ML-Driven Insights into the Geosocial Dynamics of Norman Sicily

Nischith Srikanth
MehtA+

Vivian Tang
MehtA+

Carmen Wang
MehtA+

Selina Zhang
MehtA+

July 31, 2025

Abstract

In this 3-part research project, we will analyze elevation patterns, analyze settlement patterns, and construct a sustainable AI chatbot to facilitate interactive learning of Sicilian history. For the first part, we used a combination of statistical and machine learning techniques to identify patterns and correlations between elevation and other characteristics of the sites. Additionally, we employed unsupervised machine learning methods such as k-means clustering, which allows spatial visualization of these historical sites that can reveal geographic groupings and potential settlement trends. Moreover, to help users track and find information more efficiently, we designed and developed a chatbot powered by large language models (LLM), specifically GPT-4o from OpenAI. To ensure factual accuracy and minimize hallucination and inferences from the chatbot, we integrated Retrieval-Augmented Generation (RAG), allowing the chatbot to retrieve relevant information directly from a curated dataset. The resulting system offers a user-friendly interface hosted on Streamlit. Ultimately, this research bridges the disciplines of history and data science, offering an effective and interactive solution to historical researchers and educators interested in both geographic and cultural insights.

1 Introduction

Norman Sicily refers to the period (circa 1061–1194) when the island of Sicily and parts of southern Italy were ruled by the Normans, a people of Viking descent who had settled in Normandy (modern-day France) and later expanded into southern Europe. Furthermore, the Norman Kingdom was created on Christmas Day, 1130, by Roger II of Sicily, with the agreement of Pope Innocent II. This era marked a transformative period in the island's political, cultural, and architectural development. The Norman Sicily project, under the lead of Dawn Marie Hayes, seeks to compile and preserve information related to this significant historical period. Professor Hayes's team aims to document the cultural heritage of Sicily from c. 1061-1194 by using print, photographic, web, and geolocation technologies to identify and explicate dilapidated, at-risk, and/or hard-to-access monuments [1]. By offering digital access to the cultural heritage, Professor Hayes hopes to foster interest in preserving these ancient sites and monuments.

This research builds upon the foundational data collected by Professor Hayes's team and aims to conduct an in-depth analysis to identify any patterns that may give the researchers some interesting insights into Norman Sicily's history. While previous work on this project focused on collecting datasets, our research targets analysis and interpretation from a broader perspective. We believe that machine learning is capable of finding patterns and correlations within the extensive datasets efficiently, revealing historical trends that might otherwise go unnoticed. Using both traditional statistical and machine learning methods, we performed two key case studies analyzing elevation patterns and settlement patterns among the monasteries and fortifications of Norman Sicily. Through our case studies, we were able to identify and visualize different patterns between monastery and fortification sites and use our correlational results to hypothesize insights into Norman Sicily's geosocial dynamics.

As the dataset continues to grow, however, navigating and retrieving specific data to produce such insights becomes increasingly difficult. To address this, we developed an AI chatbot that can assist scholars in documenting patterns and relationships when fed data. This could significantly improve their workflow and enable users to interact with databases more productively. Additionally, our chatbot's use of LLMs to understand, retrieve, and interpret information allows

researchers to focus on data collection over organization. Ultimately, this research bridges the disciplines of history and data science, offering a user-friendly solution that enhances both historical inquiry and data assessments.

2 Related Work

In general, there is a notable lack of prior research specifically focused on Norman Sicily. While much research has been done on medieval Italian history, we found little scholarly work that centers on the monasteries and fortifications from the Norman period. A key challenge to this research gap is the inaccessibility of many monasteries and fortifications, which limits their study. One of the few ongoing projects dedicated to this subject is led by Professor Dawn Marie Hayes and her team [2], who have focused on field investigations, cataloging historical data, and processing images at various Sicilian historical sites [2]. While their work has contributed significantly to data collection, it has not yet employed advanced data analysis techniques to uncover broader patterns, such as spatial distribution, elevation, or clustering of the sites. Our research fills this gap by building on their foundation, applying statistical analysis and machine learning methods to give their team a broader view of the historical sites and provide analytical insights to the Norman Sicily society.

