude that the Random Forest algorithm gave better accuracy results. I applied both algorithms to my research to confirm the results of such past research papers.

Kuss et al. (2022) mention that their thematic analysis for all interviews follows full anonymization, meaning that their data processing technique removes or modifies personally identifiable information to prevent it from being linked to a specific person. It's also a significant factor in Google's privacy commitment.

* *Reporting, Sharing and Storing data*

According to the contractual agreement between me and ARPA, I am obligated to send them a final draft of the entire dissertation before it is published, with the citations and conclusions generated from ARPA's data being vetted by the agency. It states that the agency reserves the right to request amendments to any text that it believes, in its sole discretion, misrepresents any data it has provided or conclusions drawn from it, and I agree to make every effort to comply with such request and, in any case, not to publish any reference to which the agency has not expressly consented in writing, bearing sole responsibility for any failure to do so.

References:

Education Research, 2006, *Sampling in Research*, Available at: <https://www.thh.nhs.uk/documents/_Departments/Research/InfoSheets/16_sampling_research.pdf> (Accessed: 30 March 2022)

Kuss, D.J.; Kristensen, A.M.; Williams, A.J.; Lopez-Fernandez, O., 2022, To Be or Not to Be a Female Gamer: A Qualitative Exploration of Female Gamer Identity. *Int. J. Environ. Res. Public Health 2022*, 19(3):1169, DOI: 10.3390/ijerph19031169, <https://www.researchgate.net/deref/https%3A%2F%2Fdoi.org%2F10.3390%2Fijerph19031169>

Yang, Y, Cornelius, L, F, 2004, *Students’ Perceptions towards the Quality of Online Education: A Qualitative Approach*, Available at: <https://files.eric.ed.gov/fulltext/ED485012.pdf> (Accessed: 5 April 2022)

**TASK 6**

Experimental protocol & design

In scientific papers, Experimental protocols are frequently included in the "Materials and Methods" section and are critical for duplicability as they should contain all of the information needed to get consistent findings (Giraldo et al., 2018). Such protocols are essential for planning, conducting, and publishing research in a variety of fields, particularly when it comes to reporting materials and methods. Hence, the reader should understand the aim of the experiment, the materials required, the procedures to be employed, the controls, and the methods of interpretation at the end of the protocol (Jones, 2016). Jones mentioned that an experimental protocol consists of six components, starting off with the **purpose** which is a brief description of the question you're attempting to answer and the hypothesis you'll test. Moving to **materials** which include a list of all the necessary items/materials needed for the experiment. Followed by the **methods** to be carried out to complete the experiment which should be explained or cited so that a reader has all the information they need to duplicate an experiment and validate its findings. The **control** treatment should then be indicated along with the variables that will be modified and the control should be conducted in conditions that are unaffected by the variable being examined. A **data interpretation** is done once data is collected, which should be organized and summarized for other researchers to decide if the hypothesis has been proven correct or not. Finally, any cited published works should be included in the **references** for anyone to refer to that work. The whole protocol should not exceed two pages.

My study hypothesizes that uncultivated land having agricultural potential would be exposed and potentially repurposed for better use by comparing Sentinel-2 multi-Spectral imagery to ground truth data from Google Earth Pro. The research questions to be answered are as follows:

1. What satellite data and processing are required to prepare input for land-use classification?

2. What factors are considered to assume if potentially cultivated land is being cultivated or not?

3. What is the level of accuracy of the developed classification algorithm in relation to a ground truth?

Independent variables & Dependent variables

The elements that the researchers are interested in researching or the likely "cause" of the change in the dependent variable are referred to as **independent** variables. Hence, the researcher manipulates or alters the independent variable, which is expected to have a direct influence on the dependent variable (McLeod, 2019). The **dependent** variable is the variable being tested and evaluated and is influenced by a participant's actions or changes in the independent variables. The aim of the study is to see if and how changes in independent factors affect dependent variables.

Two algorithms are applied, and their accuracy is compared. Clouds present in the sentinel imagery affect the accuracy of the algorithm. The **independent** variable here is the clouds and the **dependent** variable is accuracy. This is because clouds covering land in the study area would intervene in the interpretation of land being cultivated or not, hence the algorithm would classify it under the wrong category, affecting the other datasets and the classification overall.

