AGILE: GIScience Series, 5, 49, 2024. https://doi.org/10.5194/agile-giss-5-49-2024
Proceedings of the 27th AGILE Conference on Geographic Information Science, 4–7 Sept. 2024.
Editors: Alison Heppenstall, Mingshu Wang, Urska Demsar, Rob Lemmens, and Jing Yao.
This contribution underwent peer review based on a full paper submission.

Unveiling historical agroecological patterns through artificial intelligence (AI) and Geographic Information Systems (GIS)

Cláudia M. Viana[1], Diogo Carvalho[1]

[1] Center for Geographical Studies, Institute of Geography and Spatial Planning, University of Lisbon, Lisbon, 1600-276, Portugal

Correspondence: Cláudia M. Viana (claudiaviana@edu.ulisboa.pt)

Abstract. Tracing the evolution of regional agroecological specifics is crucial for comprehending the successes and setbacks of historical human intervention in land use, significantly impacting the current and future suitability of agricultural land. Despite diligent efforts to standardize diverse data sources and quantitatively reconstruct data from various periods, researchers grapple with ongoing questions about the reliability of historical agroecological data. This underscores the imperative for novel methodologies aimed at enhancing the quality and quantity of available historical agroecological data. Recognizing the pivotal role of these historical sources, this paper unveils the preliminary outcomes of the AgroecoDecipher project—dedicated to tracing geographic land patterns through historical agricultural records and artificial intelligence. The initial phase involves gathering and harmonizing data through the digitization, georeferencing, and storage of historical surveys for each of Portuguese municipality (n = 277). Employing an exploratory methodology grounded in artificial intelligence (AI) and Geographic Information Systems (GIS), the projected solutions aim to extract a comprehensive database from textual records and map files, facilitating their accessibility for geospatial analysis. The overarching results have contributed to the development of open science and collaborative solutions, embedded within enduring tools for agroecological analysis. This includes the establishment of routines based on open-source AI tools for optical character recognition (OCR), coupled with the formulation of guidelines for text parsing. These endeavors not only preserve the historical information contained in these sources but also establish an invaluable resource for researchers and future studies.

Keywords. Historical surveys, K-means, Cluster-based analysis, Optical Character Recognition

1 Introduction

Many historical sources inherently contain geographic information, addressing crucial questions about the timing, location, and spatial extent of various phenomena (Goodchild, 2002; Knowles, 2005). Historians and geographers, recognizing the value of each question, employ them to complement and structure new knowledge. Historical data, available in diverse formats such as tables, analog records, and even digital maps, can be transformed into both quantitative and qualitative forms (Gregory and Geddes, 2014; Gregory and Ell, 2007; Knowles, 2008). However, it is essential to delineate the constraints associated with historical sources in different scientific fields, including limited spatial resolution, accuracy issues, and challenges with poorly readable text (Murrieta-Flores et al, 2017; Murrieta-Flores and Martins, 2019; Boivin and Crowther, 2021). Addressing these constraints enhances our ability to evaluate historical events through scientific analysis and offers insights for developing effective solutions to contemporary and future societal challenges. Geographic Information Systems (GIS) technology has played a pivotal role in extracting geospatial data from historical sources (Viana, 2023). By converting past records into normalized and structured detailed data, GIS inspires new scenarios of understanding. This is achieved by exploring spatially the temporal evolution of historical-geographical events through the construction of visual arguments and textual typologies using maps and diagrams, thus aiding the understanding of natural and anthropogenic dynamics. Promoting quantitative and geospatial approaches in areas beyond their usual scientific context facilitates

interdisciplinary and transdisciplinary processes and dynamics. For instance, detailed data on agroecological conditions are vital for harmonizing agriculture with natural processes and correlating the impacts of historical human activities with past environmental changes. Despite ongoing efforts to standardize diverse data sources and quantitatively reconstruct historical agroecological data, challenges persist in ensuring reliability, and precise translation into contemporary agroecological standards remains elusive. Nevertheless, precise reconstructions of historical agricultural systems and ecological processes are crucial for tracing agroecological conditions and trends, playing a vital role in academic discussions and sustainable future planning. Insights into the relationship between agriculture and the environment are particularly essential for evaluating food security in any country. The persistent need for new methodologies underscores the quest to enhance the quality and quantity of available historical agroecological data.