In the chatbot aspect, large language models are proven to be highly effective in generating natural, human-like responses. Large language models (LLMs) are pretrained models on massive datasets, such as OpenAI's ChatGPT, capable of instantly generating highly realistic and convincing conversational responses [3]. However, the biggest flaw of these LLMs is that they tend to make hallucinations very often and may generate misinformation when lacking relevant knowledge [4]. This issue is particularly problematic in fields like education and medicine, where users want the LLM to only retrieve information based on the data they input, without making any inferences [5, 6]. Consequently, other researchers have introduced retrieval augmented generation (RAG) to their projects. In short, RAG uses the input sequence x to retrieve text documents z and uses them as additional context when generating the target sequence y. [7] There has been some prior research that explores RAG chatbot's applications and has shown success in reducing hallucination and inference errors [5, 6]. This matches with the goal of our research, which is to help other researchers locate their data more efficiently. Therefore, developing a chatbot using RAG and LLM ensures that the bot does not make any inferences or output misinformation. Ultimately, it gives Professor Hayes's team a reliable and efficient solution to navigate and organize the data they've collected. We believe this could greatly improve their work flow and provide valuable educational resources to other researchers in the related field.

3 Methodology

3.1 Dataset

All the data used in this research, from finding patterns to training LLM, is provided by our collaborating professors from Montclair State University. They include datasets on monasteries, fortifications, churches, bridges, and people and places. To identify the patterns in elevation and settlements, we primarily focused on the "Sites of Norman Sicily" dataset and specifically the monasteries and fortification subdatasets. These datasets contain around 200 identified monasteries and 150 fortifications, and each of them has a unique ID. Within each dataset, there is a wide range of information, such as locations, names, elevations, seismic classifications, status, and many more. Besides that, we also used "the CSSI Surveys" dataset, which quantifies various forms of environmental damage that exist at these sites. As noted by our professor, these surveys were just conducted in May 2025, which shows the high relevance and accuracy of the dataset. We also incorporated "People to Places" and "Places to Places" datasets, especially for our chatbot: the former links historical individuals to specific sites, and the latter maps the relationships between two sites (for example, indicating dependencies such as "Monastery A controls Church B"). Additionally, we used the "Norman Conquest" dataset, which provides valuable information about the wars and events related to Muslim towns, making our story fuller.

3.2 Preprocessing

To begin preprocessing the datasets for the first part of the research, which is finding settlement and elevation patterns, we start by cleaning the data and transforming it into something that the computer can understand. The first step is to remove the unnecessary columns that do not contribute to our analysis. This includes columns like "Site Spelling Variations" and "Location Spelling Variations", which are text-based notes that are not required for the quantitative modeling.

Next, for categorical data like "Modern Comune", "Modern Province", and "Probable Status", we convert them into one-hot encoded vectors. Each row gets a 1 in the column corresponding to its category and 0s in the other columns.

This allows the model to understand the categorical data numerically without any ordering or ranking between the categories. As shown in Table 1, the one-hot encoded values represent different categories.

Table 1: One-Hot Encoding for Province Data

Province_Agrigento	Province_Caltanissetta	Province_Catania
0	1	0
0	0	0
1	0	0
0	0	1

Finally, we need to normalize the numerical data, such as "Elevation", "Seismic Classification", and "Date Range of Attestation" in the dataset. Different features in a dataset can have significantly different ranges, units, and scales. For example, seismic classification ranges from 0 to 5, while elevation might range from 0 to 5000. If these features are not normalized, features with larger numerical values or wider ranges tend to dominate the model's calculation, leading to biased results. Therefore, we perform standardization, which scales the data so it has a mean of 0 and a standard deviation of 1. The formula for normalization is given by:

$$x_{\text{normalized}} = \frac{x - \mu}{\sigma}$$

where x is the original value, μ is the mean of the feature, and σ is the standard deviation of the feature.