External validity of an experiment & Sampling types

The extent to which the findings are applicable in the actual world is referred to as external validity and relates to how effectively a study's findings may be anticipated to apply in other situations. In other words, the generalizability of the findings is determined by this form of validity. Here, the outcomes are applicable to real-life circumstances and to the entire world and can be applied to different situations. The three main types of external validity are population validity, which relates to whether you can properly extrapolate your sample's findings to a bigger group of people (the population), using non-probability sampling methods which include **convenience sample**, **purposive sample**, **snowball sample** and **quota sample** (McCombes, 2022); ecological/environmental validity, which show whether the results can be applied to contexts or places other than those in which you conducted your experiment; and temporal validity, to check whether it is possible for a study’s findings to be used at any time, not simply during the period or season in which your experiment was conducted (Williams, 2022), (Bhandari, 2021). A sample that represents the population in terms of external validity is known as a **representative sample**, where every individual in the population has an equal probability of getting chosen (Noob School, n.d.). If obtaining a Representative Sample Isn't a Priority, External validity is connected to **random sampling**, and this is frequently the key concern in frequency assertions. External validity is improved by random sampling.

By comparing to Google Earth Pro, I validate that my findings mirror the imagery found of past years on Google Earth Pro. In the case of my study, the findings can be applied for future research related to agricultural land, for other researchers to know what material and algorithms to use to obtain data on land changes throughout a period of time. Hence, this study falls under ecological and temporal generalization/validities. The Maltese agency ARPA would benefit greatly from my findings to provide the Commission, local entities, and the farming community with accurate and timely information.

Coding

Codes are identifiers given to text fragments that may be used on paragraphs, sentences, phrases, or single words (Delve, n.d.). In Qualitative coding, there is an approach for finding themes and patterns in unstructured material by classifying extracts in a methodical manner. This occurs during the study analysis and report writing stage after the research strategy has been developed, interviews have been done, and data has been gathered and prepared through transcribing interviews. When it's time to analyze and write, qualitative coding comes into play. Different types of qualitative data which are useful for qualitative coding include transcriptions, diary accounts, documents, case studies, audio and video recordings, notes and observations, and anything that can be analyzed using qualitative coding. Qualitative coding has several advantages, including being systematic and rigorous, assisting in the speedy identification of quotes, and serving as a tool for identifying patterns and themes as well as checking for biases.

Qualitative coding is divided into two types: **inductive** and **deductive** (Crosley & Jansen, 2020). Either one can be applied or combined as a hybrid. The inductive coding method derives codes from data from the ground up. At the start, there are no predetermined preconceptions about what the codes should be; instead, the story or theory is left to develop naturally from the raw data. This is ideal for exploratory studies or when trying to come up with fresh hypotheses, notions, or ideas. On the other hand, deductive coding is a top-down strategy in which a coding scheme with the first set of codes is created. This collection might be based on the research’s questions or a theory or framework that already exists. The data is then looked over and codes are assigned to extracts. The codes at the end of your analysis should seem quite similar to the coding scheme applied at the beginning. This is ideal to having a pre-determined format of how the final findings should look like. For instance, a deductive technique could be used in program assessment studies.

In Kuss et al.’s study (2022), a thematic analysis was utilized to analyse the data and it is stated that semantic codes were employed to develop independent frameworks after reading and familiarizing with the data. Recurring patterns were grouped together into larger themes and subthemes after comparison. Themes were compacted and redistributed throughout the research team during the reviewing phase to ensure internal homogeneity and external heterogeneity. This was done to ensure that there was internal consistency and that the themes were unique. During this phase, the data was rigorously recoded to ensure that the themes in use covered the entire dataset and were refined at the same time. As seen in Table 1, the participants were identified by the codes 'FG' for female gamer followed by a number for identification (FG1 is the first female gamer interviewed). Participants' quotes were referred to using the codes, which included their age in numbers and their professional position as key variables that indicate characteristics of the participants in relation to the meaning offered in an analytic topic or subtheme (ex. FG1 (33, Researcher).