This paper unveils the initial findings of the AgroecoDecipher project, which focuses on tracing geographic land patterns through historical agricultural records and artificial intelligence. The study specifically explores the utilization of transcribed historical surveys as data sources for reconstructing and reinterpreting regional agroecological conditions and trends in Portugal throughout the 20th century. The research pursues three primary objectives: a) Digitalization of Surveys: this involves converting survey information into digital formats, facilitating systematic storage and retrieval of data; b) Optical Character Recognition (OCR): the implementation of OCR aims to recognize and extract text from scanned documents, enabling the conversion of textual data into machine-readable formats; and c) Exploratory K-Means Cluster-Based Analysis: employing exploratory K-means clustering analysis, the study categorizes the textual data of each municipality survey into clusters with similar terms. This approach enhances the understanding of patterns and relationships within each municipality, revealing insights that might not be immediately apparent through traditional data examination methods.

2 Data and Software

2.1 Data

In this study, the primary data source comprises historical surveys from the Agricultural and Forestry Surveys of the Agricultural Promotion Plan (Fig. 1), published at various times throughout the 1950s. These surveys offer comprehensive historical information presented in textual, tabular, and cartographic formats, detailing the edaphoclimatic, agroecological, and physiographic characteristics (e.g., climate, soil, water, land use, road networks) of each municipality (n=277) in mainland Portugal.

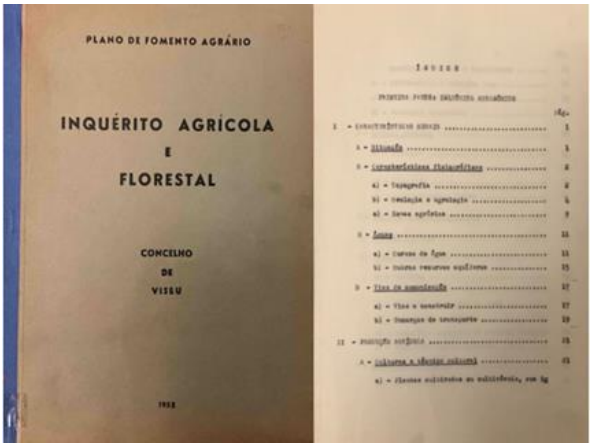


Figure 1. Copy of the Agricultural and forestry Survey – agrarian development plan for the municipality of Viseu (1953).

Figure 1 illustrates a survey, specifically the agrarian development plan for the municipality of Viseu (1953). These surveys meticulously compile information on almost every facet of rural Portugal's economy, providing a detailed account of agricultural and forestry uses along with their suitability. Extracting this detailed information contributes significantly to reconstructing historical agroecological conditions and trends, enabling an understanding of potential complex behaviors and dynamics between different variables that describe agricultural systems in Portugal.

For this exploratory analysis, 10 surveys were considered (Fig. 2).