3.3 Elevation Case Study

To investigate elevation patterns across monasteries and fortifications, we implemented a combination of statistical and machine learning techniques depending on our feature types and distributions. For comparisons between elevation, a continuous variable, and another continuous variable such as latitude or longitude, we used Pearson Correlation Coefficients, which measure the strength and direction of the variables' linear relationships on a scale from -1 to 1. The formula for the Pearson correlation is given by:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

In other cases involving a dichotomous feature, such as probable gender (coded as male or female), we used point-biserial correlation, a specialized form of Pearson's Correlation meant to measure associations between a continuous variable and a dichotomous variable. The formula for point-biserial correlation is:

$$r_{pb} = \frac{M_1 - M_2}{s_p} \cdot \sqrt{\frac{n_1 n_2}{n(n-1)}}$$

where M_1 and M_2 are the means of the two groups, s_p is the pooled standard deviation, and n_1, n_2, and n are the sample sizes.

Oftentimes, however, the process wasn't as simple. Our data consisted mainly of categorical features, such as monastic identity, modern province, and historical region. Since traditional correlation measurement methods do not apply to categorical variables, we first transformed them into one-hot encoded vectors and then performed one-way ANOVA (analysis of variance) tests.

ANOVA assesses whether the between-group variance is significantly greater than the within-group variance, allowing us to compare the mean elevation among each category and determine whether there is a statistically significant difference in elevation among them. The formula for ANOVA is:

$$F = \frac{\text{Variance Between Groups}}{\text{Variance Within Groups}} = \frac{\frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2}{\frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}$$

where k is the number of groups, n_i is the number of observations in group i, $\bar{y}_i$ is the mean of group i, and $\bar{y}$ is the overall mean.

To visualize these differences, we constructed side-by-side boxplots, allowing us to compare the differences in elevation distributions among different categories.

We also wanted to identify, among the given monastery and fortification features, which ones were most correlated with elevation. To explore this, we constructed a multivariate linear regression model with Lasso regularization (Least Absolute Shrinkage and Selection Operator). Lasso regressions apply a penalty on large regression coefficients, causing some coefficients to shrink to zero, which effectively performs both variable selection and regularization to prevent overfitting. The formula for Lasso regression is:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} (y_i - \mathbf{x}_i \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

By examining the nonzero coefficients retained by the Lasso model, we identified the features most predictive of elevation.

Similarly, we also explored using SVM (Support Vector Machine) with feature importance to predict elevation. We used a Radial Basis Function (RBF) kernel, which captures non-linear relationships between the features and the target variable, the elevation. By evaluating the feature importance using the permutation importance method we imported from sklearn, we identified the top impactful features on the model prediction, and whether they contributed positively or negatively. The formula for permutation importance is:

$$\text{Importance}_j = \frac{1}{n_{perm}} \sum_{i=1}^{n_{perm}} \left(\text{Score}(X) - \text{Score}(X_j^{perm}) \right)$$

where X_j^{perm} represents the permuted values of feature j, and n_{perm} is the number of permutations.

Finally, we compare the feature importance results from both the Lasso regression and SVM models, and features that appear as important in both models can be considered highly predictive of elevation.

3.4 K-means clustering

To explore settlement patterns, we employed K-means clustering to plot all the monasteries and fortifications on a map, which enabled us to analyze any potential settlement patterns. K-means clustering is a type of unsupervised learning and uses vector quantization. It aims to partition n observations to k clusters in which each observation belongs to the cluster with the nearest mean, which minimizes the within-cluster variance. Formally, the objective is to find:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \mu_i\|^2 = \arg\min_{S} \sum_{i=1}^{k} |S_i| \operatorname{Var}(S_i)$$

This approach allows us to assess if any part of the island exhibits a higher concentration of monasteries and fortifications while also providing a broader overview of their distribution. Furthermore, we will divide the clusters into different colors to investigate any specific trend or pattern in the monasteries and fortifications.

3.5 LLMs

A common challenge with existing large language models, such as ChatGPT-4o or Gemini, is their inability to answer questions about information that was not included in the training data. When prompted with queries relating to specific internal data that a university or company may have, these models often fail to provide accurate responses. To address this limitation, we incorporate Retrieval-Augmented Generation (RAG) and query tools. These tools enable the model to retrieve relevant information from external or internal databases in real time, enhancing its ability to answer questions about data that was not part of its original training set.

