Tutorial for digitizing and analyzing historical agroecological data

Introductory Note

This tutorial was developed within the scope of the **AgroecoDecipher project (ref. 2022.09372.PTDC)** to support researchers and technicians in the process of transforming historical paper-based information into structured digital data. By combining accessible tools with emerging technologies such as Optical Character Recognition (OCR) and Artificial Intelligence, this guide provides step-by-step instructions for:

1. Digitizing historical documents (text and maps);
2. Georeferencing and vectorization using GIS platforms (ArcGIS Pro and QGIS);
3. Text processing using OCR tools and AI-based correction (OpenAI/ChatGPT).

The tutorial adopts a practical, application-oriented approach aimed at ensuring reproducibility and efficiency in the digitization and spatial analysis of historical sources. It is intended for researchers, students, and professionals in the fields of geography, environmental history, agriculture, and geographic information systems.

Requirements

Before starting, make sure you have:

- A tablet or smartphone with the Adobe Scan app
- An Amazon Web Services (AWS) account
- An OpenAI account
- Access to ArcGIS Pro or QGIS

How to cite:
Viana, C. M., Carvalho, D. (2022). Tutorial for digitizing and analyzing historical agroecological data. AgroecoDecipher Project (Ref. 2022.09372.PTDC), CEG/IGOT – Universidade de Lisboa.

Note: The tutorial was developed during the public release of ChatGPT.

December, 2022

2022.09372.PTDC
LA/P/0092/2020
UIDB/00295/2020
UIDP/00295/2020

Table of Contents

1. Digitizing historical documents (text and maps)

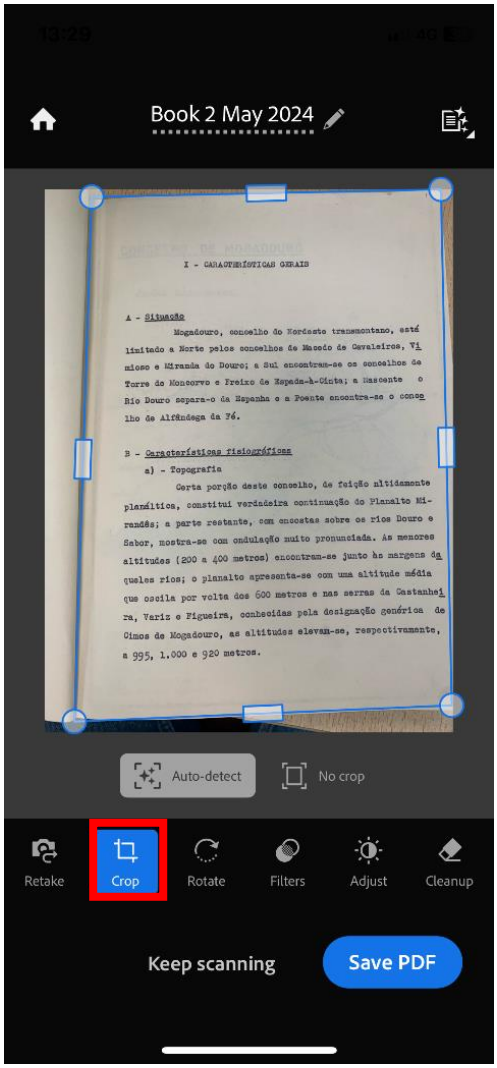
The primary goal of digitization is to convert information from paper format into digital form for study or other purposes. To achieve this efficiently, we can use the Adobe Scan app, available for download on both the App Store and Google Play Store.

Part I - Convert text and paper maps to PDF with improved quality for OCR

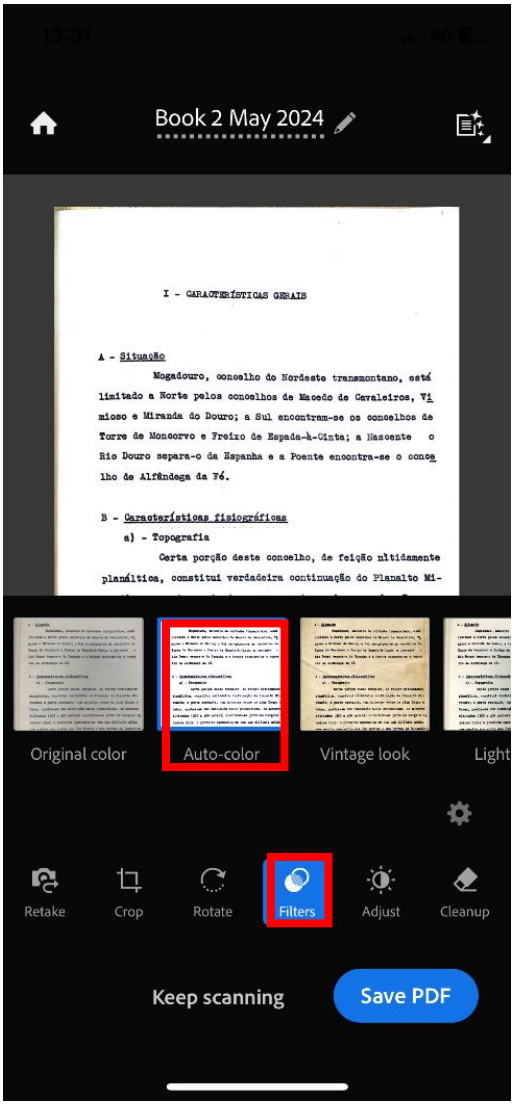
1.1. Open the Adobe Scan app.

1.2. Align and crop pages properly. Ensure the page is well-centered by adjusting the four corners. While the app usually detects this automatically, in some cases, you may need to use the "crop" tool to manually adjust the points to their respective corners.

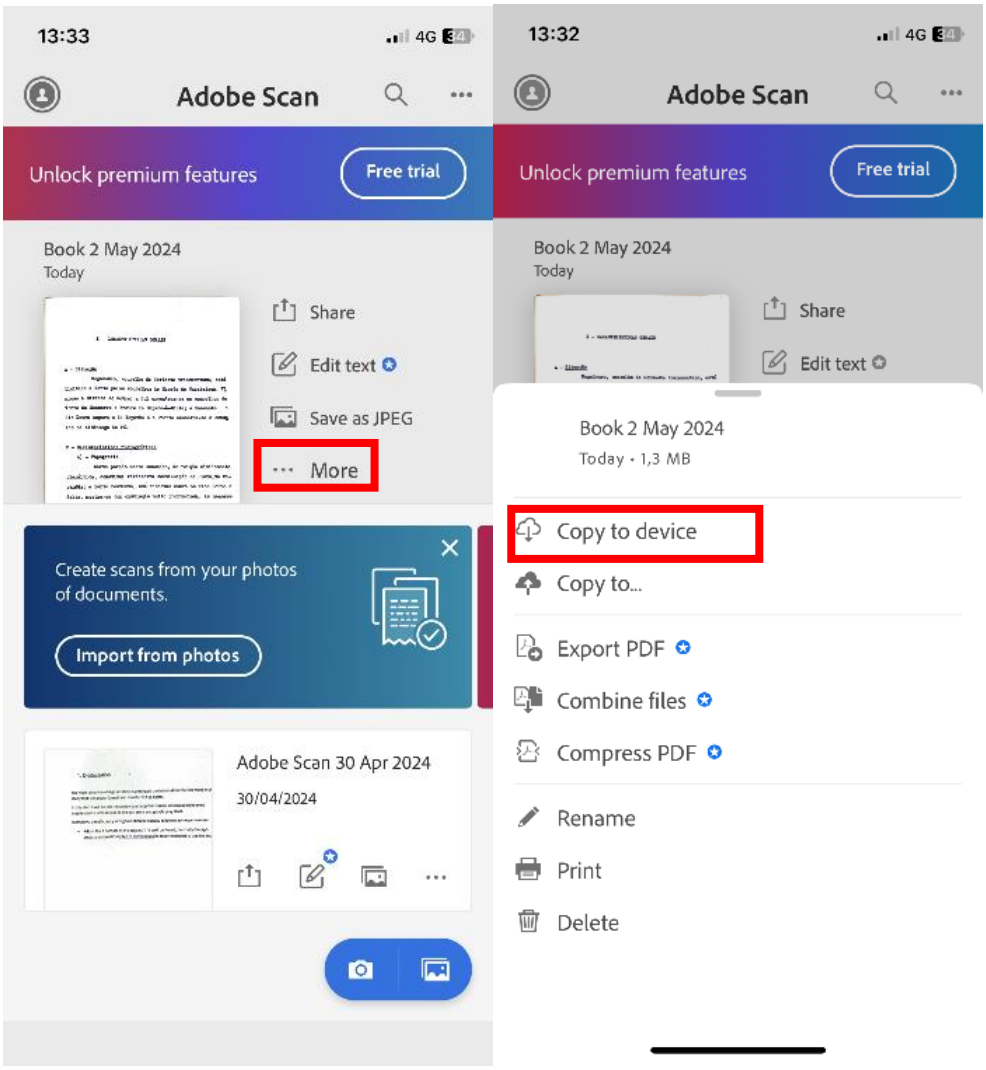
1.3. Continuous Scanning: After correctly aligning the page, click "keep scanning" to capture additional pages. Repeat this process until all pages are digitized.



1.4. Use the auto-color filter to enhance contrast. Utilize the "filters" tool and select the auto-color option. This filter enhances contrast, facilitating Optical Character Recognition (OCR) processes. Apply this filter to all pages by clicking "apply to all pages."

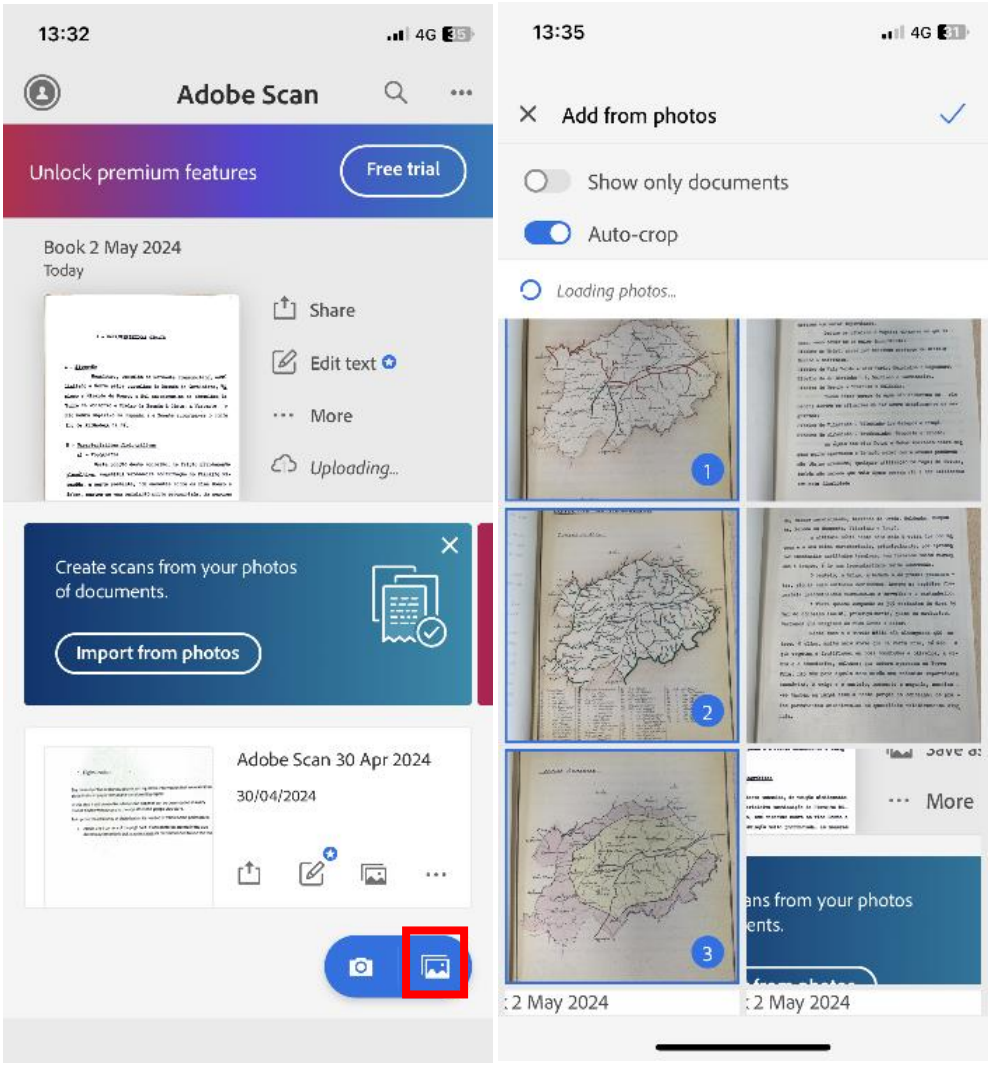


1.5. Scan all pages and save/export as PDF. Save the scanned documents as a PDF file. Export it to your desired location by selecting "more," then "copy to device," and finally, save it. You can further export it from your device's files to a cloud drive for access on your computer.