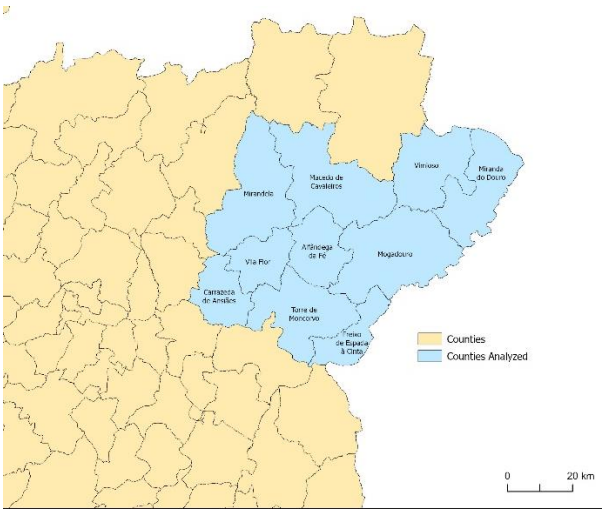
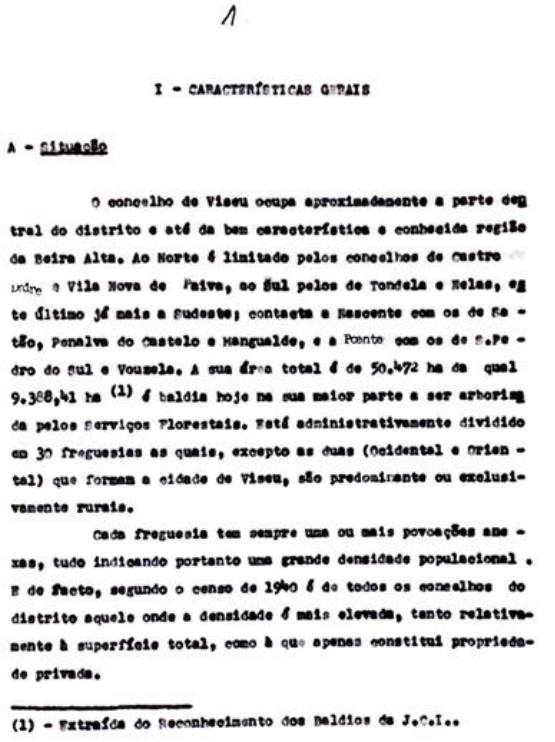


Figure 2. Geographic distribution of surveyed municipalities (n=10)

2.2 Methods

To digitally transcribe historical text records, a methodological approach was devised that seamlessly integrated Artificial Intelligence (AI) to convert written characters into machine-readable form and Geographic Information System (GIS) to scan, georeference, and disseminate the rich historical and geographic information contained within. For the digitization process, the strategic decision was made to leverage Adobe Scan software utilizing an iPad Pro. This deliberate choice underscores the adoption of a specific technological approach, where the utilization of Adobe Scan software, in tandem with the advanced capabilities of an iPad Pro device, played a pivotal role in ensuring the efficiency of the digitalization endeavor. Subsequent to the digitization process, a two-color system filter—comprising black and white—was systematically applied. This approach employed an effective binarization technique to the colors of each survey sheet (Fig. 3). Through this step, the survey sheets were transformed into a binary format, where pixels were represented exclusively in black or white. This process streamlined the data for further analysis, offering enhanced clarity and facilitating the extraction of meaningful insights.



I - CARACTERÍSTICAS GERAIS

A - Situação

O concelho de Viseu ocupa aproximadamente a parte cen
tral do distrito e até da bem característica e conhecida região
da Beira Alta. Ao Norte é limitado pelos concelhos de Castro
Daire e Vila Nova de Paiva, ao Sul pelos de Tondela e Nelas, es
te último já mais a Sudoeste; contacta a Nascente com os de Fa -
lão, Penalva do Castelo e Mangualde, e a Poente com os de S.Pe -
dro do Sul e Vouzela. A sua área total é de 50.472 ha da qual
9.388,41 ha (1) é baldia hoje na sua maior parte a ser arboriza
da pelos Serviços Florestais. Está administrativamente dividido
em 30 freguesias as quais, excepto as duas (Ocidental e Orien -
tal) que formam a cidade de Viseu, são predominante ou exclusi-
vamente rurais.

Cada freguesia tem sempre uma ou mais povoações ane -
xas, tudo indicando portanto uma grande densidade populacional .
E de facto, segundo o censo de 1940 é do todos os concelhos do
distrito aquele onde a densidade é mais elevada, tanto relativa-
mente à superfície total, como à que apenas constitui propriedade
de privada.

(1) - Extraída do Reconhecimento dos Baldios da J.C.I..

Figure 3. Illustration of a digitally transcribed sheet through the utilization of Adobe Scan software and a two-color system filter—employing a black and white scheme

The second phase involved determining the optimal Optical Character Recognition (OCR) software for extracting text from images/PDFs. In this era of emerging Artificial Intelligence (AI), we employed the AI tool 'ChatGPT version 3.5' to inquire about the most effective OCR software for our process. 'ChatGPT' identified the top five contenders as Google's Tesseract, Adobe Reader, Abby FineReader, OneNote, and Amazon Textract. Subsequently, we conducted a comparative assessment by testing all five software options on a single sheet (Fig. 3) to determine which would produce the most accurate results in text recognition. Accuracy was evaluated by comparing the number of characters present on the sheet with the OCR recognition results—determining the software's capability to accurately recognize each character.