The chatbot is implemented using the LangChain framework and integrates Azure OpenAI's GPT-4o model via tool-calling. To reduce hallucinations and promote reproducible outputs, the model temperature is fixed at zero. Upon receiving a specific user question, the chatbot invokes a structured run_query tool. The model is explicitly prompted to infer which dataset best answers the question, generate the required query, and execute the tool only when the input

appears to map to structured data. If the input is a general inquiry or outside the scope of the tabular data, the chatbot defers to an external retrieval-augmented generation (RAG) tool.

To ensure precision and maintain a tight feedback loop between generation and execution, all interactions are mediated through a LangChain agent constructed using AgentExecutor and a custom schema-aware system prompt. This prompt outlines the available datasets and their columns, allowing the LLM to reason about table structure even when user phrasing is ambiguous.

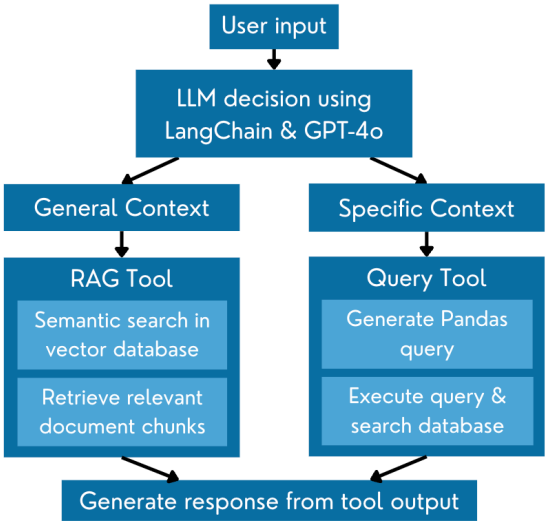

Figure 1: Flowchart of the Chatbot

3.5.1 Query Tool

To support deeper exploration of relationships among historical entities, we developed a natural language chatbot that can be incrementally fed new data and respond to user queries with context-aware, data-driven explanations. The chatbot translates natural language questions into executable pandas queries over two core datasets, `people_to_places_df` and `places_to_places_df`, which were curated to represent connections between individuals, sites, and their affiliations in Norman Sicily.

The datasets initially contained redundant or inconsistent attributes across related tables, including overlapping records on religious structures and geographic metadata. These were preprocessed before being merged with domain-specific data on monasteries, churches, fortifications, etc. This cleaning phase also included schema harmonization to enable reliable joins, reduce ambiguity in query targets, and support inference at multiple levels. Finally, all columns were renamed to a detailed description that an LLM can interpret.

Once preproccessing was complete, we wrote a Python function that requires two arguments: the target dataframe and the corresponding Pandas query string. The LLM can call and run the function, receiving a result that is later refined and given to the user.

3.5.2 Retrieval Augmented Generation (RAG)

Before implementing RAG, we need to convert the CSV dataset into readable documents that can be processed. Each row in the dataset corresponds to a unique document (e.g. information about a specific monastery or a historical figure). These documents were then structured into paragraphs, where each sentence contained relevant details that were derived from the datasets' columns. For example: "This monastery is named X. Its patron saint is Y, and it was founded in XXXX." By matching the site ID or person ID, we combined the relevant data into a comprehensive paragraph for each entity. Later, we used the Recursive Character Text Splitter to break the documents into smaller chunks (500 characters each with no overlap). Each chunk is then converted into a vector representation using HuggingFace embeddings, which is a pre-trained model that captures semantic meaning. After generating embeddings, we stored the vectors in a vector store (Chroma). When a user submits a query, the RAG pipeline searches the vector store for the most similar document chunks based on the query. We used similarity search to retrieve the top k most relevant chunks. These chunks are then combined and passed to the language model to generate a coherent and contextually accurate response.