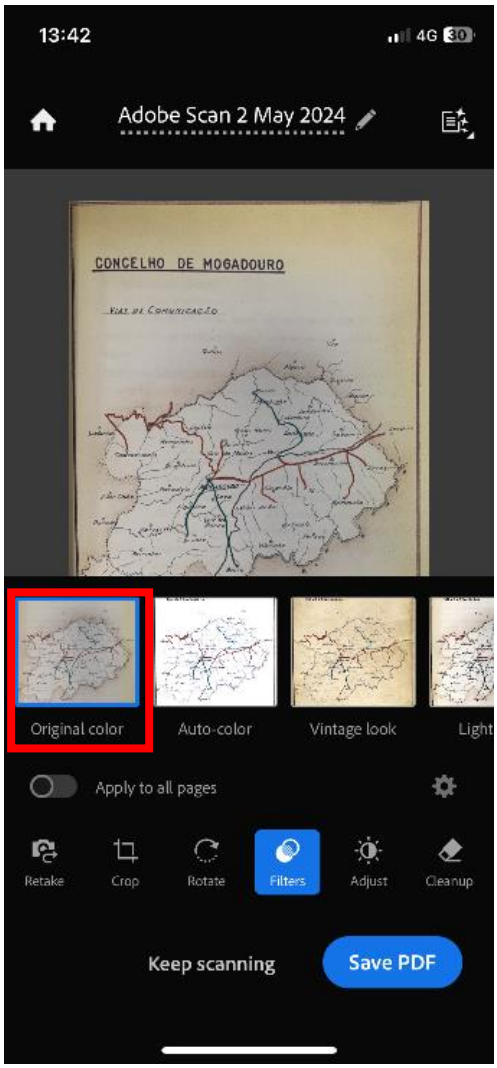


Part II - Vectorizing Historical Maps

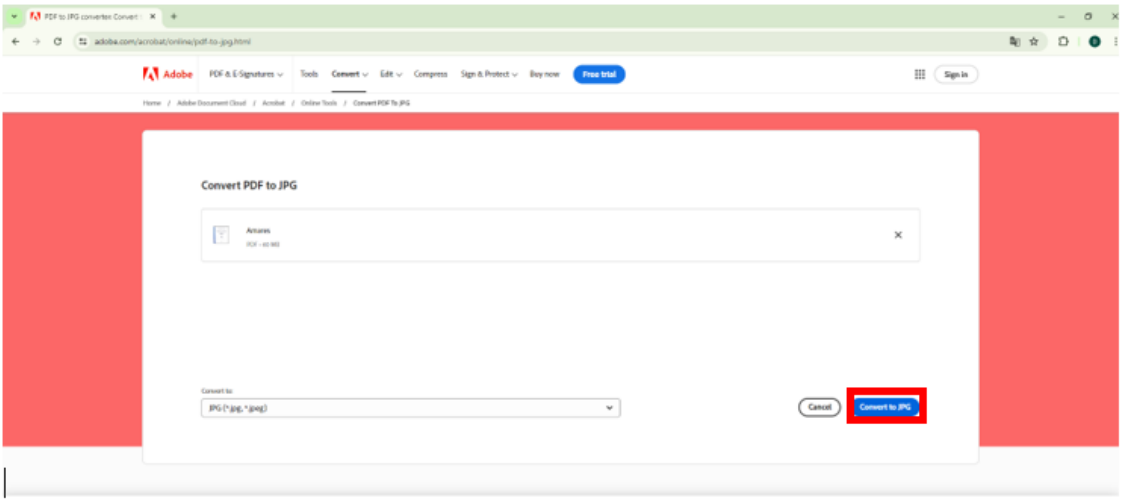
2.1. Convert PDFs to JPG. Open the Adobe Scan app and choose the option with a camera icon. This will take you to your photo library, where you can select the maps you want to work with.



2.2. Upload your scanned maps and convert. Once you've selected your maps, apply the original color filter, and save them as PDF files.

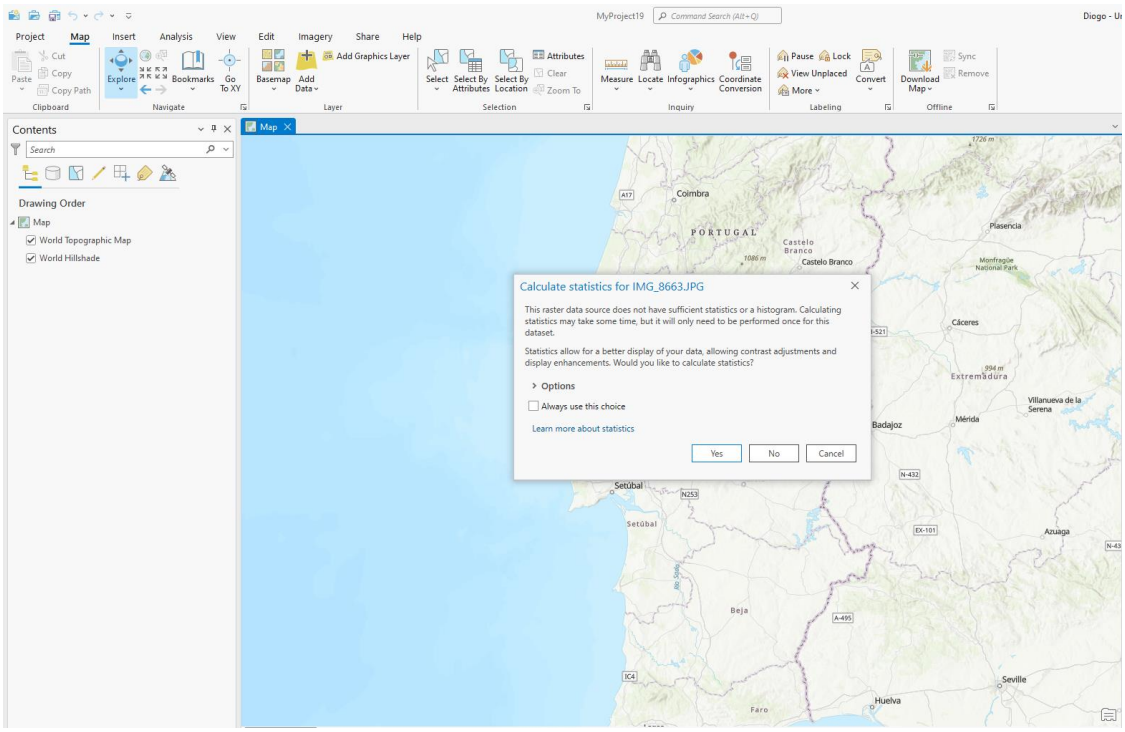


2.3. Open your web browser and search for "Adobe Convert PDF to JPG" (https://www.adobe.com/acrobat/online/pdf-to-jpg.html). Drag and drop your PDF file into the converter or use the file selection option. Then, click "Convert to JPG" and download the converted files.

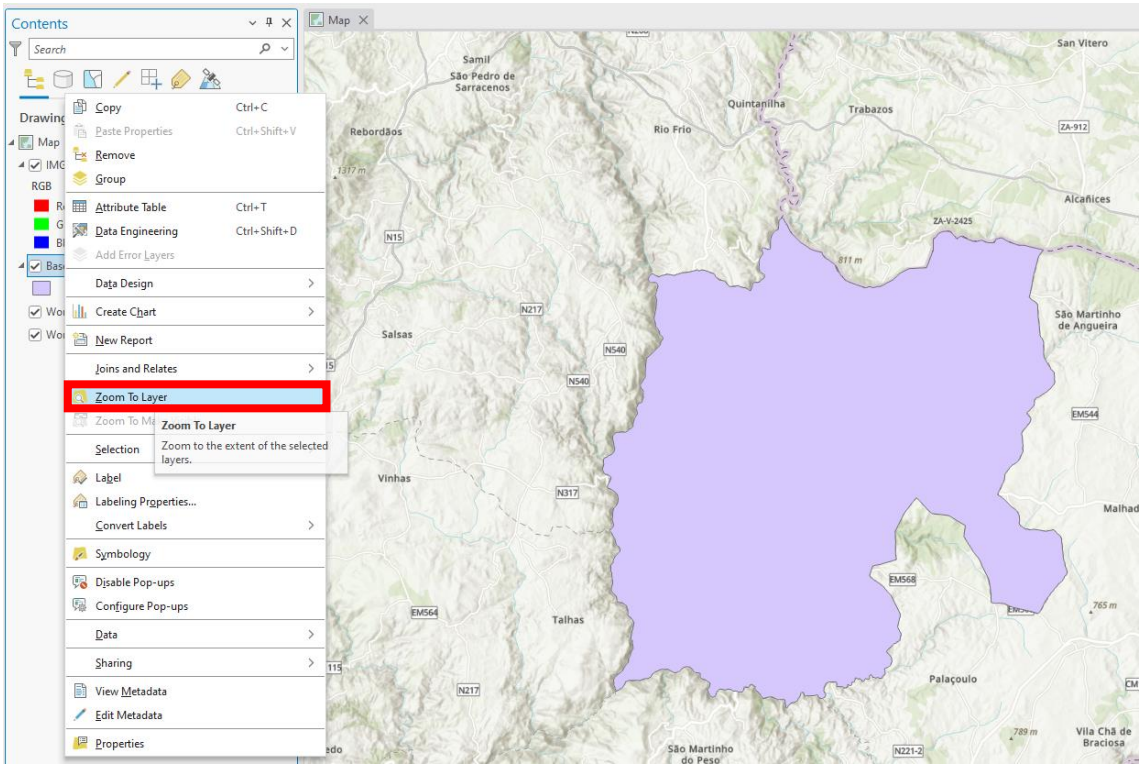


PART III - Georeferencing and Editing Maps

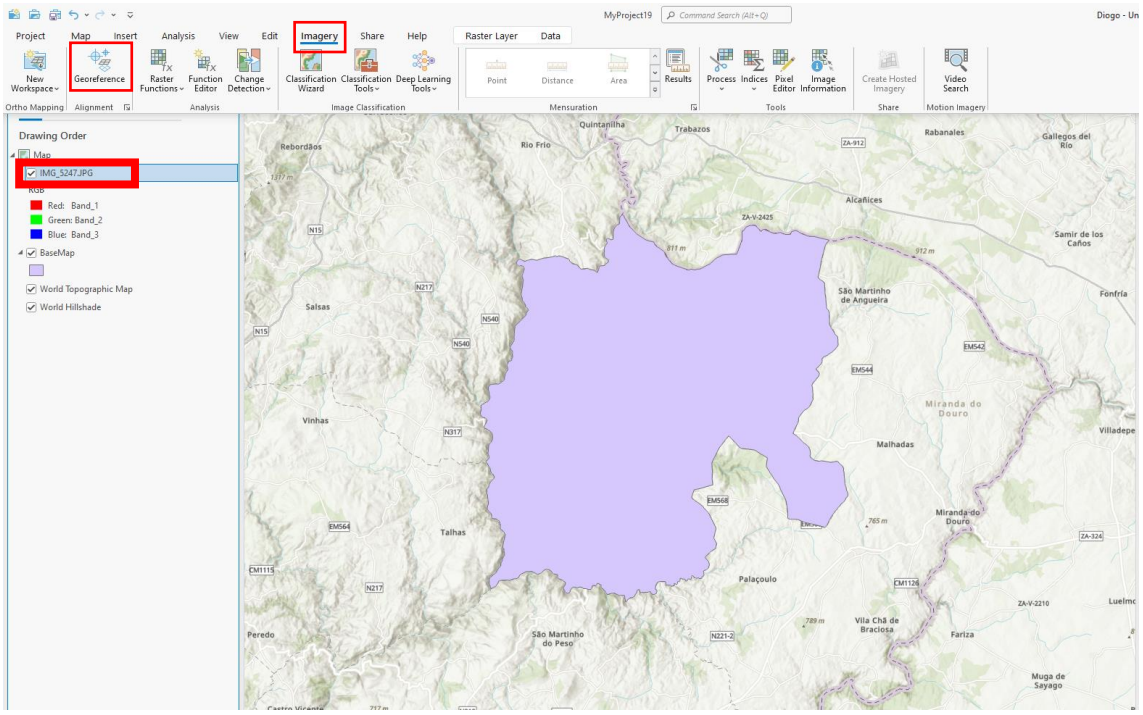
3.1. **In ArcGIS Pro**. Import JPG into the environment. Open a shapefile that has base format of your map and drag the jpg map to the environment.