Table

Description automatically generated

An inductive coding strategy was used because the majority of their interview questions were open-ended. The reason for this is that the research being conducted is not yet completely understood, and codes are progressed with their analysis, therefore the coding obtained aids in the exploration of the issue and allows researchers to investigate new theories or concepts. Following the collection of the requisite codes, four main themes with accompanying subthemes were identified:Diagram

Description automatically generated

Research triangulation

The approach of triangulation is used to boost the credibility and validity of a study’s findings (Noble, 2019), by combining numerous types and bits of data, which can result in a more accurate analysis or interpretation of a scenario (Yang and Cornelius, 2004). Validity refers to the amount to which research accurately depicts or assesses the notions or concepts being explored, while credibility relates to the trustworthiness and how convincing a study is (Noble & Heale, 2019). Researchers can benefit from triangulation since it provides a variety of datasets to explain various aspects of a situation. It may help in hypothesis rejection when one dataset invalidates a theory proposed by another, as well as hypothesis confirmation when one set of facts verifies another. Finally, triangulation can serve in the interpretation of research findings. The idea behind triangulation is that approaches that produce the same results offer researchers more confidence in their conclusions. However, triangulation comes with its limitations, for instance, it increases the complexity of the investigation, making it more time-consuming. When used to combine research methodologies, it may not be done in a uniform or consistent manner, hence researchers may not be able to completely characterize their mixing tactics. There are times when comparing the conclusions of two sources is unclear or conflicting. Addressing problems in a study design isn't always enough, and some research may overestimate the necessity of triangulation. Finally, triangulation is a difficult process that needs the assistance of a skilled analyst.

There are 4 types of triangulations:

1. **Data triangulation** - considers things like time, geography, and people,
2. **Investigator triangulation** - includes the use of multiple researchers in a study,
3. **Theory triangulation** - promotes the use of several theoretical schemes to enable interpretation of a phenomenon, and
4. **Methodological triangulation** - encourages the use of several data collection methods such as interviews and observations.

After the experiment was completed, I compared ARPA's data to that of the classifiers I used to validate my findings. In doing so, I utilized the Google Earth Pro application to first validate the results produced by the algorithms and see which land has been found to be listed as ‘cultivated’, ‘uncultivated’ and ‘other’, and compare to the images found on Google Earth Pro as ground truth data. Since I had to give the Random Forest Classification this dataset in order for it to be trained, there is a convergence with ARPA's dataset and this algorithm. As a result, parcels tagged as 'cultivated' in ARPA's dataset are still produced as 'cultivated' after the classification has been run. Due to this, the majority of the Random Forest Classification's resultant classes are noticeably more accurate in approximating Google Earth Pro imagery than those created by the K-means clustering. Finally, I compared the classifiers' findings to the dataset provided by ARPA, which I confirmed to be mostly accurate to the Random Forest Classification results, both for multi-temporal and mono-temporal images. My research is classified as **data triangulation** since it focuses on a study area and satellite images over a duration of time.

References:

Bhandari, P, 2021, *External Validity | Types, Threats & Examples*, Scribbr, Available at: <https://www.scribbr.com/methodology/external-validity/#:~:text=What%20are%20the%20two%20types,to%20other%20situations%20and%20settings>). (Accessed: 28 March 2022)

Crosley, J, Jansen, D, 2020, Qualitative Coding 101: How to code qualitative data, explained simply, Available at: <https://gradcoach.com/qualitative-data-coding-101/#:~:text=%E2%80%9CPigeons%20attacked%20me%20and%20stole,codes%20to%20categorise%20data%20extracts> (Accessed: 3 April 2022)

Delve, n.d., *Free Qualitative Data Analysis Course by Delve*, Available at: <https://courses.delvetool.com/view/courses/introqualitative/753386-3-approach-deductive-and-inductive-coding-methods/2200089-deductive-and-inductive-approaches-to-coding> (Accessed: 29 March 2022).

Delve, n.d., *The Essential Guide to Coding Qualitative Data*, Available at: <https://delvetool.com/guide#:~:text=Qualitative%20coding%20is%20a%20process,themes%20and%20patterns%20for%20analysis> (Accessed: 29 March 2022).