4 Results

4.1 Elevation Patterns

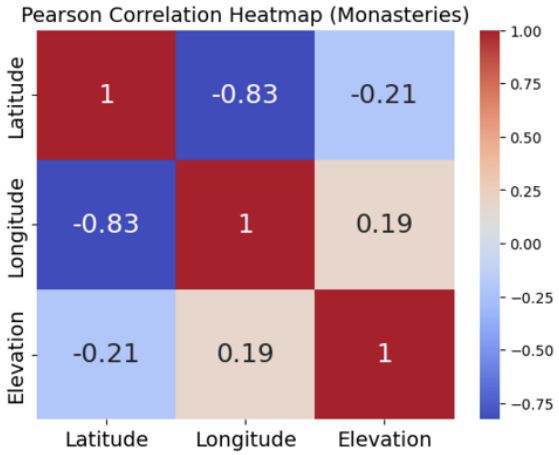

Figure 2: Heatmap showing the correlation between monastery elevation, latitude, and longitude.

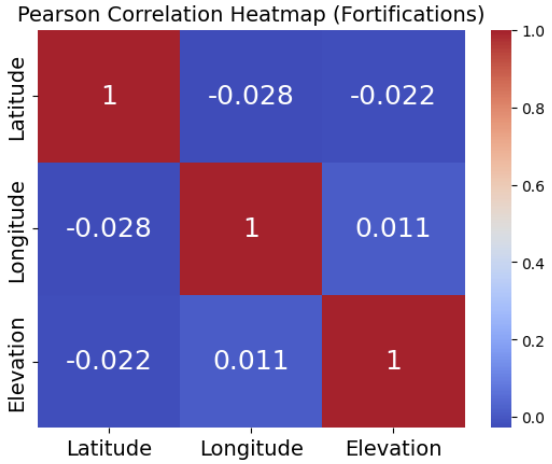

Figure 3: Heatmap showing the correlation between fortification elevation, latitude, and longitude.

As found in the correlation heatmaps above (Figures 2 and 3), monastery elevation is weakly positively correlated with longitude and weakly negatively correlated with latitude, meaning that monastery elevation increases northward and eastward of Sicily. This aligns with the topography of Sicily, where the northeastern coast is more mountainous (e.g. Madonie mountains, Nebrodi mountains). Notably, there is a strong correlation between latitude and longitude among monasteries, suggesting a particular pattern within their settlements that could be due to geographical features such as coastlines or mountain ranges.

In contrast, fortification settlements show little to no correlation among elevation, latitude, and longitude, indicating that they were more evenly distributed across Norman Sicily, regardless of terrain. This makes sense because fortifications were not always focused on elevation but were meant to provide security and defense across a broad range of locations.

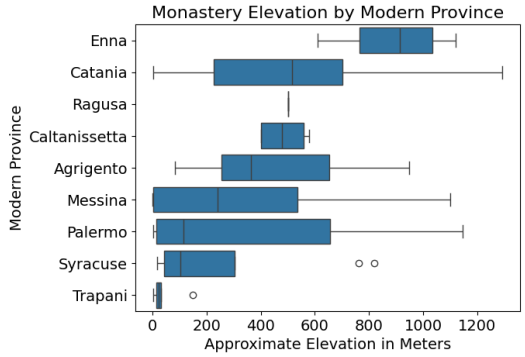

Figure 4: Boxplot of monastery elevations by province.

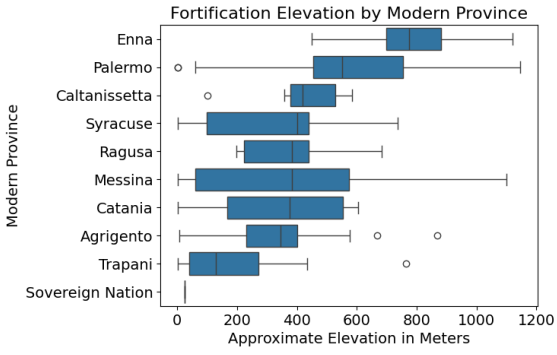

Figure 5: Boxplot of fortification elevations by province.

Figure 4 and 5 show two side-by-side boxplots for monasteries and fortifications' elevations in different modern provinces. We found that monasteries and fortifications within the same province are located at similar elevations. The highest one is Enna, situated at an altitude of approximately 3054 feet above sea level, being the highest provincial capital in Italy. Its elevated terrain made it a formidable province, often referred to as an "urbs inexpugnabilis" (impregnable city). The lowest one we found is Trapani, which is located at an elevation of 302 feet, with some areas even at sea level.