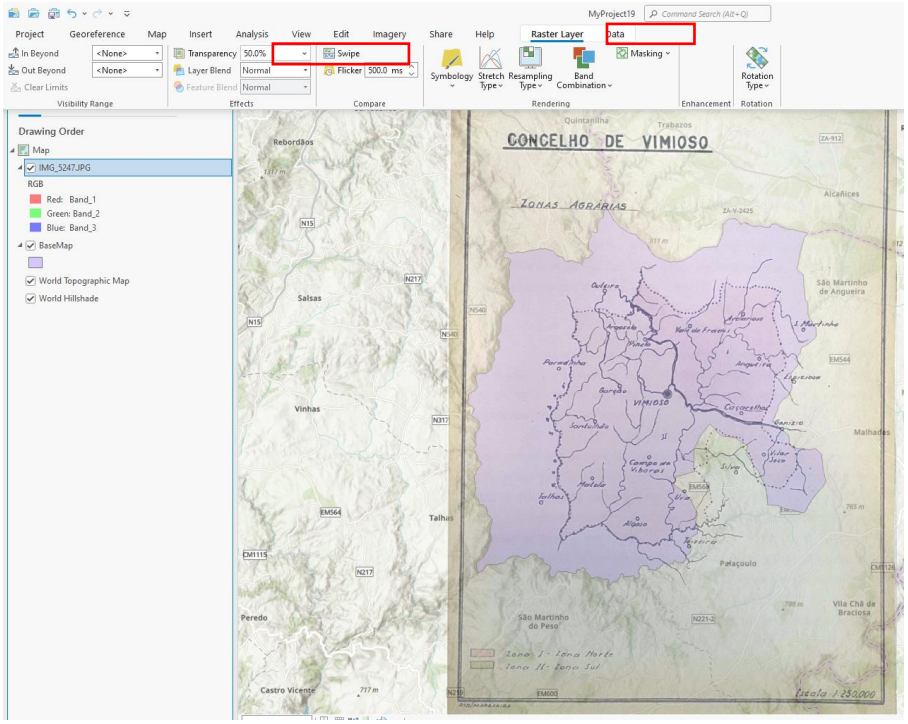
3.2. Right click on your shapefile and select Zoom to Layer.



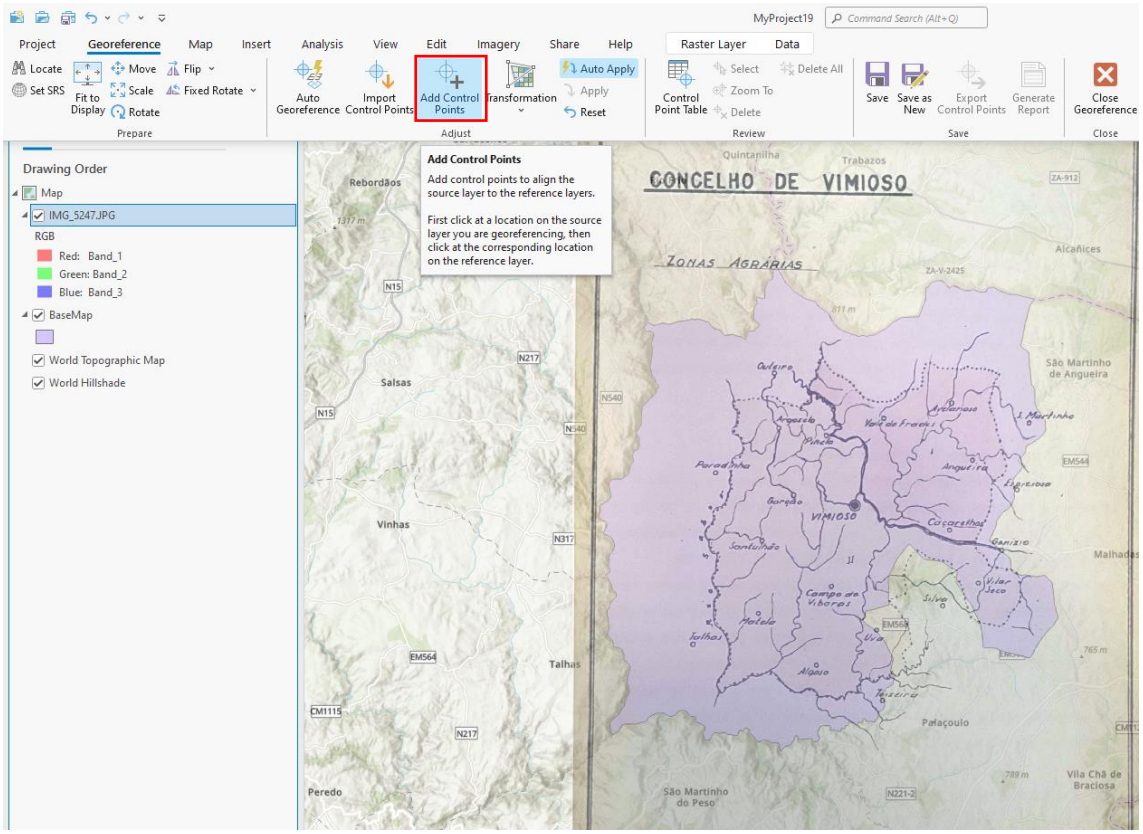
3.3. Use “Fit to Display” under the Georeference tab. Select your image and go to the Imagery tab and click georreference and click “Fit to display”.



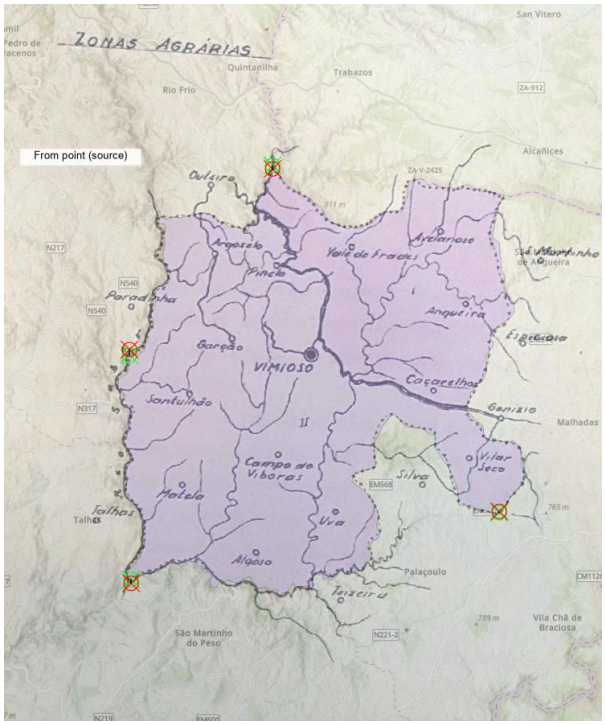
3.4. Go to “Raster layer” tab and define transparency to 50%.



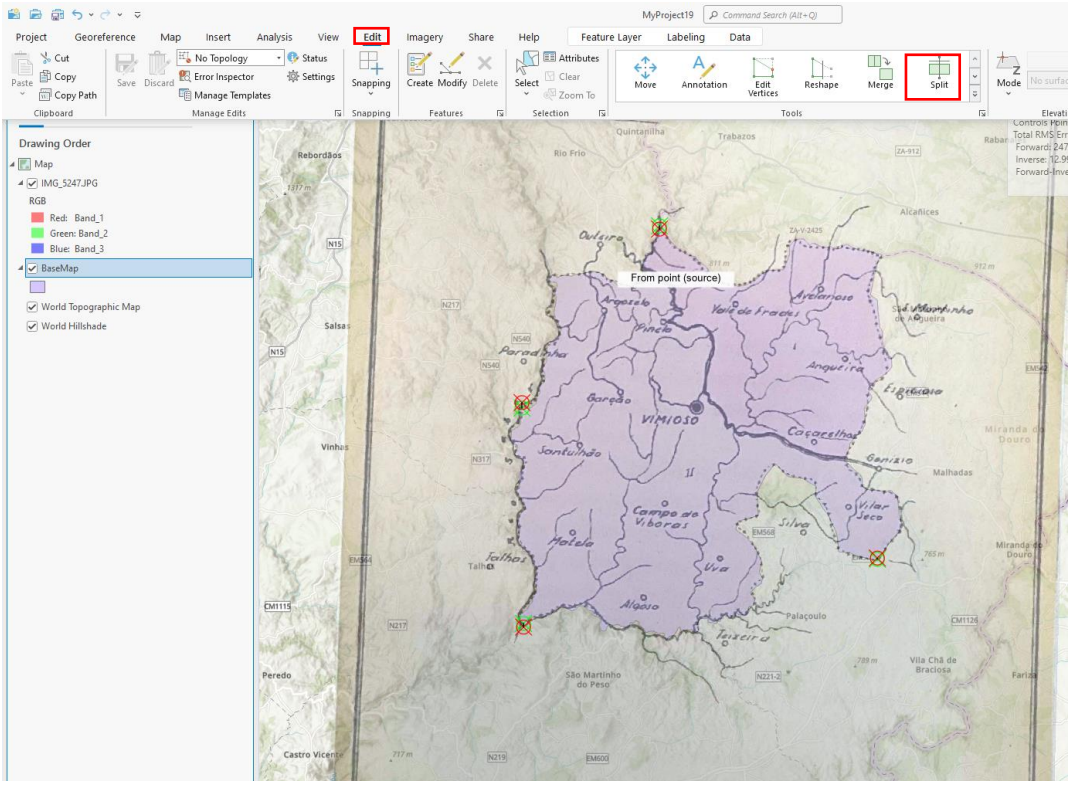
3.5. Add control points to align image with the base shapefile. Back to Georeference tab and use the tool “add control points”.



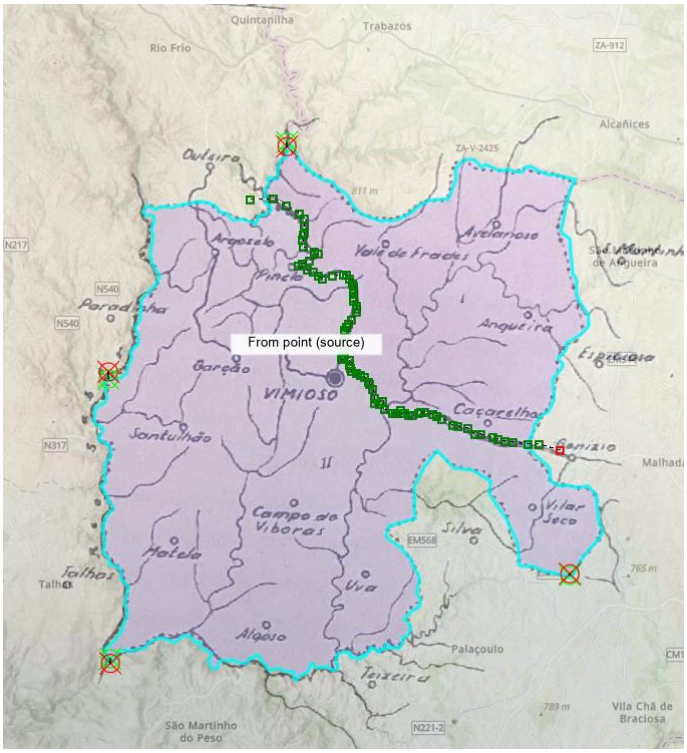
3.6. The first click will be the source point (your image) and you should make it correspond to the target location with the second click (base map).