Finally, the third step involves employing K-means cluster-based analysis to group and categorize the textual data from each municipality survey into clusters with similar characteristics, thereby fostering a more nuanced understanding of patterns and relationships within each municipality. This technique was executed in a Python environment using the 'scikit-learn' library. To initiate the clustering process, we began by removing all stopwords, a list of common words such as determiners or other frequently used words with lesser weight. This removal is pivotal, considering that a higher number of words would adversely impact clustering performance (Abualigah & Khader, 2017). The list of stopwords was sourced from 'NLTK,' a Python Natural Language Processing library. Following stopwords removal, we computed Term Frequency-Inverse Document Frequency (TFIDF). TFIDF is a widely employed metric for determining the weight of each term in a document, calculated based on the term's frequency in that document (Yao & Cong, 2012; Abualigah & Khader, 2017). The optimal number of clusters for k-means was determined using the elbow method. This involved creating a graph with outputs within the specified range of different 'k' values or centroids.

3 Results and Discussion

3.1 Digitalization and OCR

In the initial phase of our study, we successfully digitized 10 surveys, with a total sheet count surpassing 500.

Moving on to the second stage of our study, we evaluated five OCR software options. Table 1 provides an accuracy assessment for the digitized sheets. The findings revealed that the software yielding the highest accuracy was Amazon Textract, achieving an accuracy rate of 72.3%."

Table 1. OCR software accuracy assessment

Software	Accuracy
Google's Tesseract	50%
Adobe Reader	12.4%
Abby FineReader	51.5%
OneNote	16.3%
Amazon Textract	72.3%

3.2 Cluster-based analysis

Fig. 4 displays the elbow graph generated from K-means clustering. The optimal value for 'k' was determined as 4, corresponding to the point of intersection on the graph line (Cui, 2020). Consequently, the results indicated the presence of four clusters with similar characteristics

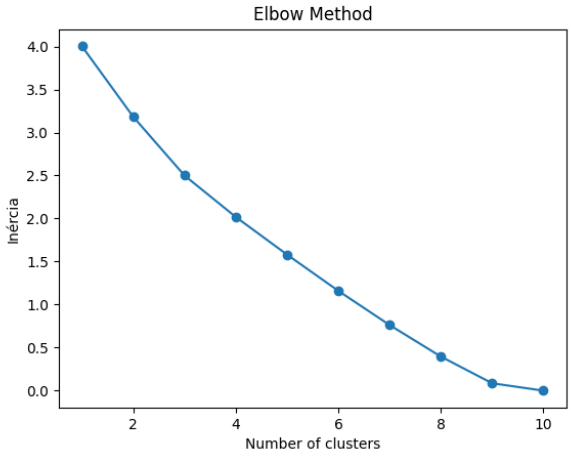


Figure 4. Elbow graph

Fig 5. unveils distinct spatial patterns, showcasing the presence of four clusters and providing valuable insights into the spatial organization and relationships within each municipality survey.