Table 2: Shared Top Features with Lasso Coefficients and SVM Importances for Monasteries)

Feature	Lasso Coefficient	SVM Importance
Latitude	-0.699606	0.013909
Province_Enna	0.374580	0.031032
Longitude	-0.307138	0.232707
Region_Val Demone	0.280775	-0.118092
Seismic2019	-0.068102	-0.384349
Province_Trapani	-0.240604	-0.227713
Seismic2024	-0.115948	-0.377781
Monastic Identity_Basilians	0.013427	-0.139996

Table 3: Shared Top Features with Lasso Coefficients and SVM Importances for Fortifications)

Feature	Lasso Coefficient	SVM Importance
Province_Enna	0.331901	0.077337
Seismic2024	0.213923	-0.399818
Province_Trapani	-0.281880	-0.251344
Latitude	-0.281654	-0.407583
Province_Syracuse	-0.248947	0.153485
Province_Palermo	0.197668	-0.178530
Longitude	-0.196591	-0.251344
Region_Val Demone	0.170824	-0.015396
Province_Catania	-0.168562	0.243675
Province_Ragusa	-0.148607	-0.025513

In Table 2, the top features Latitude, Province_Enna, and Longitude match our former findings. Another notable feature, Region_Val Demone, shows a moderately positive coefficient, and since this region encompasses areas such as the Nebrodi mountain range, its positive association with higher elevation is justified.

Fortification results are shown in Table 3, with the top feature again being Province_Enna, suggesting a similar conclusion that fortifications in this province are located at higher elevations. Interestingly, the feature Seismic2024 has a coefficient of 0.2139, indicating that fortifications in regions with higher seismic activity (such as the western parts of Sicily) tend to be located at slightly higher elevations. The Seismic2019 feature, however, has a less significant coefficient of 0.0232, which may imply that more recent seismic events did not play a major role in determining the location of fortifications compared to other geographic and defense considerations.

Both the monastery and fortification datasets show that location-based features, such as province and historical region, are primary determinants of elevation. Our resulting coefficient magnitudes and feature importance suggest that monasteries tend to be situated in higher elevation areas compared to fortifications, which could also be attributed to religious motivations of the Normans, such as a desire to be closer to the divine.

4.2 Settlement Patterns

Before applying k-means clustering, it is essential to first determine the optimal value for k. To accomplish this, one of the most widely used methods is the Elbow method, a graphical approach relying on the idea that as you increase the number of clusters, the sum of squared distances between points and their cluster centers (WCSS) will continue to decrease. The goal is to identify the value of k where this reduction starts to level off–indicating a good balance between model accuracy and complexity while avoiding overfitting. Below are the two Elbow curves for monasteries 6 and fortifications 7:

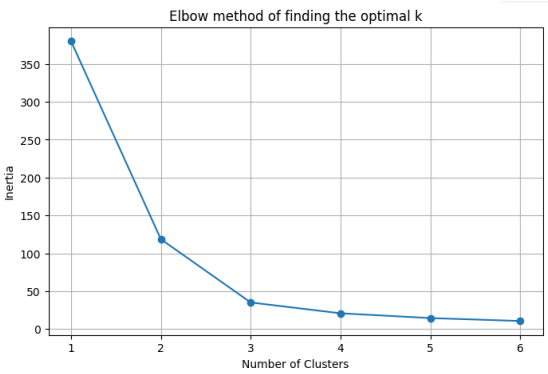

Figure 6: Elbow for monasteries

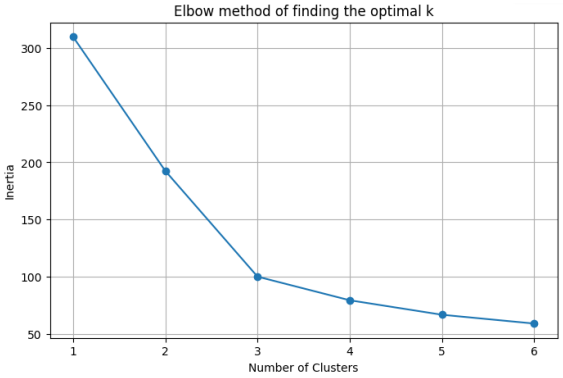

Figure 7: Elbow graph for fortifications

From these two graphs, it is clear that k=3 will be our most optimal number of clusters.