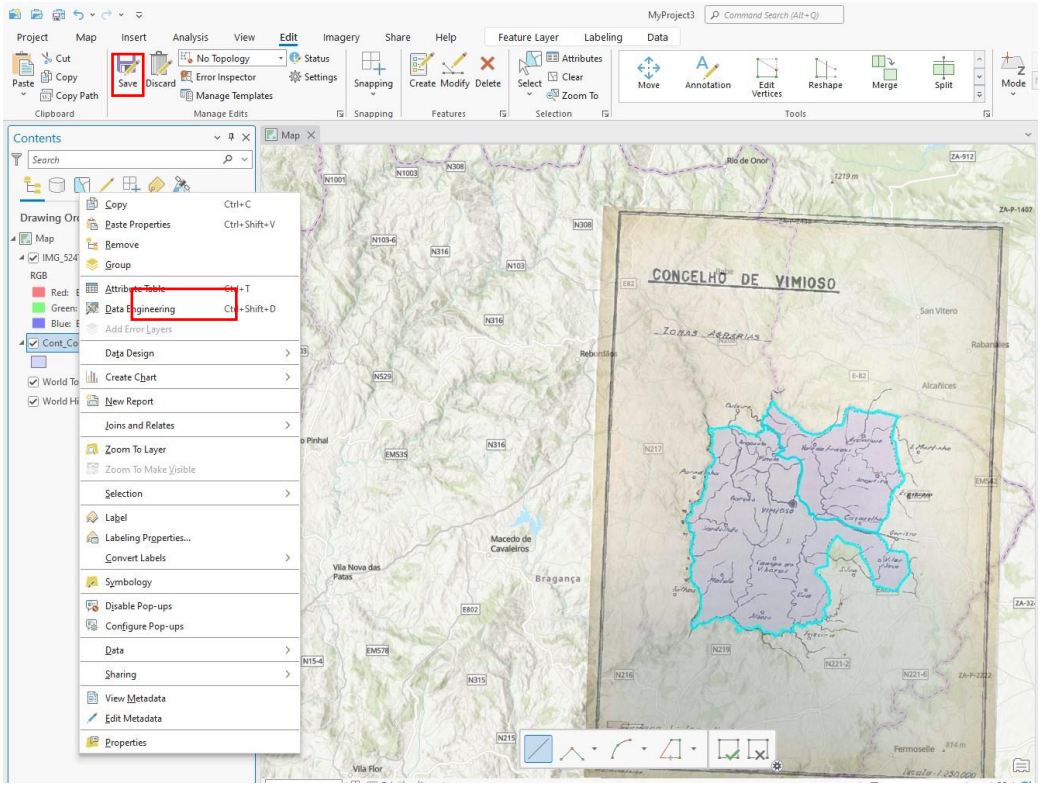
3.7. Save and close georeferenced. Use the “Split” tool to digitize and classify regions.



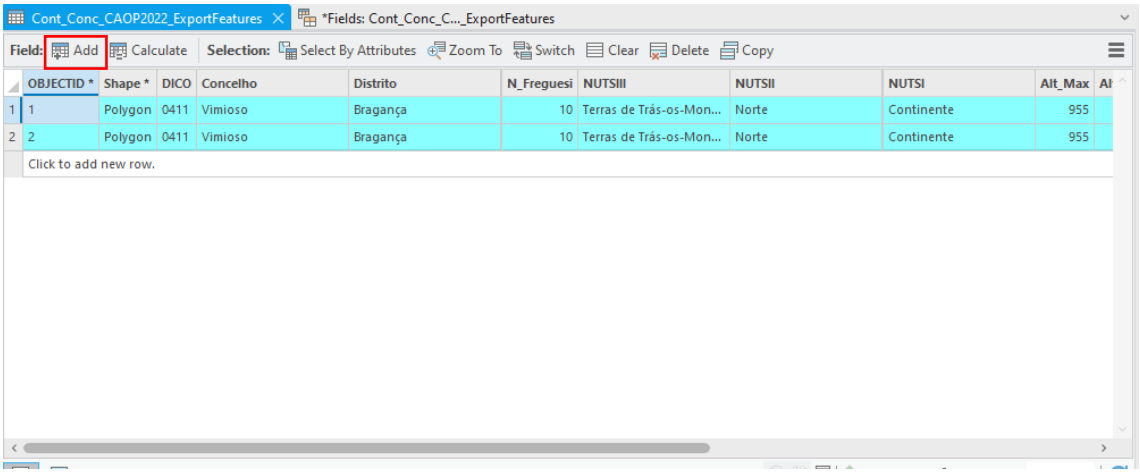
3.8. Select your base map by clicking on it and trace the the limits of the features that you want to define (pay attention that the line that you trace must completely cross the shapefile), click 2 times to confirm that your shapefile is divided as you want.



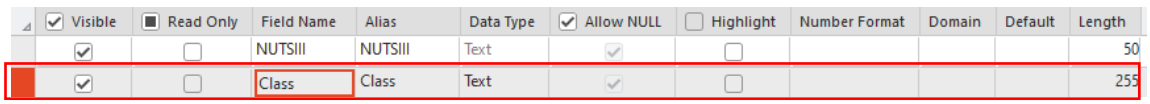
3.9. Save your edits and right click again on your shapefile and select attribute table.



3.10. Click “add”.

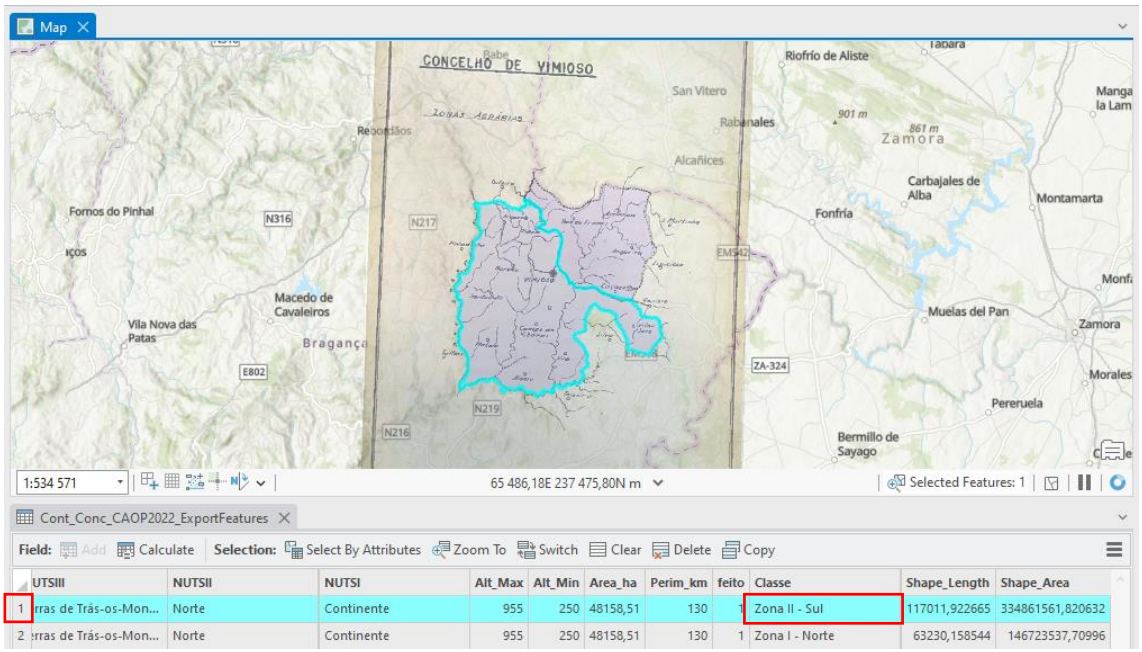


3.11. Create a new column where you will name the division that you made on your shapefile named class,close the temporary table and back to attribute table.

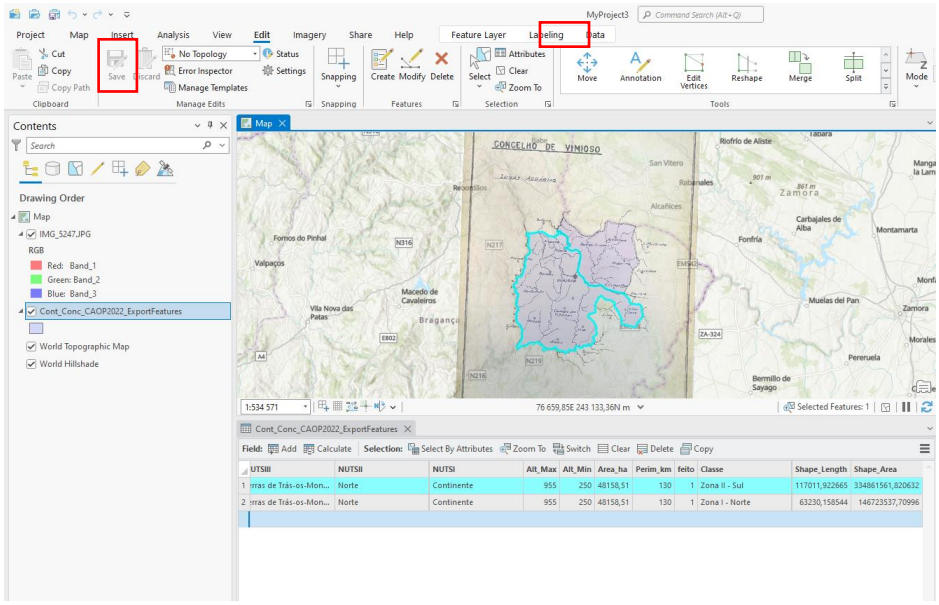


3.12. Click in one line to see which part of your shapefile you are selecting and, in the column, “class” make it correspond to the desire category and how you want to named it.

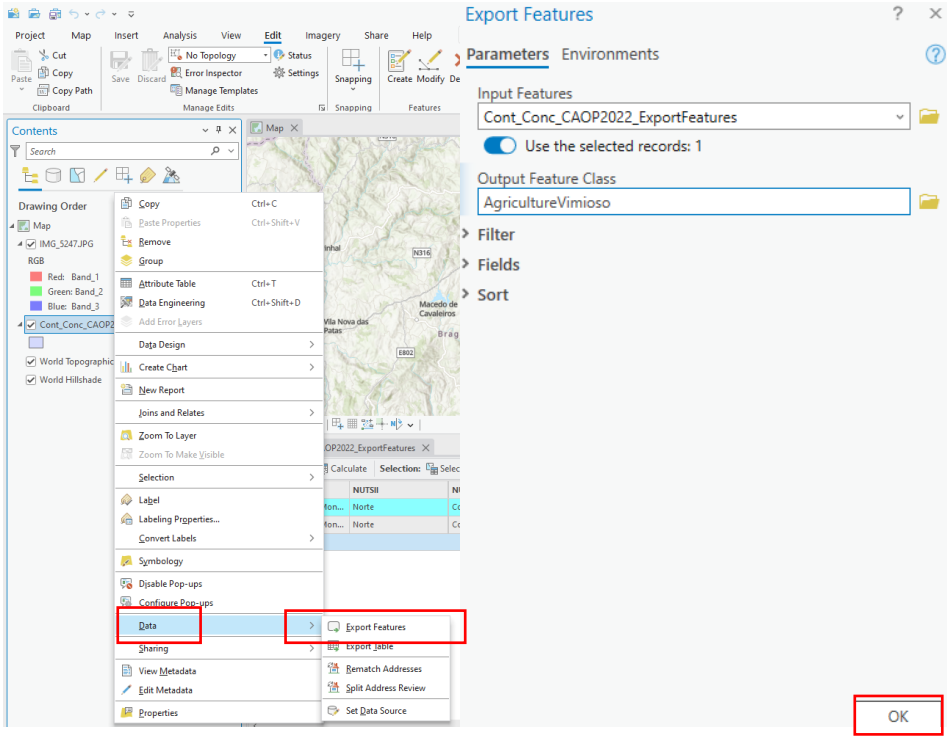
	UTSIII	NUTSII	NUTSI	Alt_Max	Alt_Min	Area_ha	Perim_km	feito	Classe	Shape_Length	Shape_Area
1	rras de Trás-os-Mon...	Norte	Continente	955	250	48158,51	130	1		117011,922665	334861561,820632
2	rras de Trás-os-Mon...	Norte	Continente	955	250	48158,51	130	1		63230,158544	146723537,70996