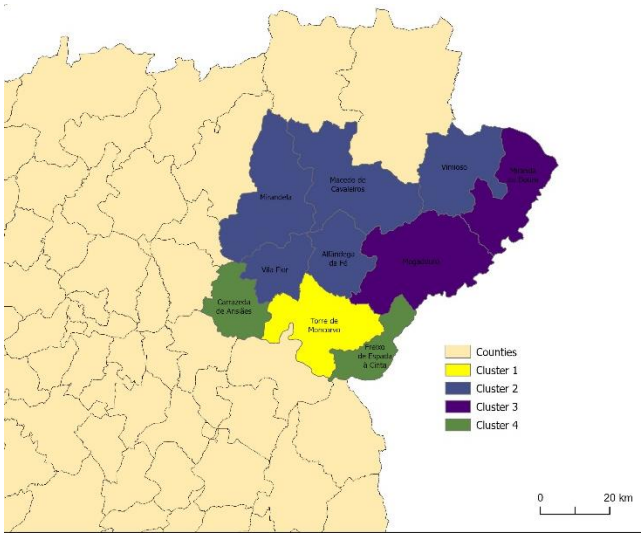


Figure 5. Cluters-based analysis spatial patterns

When analyzing the most frequently used terms in each cluster, the exploratory results reveal distinctive agricultural themes. In Cluster 1, the predominant term was “wine”, while in Cluster 2 it was “rye”. Cluster 3 prominently featured the term “Wine”, and Cluster 4 showcased “maize” and “rye”. In summary, the two-phase approach employed in this study enhances the usability of the survey data, rendering it conducive to advanced analytical techniques for deriving meaningful insights and patterns.

4 Conclusions

The utilization of historical sources for detecting spatiotemporal patterns and trends presents challenges in terms of transcription and transformation into normalized, detailed, and structured data. Consequently, it becomes imperative to explore approaches facilitating the normalization of diverse information sources and the quantitative reconstruction of data from distinct periods. The exploratory clustering analysis played a pivotal role in unveiling nuanced insights and identifying underlying structures within the textual data derived from surveys. In essence, the exploratory analysis presented demonstrates efficiency and simplicity, establishing itself as a promising and effective approach.

Funding

This research was funded by the FCT - Portuguese Foundation for Science and Technology [2022.09372.PTDC] and Center for Geographical Studies [LA/P/0092/2020, CUIDB/00295/2020, UIDP/00295/2020].

Acknowledgments

We acknowledge the GEOMODLAB (Laboratory for Remote Sensing, Geographical Analysis and Modelling) of the Centre of Geographical Studies/ Institute of Geography and Spatial Planning for providing the computational infrastructure support for this research.

References

Abualigah, L. M., & Khader, A. T. Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering. The Journal of Supercomputing, 73, 4773-4795. https://doi.org/10.1007/s11227-017-2046-2, 2017.

Boivin, N.; Crowther, A. Mobilizing the Past to Shape a Better Anthropocene. Nat. Ecol. Evol., 1–12. https://doi.org/10.1038/s41559-020-01361-4, 2021.

Cui, M. Introduction to the k-means clustering algorithm based on the elbow method. Accounting, Auditing and Finance, 1(1), 5-8. doi: 10.23977/accaf.2020.010102, 2020.

Goodchild, M. F. Combining Space and Time: New Potential for Temporal GIS. In A. K. Knowles (Ed.), Past time, past place: GIS for history (pp. 179–197). 2022.

Gregory, I.; Geddes, A. Towards Spatial Humanities: Historical GIS and Spatial History; Indiana University Press, Bloomington. https://www.jstor.org/stable/j.ctt16gz7s5, 2014.

Gregory, I.; Ell, P. Historical GIS Technologies, Methodologies and Scholarship. Archaeology, No. CUP. https://doi.org/10.1017/cbo9780511493645, 2007.

Knowles, A. K. Emerging Trends in Historical GIS. Hist. Geogr., 33, 7–13. 2005.

Knowles, A. K. GIS and History. Placing History. How Maps, Spatial Data, and GIS are changing Historical Scholarship, pp 1–24. 2008.

Murrieta-Flores, P., Donaldson, C., & Gregory, I. GIS and Literary History: Advancing Digital Humanities research through the Spatial Analysis of historical travel. DHQ: Digital Humanities Quarterly, 11(1), 1–18. http://www.digitalhumanities.org/dhq/vol/11/1/000283/000283.html, 2017.

Murrieta-Flores, P.; Martins, B. The Geospatial Humanities: Past, Present and Future. International Journal of Geographical Information Science. Taylor and Francis Ltd. December 2, pp 2424–2429. https://doi.org/10.1080/13658816.2019.1645336, 2019.

Viana, C. M. Reflexão sobre as abordagens geoespaciais da investigação geográfica a aplicadas à MODELAÇÃO DOS SISTEMAS AGRÍCOLAS. Finisterra, 58(124), 181–196. https://doi.org/10.18055/Finis33462, 2023.

Yao, M., Pi, D., & Cong, X. Chinese text clustering algorithm-based k-means. Physics Procedia, 33, 301-307. doi: 10.1016/j.phpro.2012.05.066, 2012