Using the k value we found, we plotted all the monasteries and fortifications by their coordinates on a map centered around Italy using the Folium library. The Folium library is a powerful Python library that helps users create different types of Leaflet maps, which allows us to visualize our clusters on a map. We plotted two kinds of visualization: one categorized the sites by type (monastery and fortification); the other used the computed k value to display clusters, with each cluster represented by a distinct color. It allows us to visualize any potential relationship between monasteries and fortifications and whether there are any clusters in general. Below are the results classified by the type of site: Red represents fortifications, and blue represents monasteries.

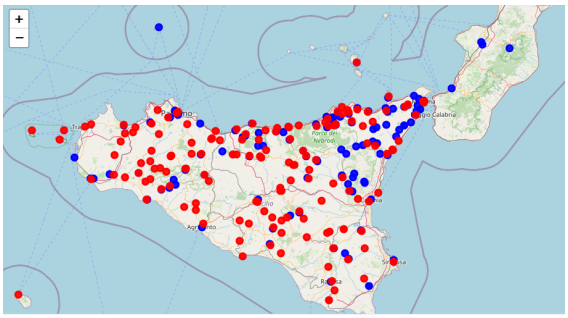

Figure 8: Monasteries k-mean clustering 1

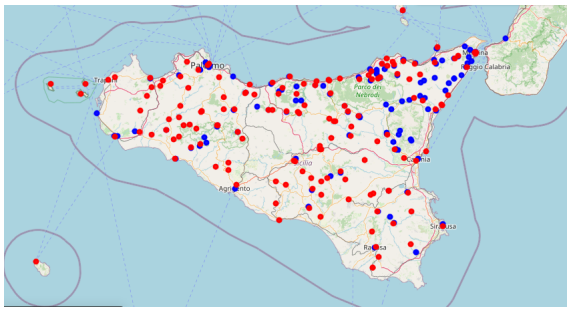

Figure 9: Monasteries k-means clustering 2

From the graphs above, we observed that many monasteries often follow another fortification nearby, suggesting potential spatial relationships between the two. However, this does not apply to all the sites, as there are, in general, more fortifications than monasteries. Moreover, most of the monasteries are located along the coastal areas, especially along the northeast coastline. On the other hand, fortifications are more evenly spread out across the island. Additionally, fortifications are primarily located near border lines (grey lines on the map). This distribution aligns with their strategic purpose, as fortifications are typically constructed to defend territorial boundaries and deter potential intrusions.

Below is the second graph classified into 3 clusters (k=3) with different colors (Figure 10 and Figure 11). This graph could be helpful to see if the monasteries and fortifications are clustered in a particular place on the island.