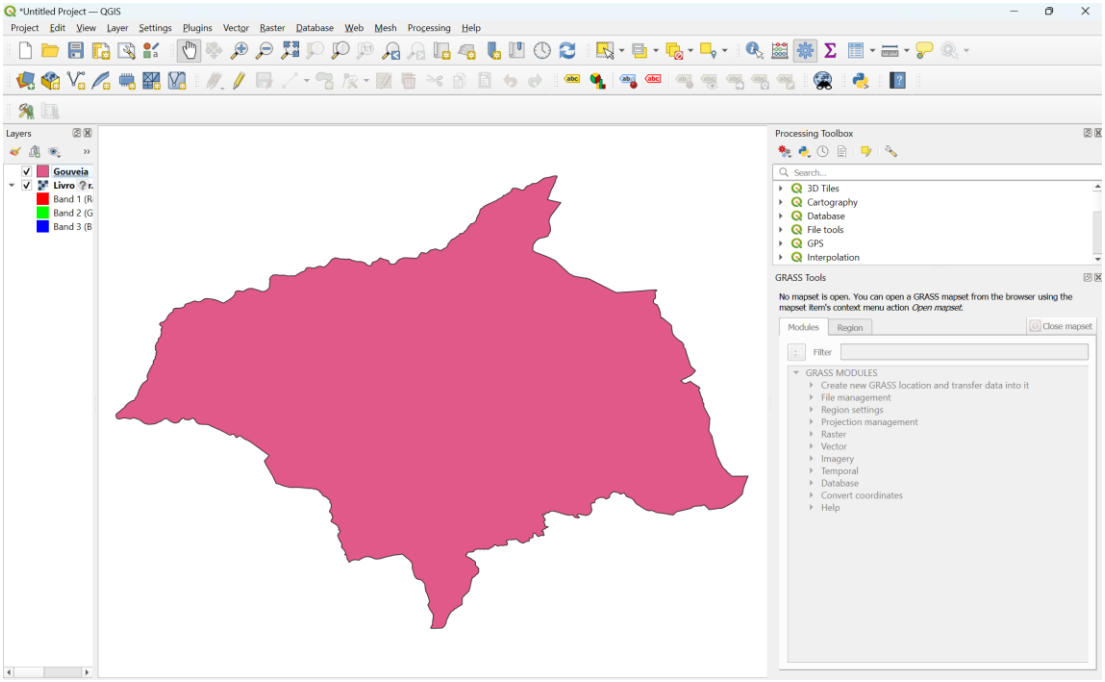
3.13. Save your edits and clear all selection.



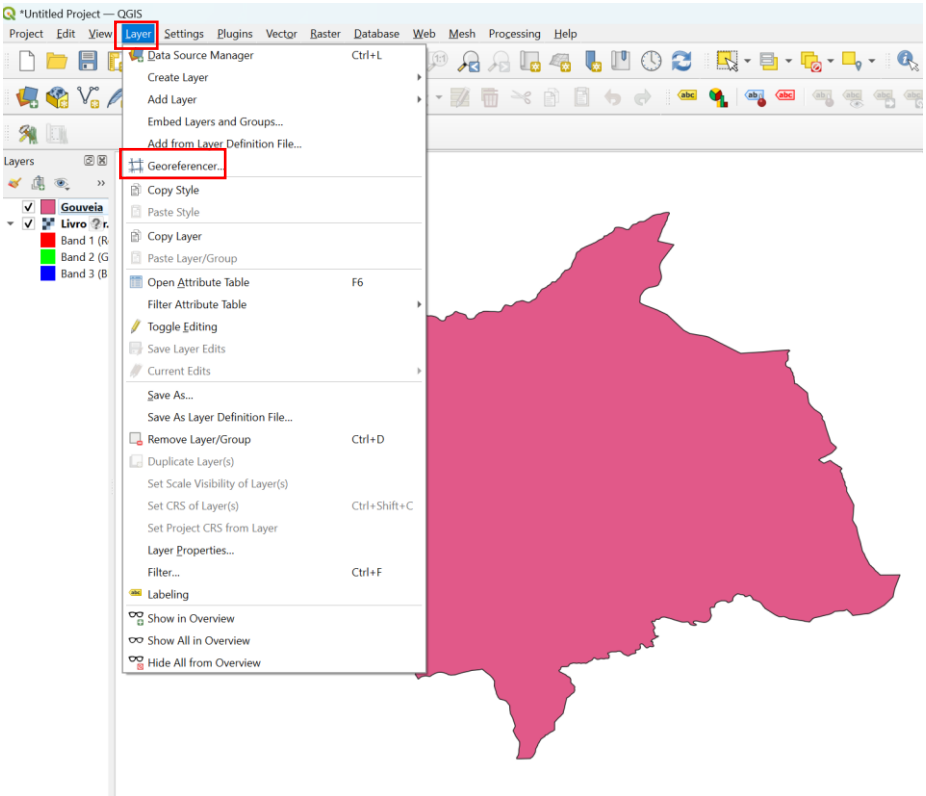
3.14. Export your shapefile (Data > Export Features).



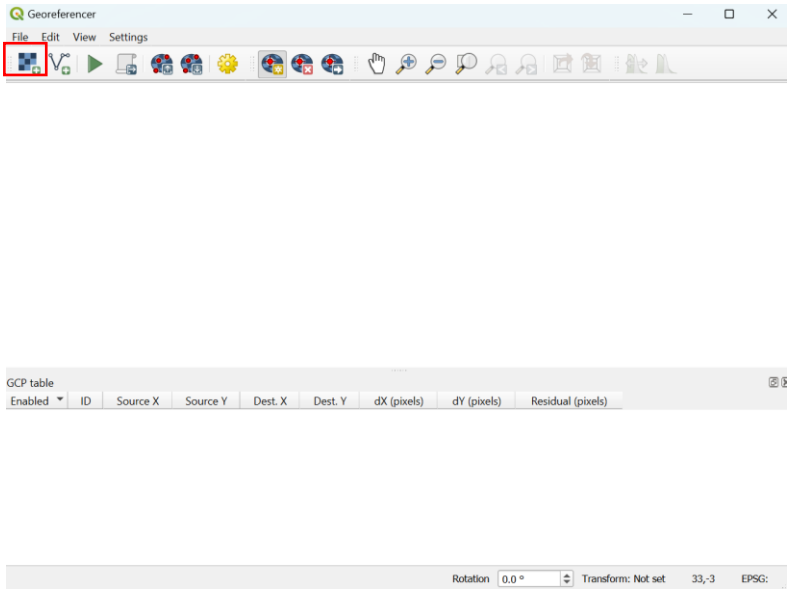
4.1. In QGIS. Open a shapefile that has base format of your map and drag to the environment.



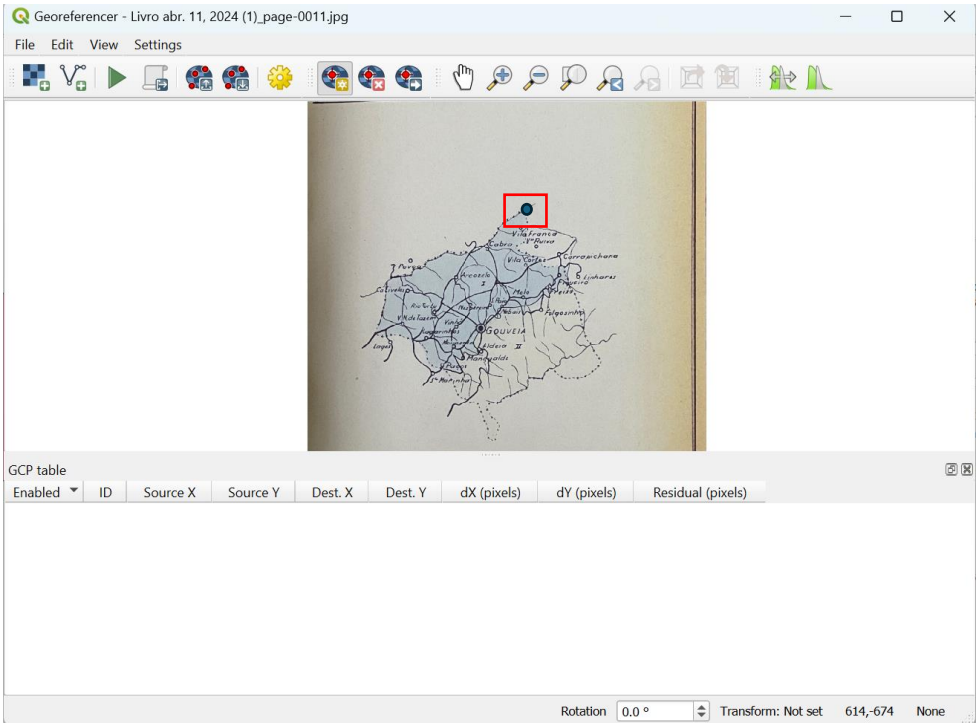
4.2. Open the Georeferencer tool. Go to the layers tab and select the Georeferencer Option.