Giraldo, O, Garcia, A, Corcho, O, 2018, *A guideline for reporting experimental protocols in life sciences*, PeerJ, PMC5978404, 6: e4795 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5978404/>

Jones, B, 2016, *How to write an experiment protocol*, Available at: <https://bridger-jones.com/how-to-write-an-experiment-protocol/> (Accessed: 25 March 2022).

McCombes, S, 2022, *Sampling Methods | Types and Techniques Explained, Scribbr*, Available at: <https://www.scribbr.com/methodology/sampling-methods/#non-probability-sampling> (Accessed: 28 March 2022).

McLeod, S, 2019, *What are Independent and Dependent Variables?*, Available at: <https://www.simplypsychology.org/variables.html> (Accessed: 28 March 2022).

Noble, H, Heale, R, 2019, Triangulation in research, with examples, *Evidence-Based Nursing*, <http://dx.doi.org/10.1136/ebnurs-2019-103145>, 22:67-68.

Noob School, n.d., *Chapter 7 Quiz*, Available at: <https://quizlet.com/124154964/chapter-7-quiz-flash-cards/> (Accessed: 28 March 2022).

Williams, y, 2022, *External Validity in Psychology: Threats, Definition & Examples*, Available at: <https://study.com/academy/lesson/external-validity-in-psychology-threats-definition-examples.html> (Accessed 28 March 2022).

**TASK 7**

Tong et al.’s (2020) study has a well-structured conclusion, providing the necessary information for other researchers to understand what materials and methods were applied, the findings and how the study can be utilized for further research. The authors say that *“The accuracy of the cropland products, serving as a point of departure for the numbers reported here, should however be kept in mind when interpreting the fallow extent, as misclassifications of natural vegetation in the cropland class will propagate to the estimated extent of fallow fields.”* This rings true as pixels with natural vegetation in the cropland products would affect the classification of identifying fallow land since they will be similar.

Abdi’s (2019) conclusion gives a very in-depth overview of his study and results including recommendations. He says that algorithms executed with hyper-parameter values from a priori knowledge (knowledge that comes before the facts) can introduce bias, however, *“Random iteration across a defined number of parameter combinations can be used to eliminate a priori knowledge but at the cost of the algorithms not reaching optimum accuracies.”* I agree with the said statement since having a machine learning parameter whose value is determined prior to the training of a learning algorithm (hyper-parameter values) from an already known knowledge would create systematic and repeatable errors in a computer system that create unfair outcomes. Hence, when applying random iterations, training data will give models that make different predictions and have a different estimate of performance, in this case decreasing the accuracy, but the bias is eliminated. I also agree when he suggests that *“the presence of the red edge bands in the Sentinel-2 satellites might render the use of vegetation indices obsolete in boreal landscapes”* after finding out that red edge made up a fourth of the top twenty essential bands, but none of the spectral indices was highly ranked.

He and Zhao’s (2019) conclusion gives a very short and vague overview of their study, which I don't think is appropriate for a conclusion. In the sentence, *“Our TCN-based approach works well on three open real-world datasets.”*, they didn’t mention what the three open real-world datasets are, so one would need to go over the paper to find these datasets themselves. I do agree, though, how they simply convey that their findings prove TCN is a *“feasible model to learn normal time-series behaviors, and can be used for anomaly detection.”* Yet, a more detailed explanation of how they arrived at such a conclusion should have been provided.

References:

Abdi, A, M, 2020, Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data, GIScience & Remote Sensing, 57:1, 1-20, DOI: 10.1080/15481603.2019.1650447

He, Y, Zhao, J 2019, Temporal Convolutional Networks for Anomaly Detection in Time Series, *J. Phys.: Conf. Ser.* **1213** 042050

Tong, X, Brandt, M, Hiernaux, P, Herrmann, S, Rasmussen, L, V, Rasmussen, K, Tien, F, Tagesson, T, Zhang, W, Fensholt, R, 2020, The forgotten land use class: Mapping of fallow fields across the Sahel using Sentinel-2, Remote Sensing of Environment 239:111598, DOI: 10.1016/j.rse.2019.111598