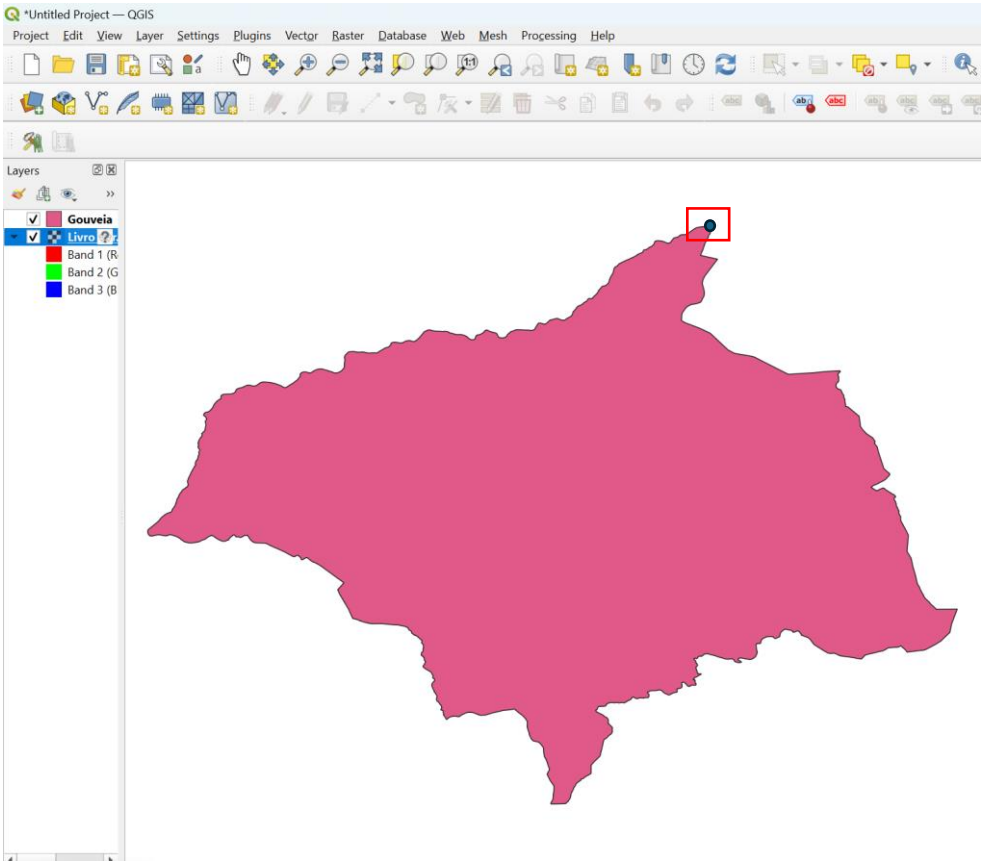


4.3. Click on the first option that is the raster layer and select the image of the map that you want to use.

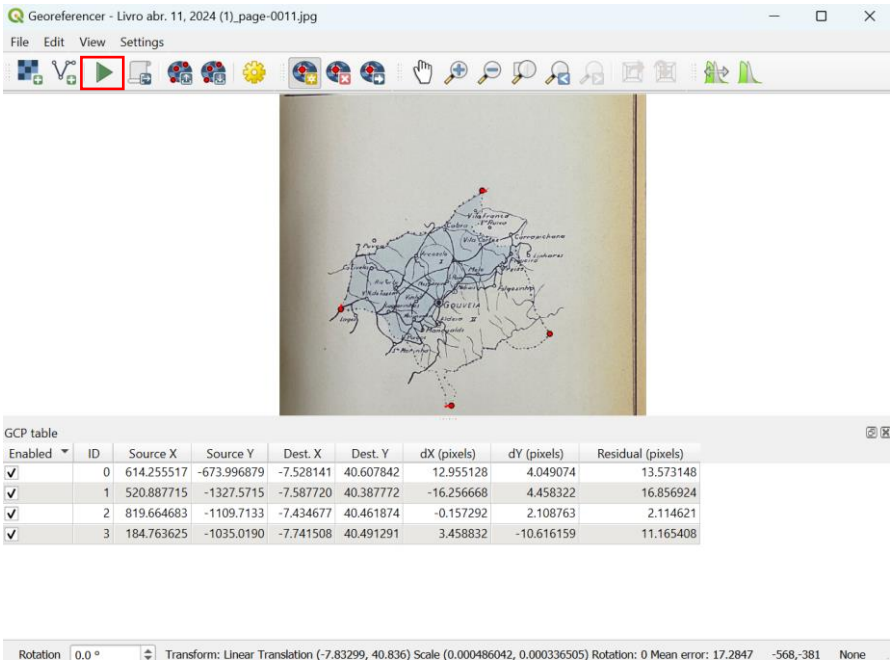


4.4. Once you have your map displayed, choose an edge of it and make it correspond to the shapefile edge.

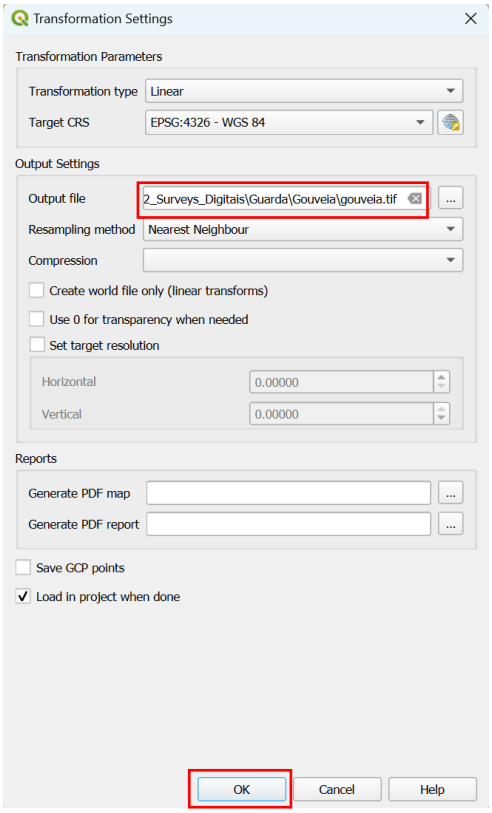




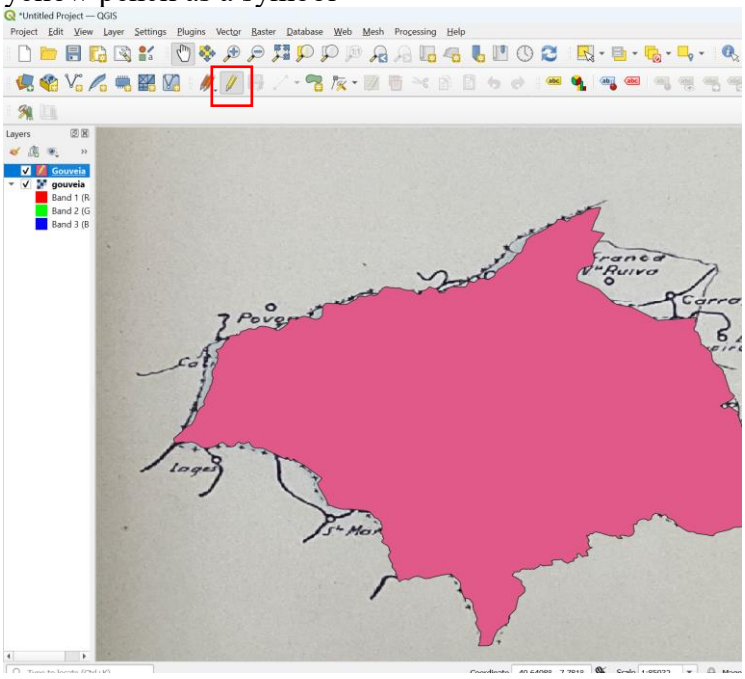
4.5. Add at least four control points to align the image and select “Start Georeferencing”.



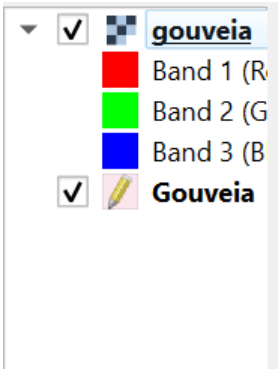
4.6. Choose a name for your output file and select “Ok”-



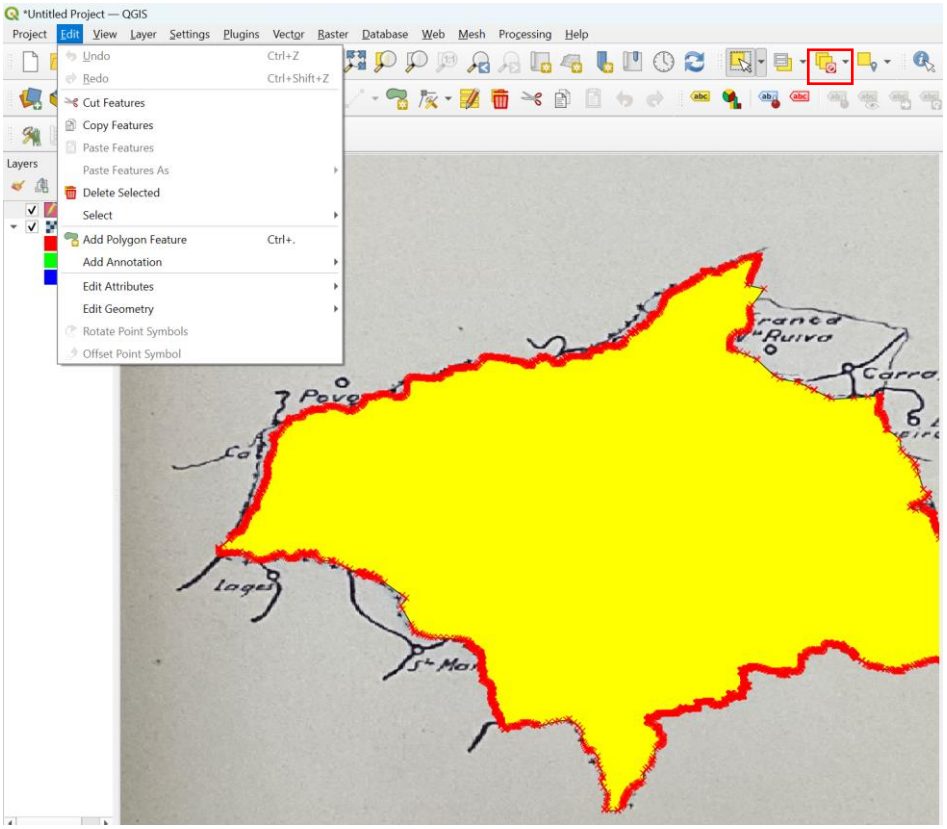
4.7. Enable editing and use the Split Features tool. Use the edit option that has a yellow pencil as a symbol



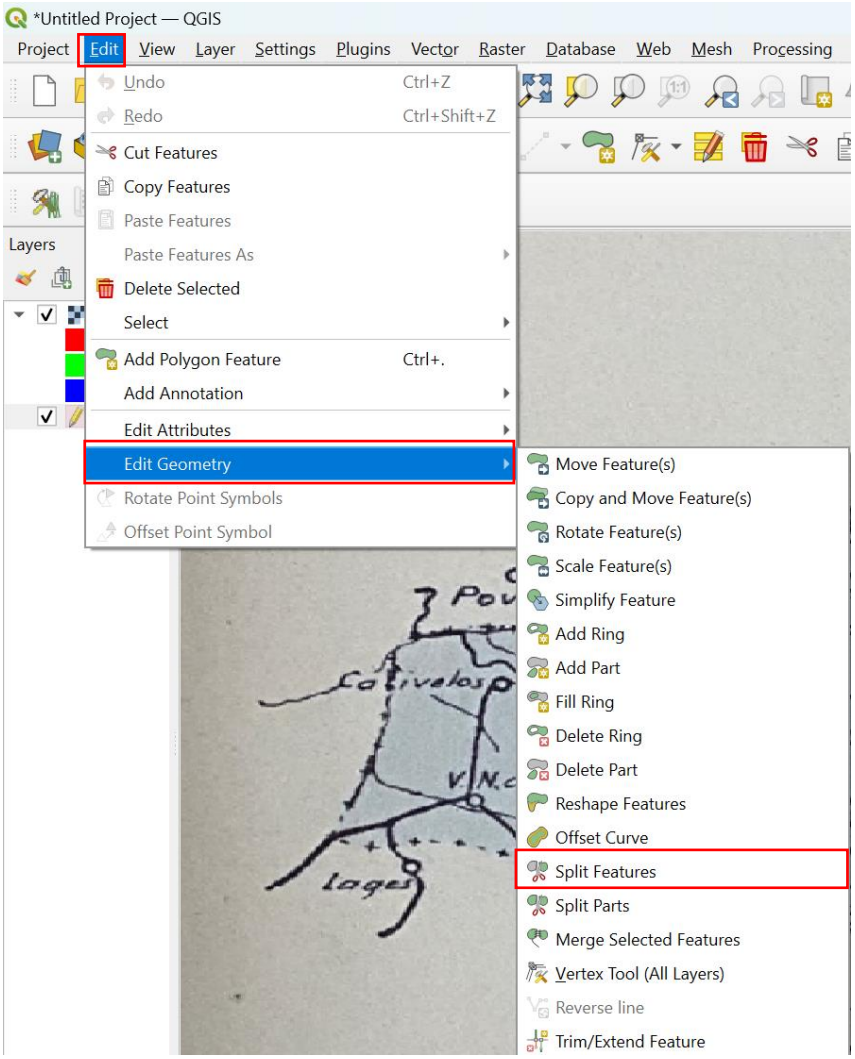
4.8. Make sure to put the map image above the shapefile in the layers option



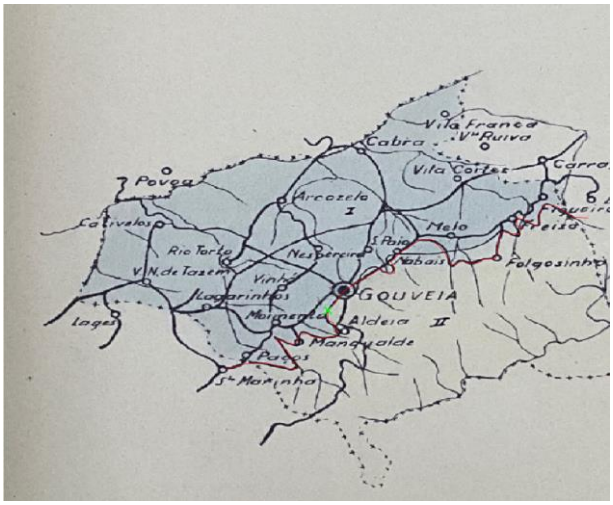
4.9. Select your shapefile



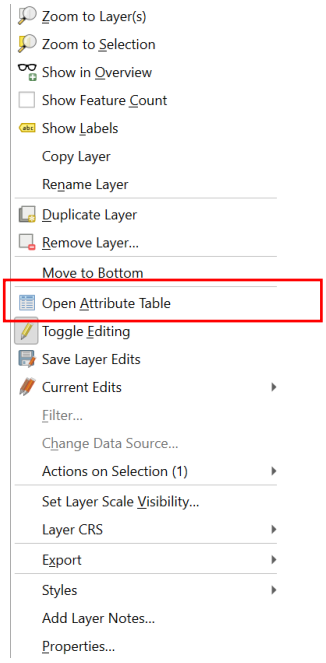
4.10. Go to the edit tab and select edit>edit geometry>splitfeatures.



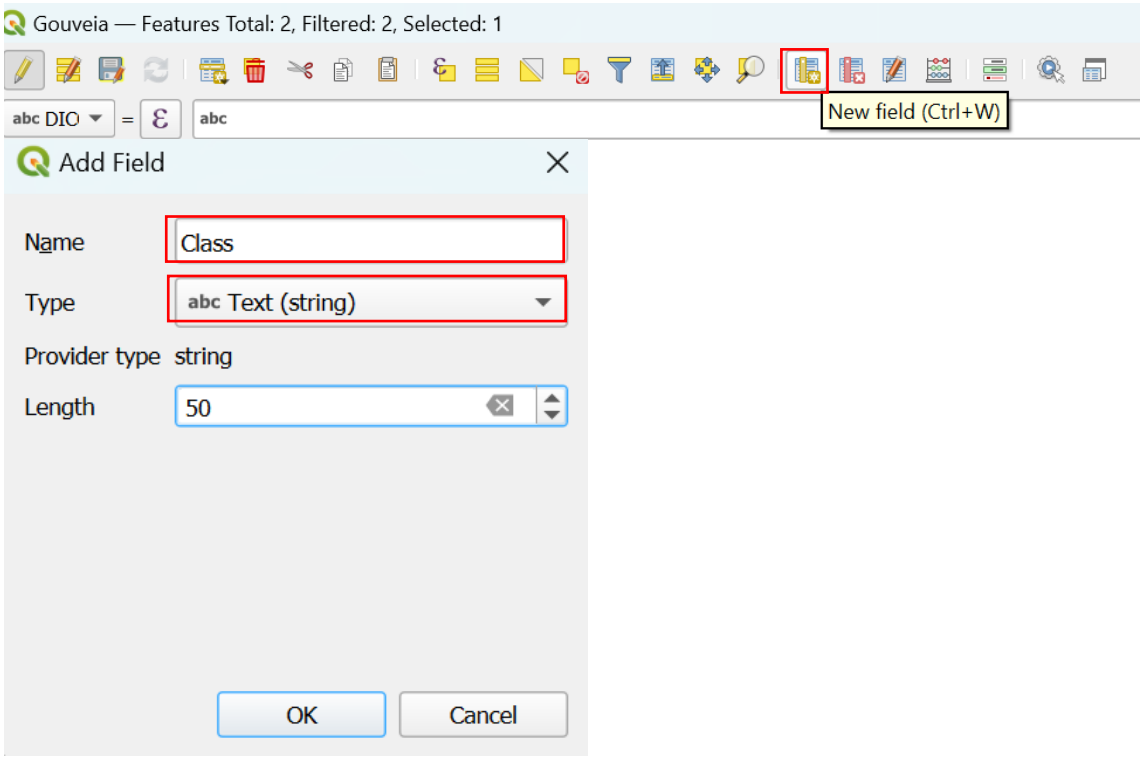
4.11. Trace the the limits of the features that you want to define (pay attention that the line that you trace must completely cross the shapefile), click one the right button of your mouse to confirm that you've finished.



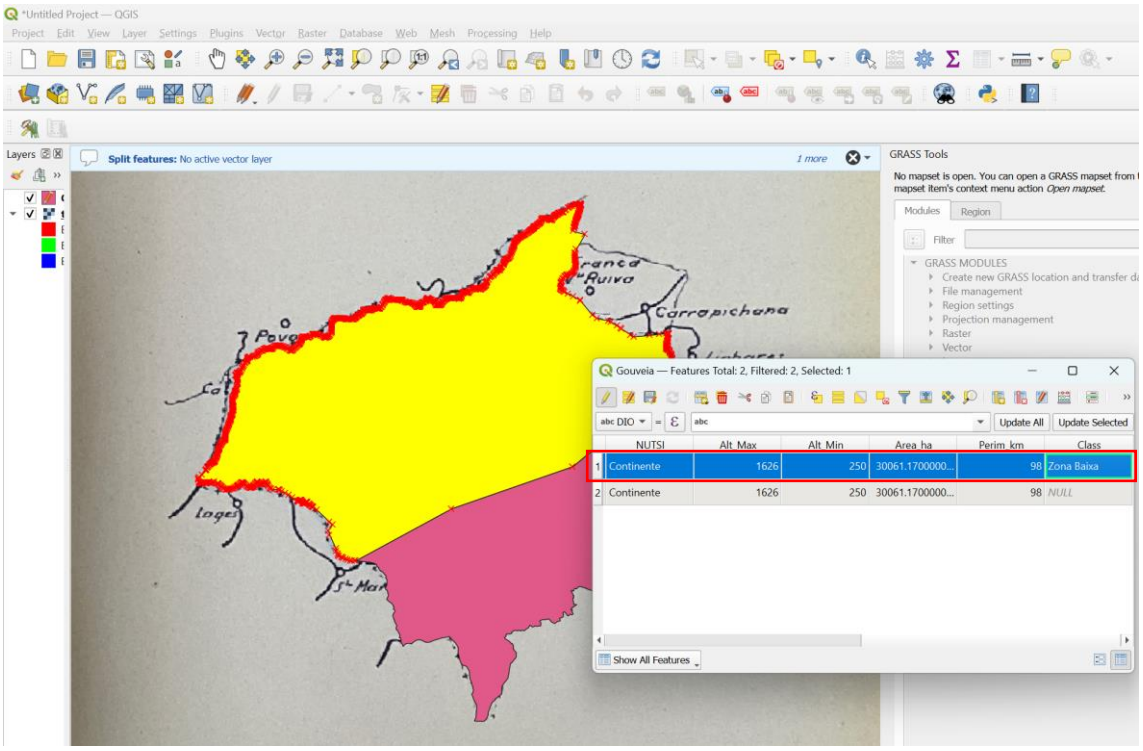
4.12. Once you have the features divided as you need right click on your shapefile and open your attribute table.



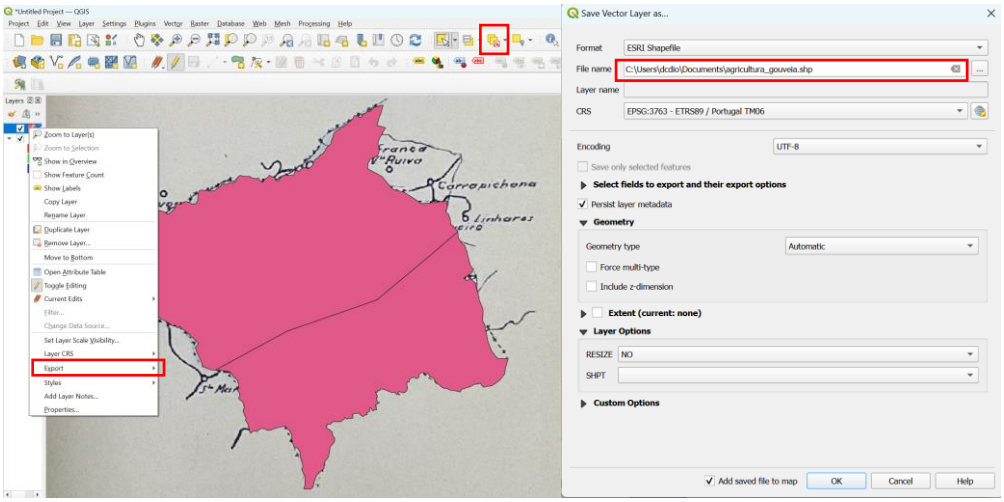
4.13. Attribute new fields (e.g., "class") for classification. Select “New Field” and create a field, you should select text(string) on the type tab.



4.14. Click in one line to see which part of your shapefile you are selecting and in the column “class” make it correspond to the desire category and how you want to named it and then save your edits.

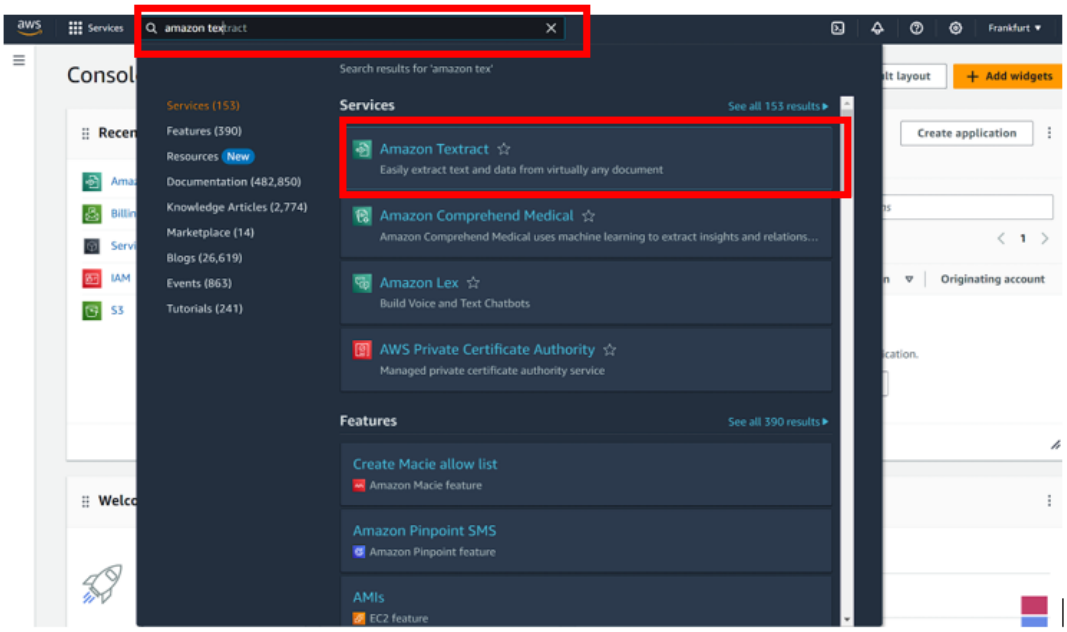


4.15. Deselect features from all layers and right click on your shapefile and click “ Export”, named it and click “Ok”

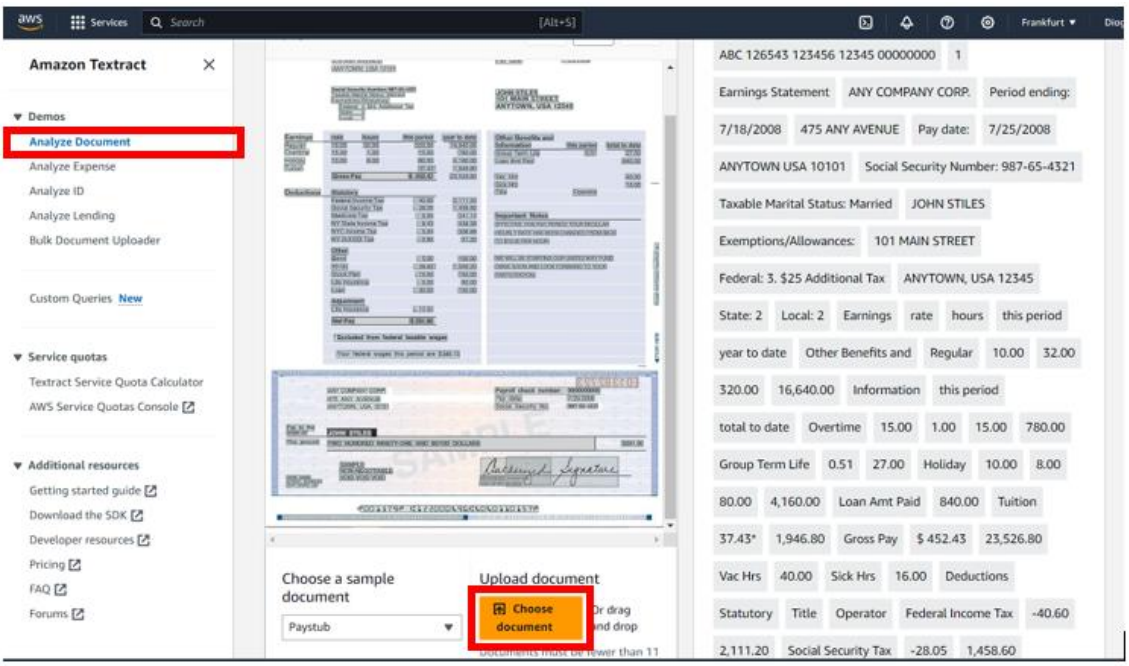


Part IV - OCR: Extracting Text from Scans

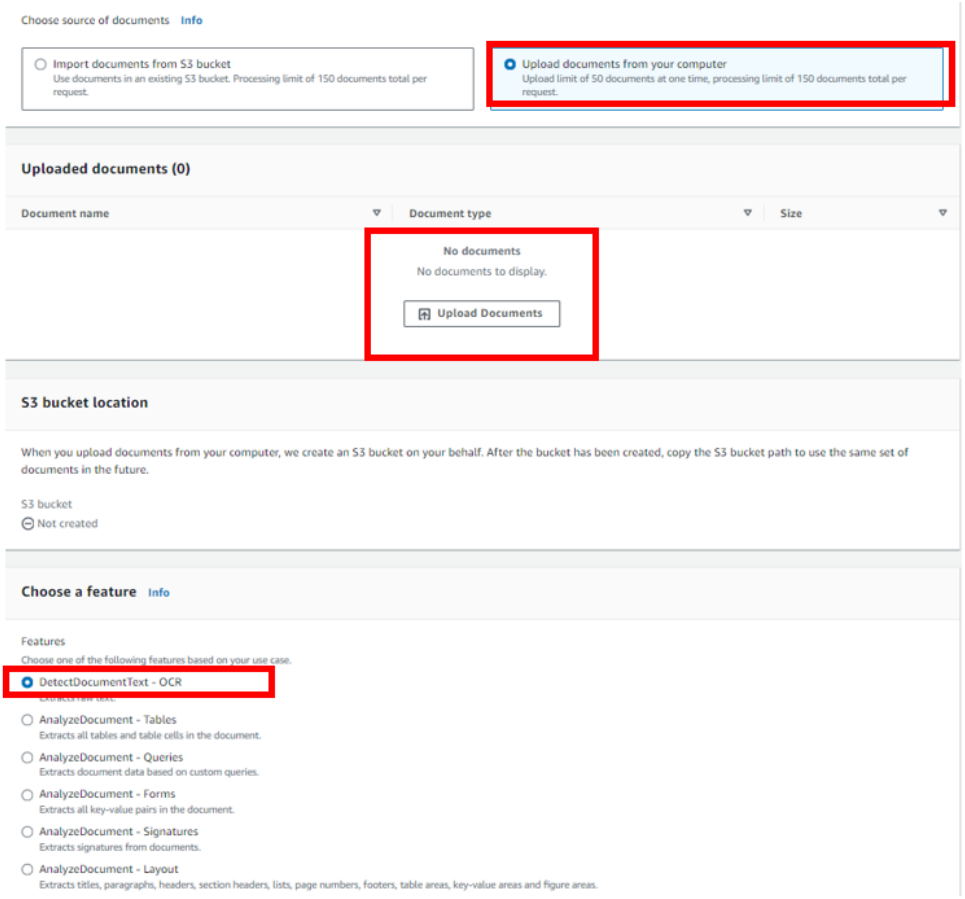
2.4. Log into your AWS account. Search for Amazon Textract.



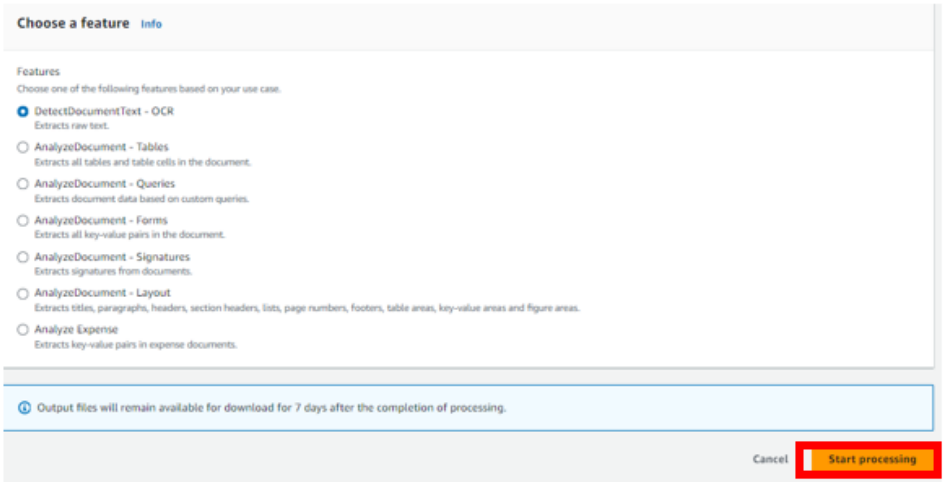
2.5. Upload your PDF in the Analyze Document section. Click it and select the option "Try Amazon Textract" in some regions the textract won't be available, try to choose one that fits your case. In the Analyze Document tab, go to the bottom of the page and select "Choose Document".



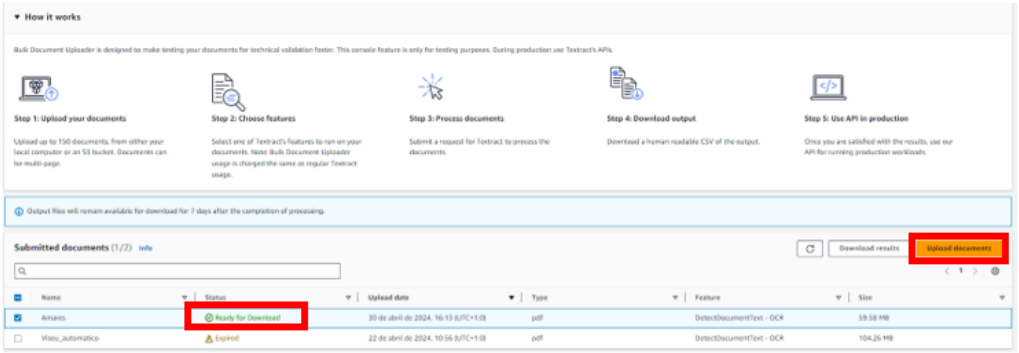
2.6. Select DetectDocumentText and start processing. The source of documents will be your computer and the feature that will be used is DetectDocumentText- OCR. Choose the pdf that you want to apply the OCR by selecting “Upload Documents”.



2.7. Select “Start Processing” at the end of the page.



2.8. Download the TXT output file. When the document status appears as ready for download, select it and click the option “Download Results”.

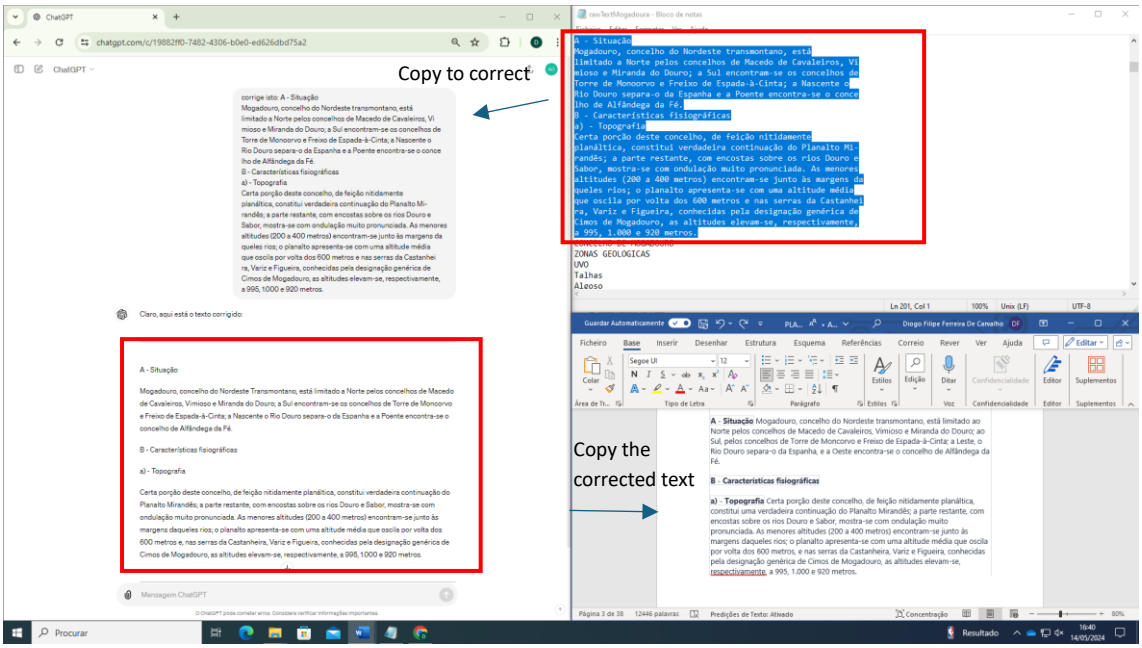


Part V - AI-based Text Correction

6.1. Log in to your OpenAI account (ChatGPT). If you haven't a OpenAI account, you need to sign up. After creating your account, start correcting your text by writing to ChatGPT, type "Correct this: (insert the text of the txt file)" pay attention to the limit of text, if you ask for too much text the results won't be precise.

For this step you'll need to have the results of OCR

- The output will be 3 different files, for this purpose you'll only use the txt file;
- Create one more file where you'll copy the corrected text.



6.2. Open the TXT file and paste in small sections (due to token limits). Prompt: "Correct this text: [paste text]". Finally, copy the corrected version into a new document (e.g., Word or .txt). Tip: Review the corrected text manually for OCR anomalies like symbols or misread numbers.

EXTRA: Troubleshooting Tips

Scanning & Image Quality

- Scanned text appears blurry or low contrast? Ensure good lighting when scanning and apply the “Auto-color” filter in Adobe Scan. Rescan pages if needed.
- Pages are misaligned or cropped? Use manual corner adjustment in Adobe Scan before saving. Check margins visually.
- PDF is too large to process? Compress the PDF using free online tools or reduce the number of pages per file.

OCR Processing (Amazon Textract)

- Textract not available in your AWS region? Try changing your region (e.g., to "US East (N. Virginia)" or another supported one).
- OCR output is disorganized or text blocks are missing? Try increasing the scan contrast or rescan using higher resolution.
- Text full of strange symbols or broken words? Likely due to poor scan quality, ensure flat page scans and good lighting.

AI Text Correction (ChatGPT/OpenAI)

- Model refuses to process the full document? Split the input into smaller chunks (around 500–1000 words).
- Corrections are inaccurate or misinterpret meaning? Avoid including headers/footers or formatting artifacts. Prompt clearly, e.g., “Correct this paragraph for OCR mistakes.”
- Unwanted rephrasing instead of correction? Use more precise prompts like:
- "Correct OCR errors only. Do not rewrite or paraphrase."

GIS (ArcGIS Pro / QGIS)

- Georeferencing points don’t match correctly? Ensure you are using the same coordinate reference system (CRS) in both the image and the shapefile.
- Image does not appear after loading? Check the layer order. Raster image should be beneath vector layers. Zoom to layer.
- Split tool fails to divide the shapefile? The split line must fully cross the polygon geometry. Incomplete lines won't work.
- Edits are not saving? Make sure you are in editing mode and that the session is saved before exiting.
- Exported shapefile is empty or incomplete? Verify that features were selected before exporting and that the attribute table is correctly filled.

File Management & Output

Lost track of files or versions? Use consistent, descriptive filenames with dates or version numbers (e.g., Map_PT_1890_v2_georef.tif).
Cannot open exported shapefile on another computer? Make sure to include all components (.shp, .dbf, .shx, etc.) when transferring.