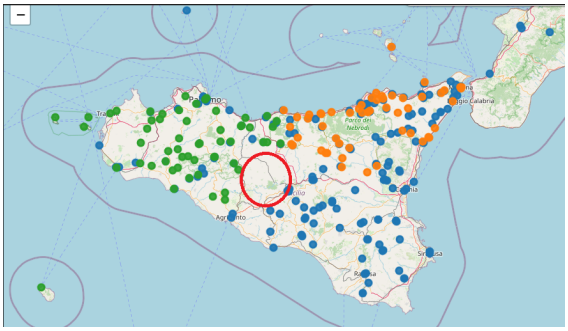

Figure 10: Fortifications k-mean clustering 1

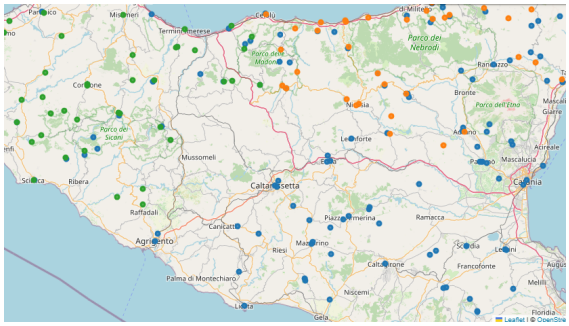

Figure 11: Fortifications k-means clustering 2

In the more zoomed-in view shown in Figure 11, the map's visual elements are color-coded as follows: the gray lines indicate borders between provinces and territories; the blue lines indicate rivers and lakes; the red lines indicate major highways or roads; the orange lines indicate secondary roads. From this graph, we identified a gap in the inland region of the island where no monasteries or fortifications were found, circled in red. In Figure 11, we noticed that most monasteries and fortifications are located near rivers, forests, and roads. Admittedly, the area circled in red does not have any signs of major greenland, rivers, or roads. Given that this research focuses on Norman Sicily during the period 1061–1194, the primary sources of transportation are maritime and overland routes. This aligns with our observation that many sites are located along the coast or near roads and rivers. Consequently, we hypothesize that the central inland region may have been relatively inaccessible to the people of Norman Sicily. Furthermore, we speculate that this finding might suggest that ancient Normans relied on access to rivers and forests for food supply and agriculture. Therefore, communities may have preferred to settle around areas with more accessible resources.

4.3 Chatbot

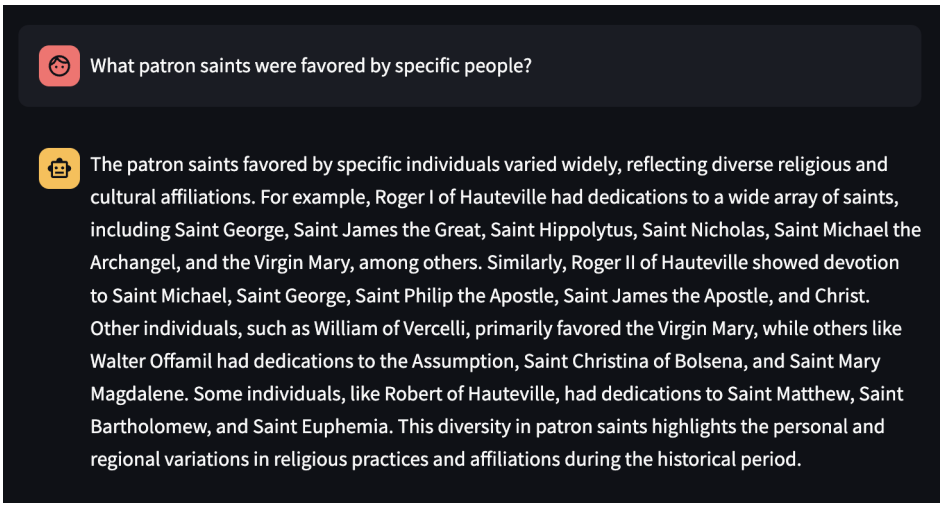

Figure 12: Chatbot Output

4.3.1 Streamlit Interface

To facilitate user interaction with the chatbot, we deployed a web-based interface using Streamlit, a lightweight and flexible Python framework that allows for the seamless creation of interactive interfaces. Streamlit enabled rapid development of a clean, responsive UI without the need for complex front-end development. Users can engage with the chatbot by asking captivating questions, and the interface dynamically displays both user input as well as AI-generated responses in a conversational format. Importantly, the chatbot also preserves history, allowing the model to incorporate context from previous turns, which is crucial for replicating human interaction and meaningful dialogue in a research setting. For instance, when asking the knowledgeable chatbot when Roger I of Hauteville's birthdate is and following

with where he was born, the chatbot can infer the subject of the subsequent question in order to intelligently answer with an accurate understanding.

4.3.2 Chatbot Edge Cases and Responses

The chatbot is designed to handle a variety of user inputs, but there are specific edge cases where its responses are pre-programmed to ensure it's on topic and ethnical correct. Below are some of the key edge cases along with the chatbot's response:

- **Non-Norman Sicily Topics:** When the user asks a question unrelated to Norman Sicily or the datasets provided, the chatbot will respond with:

  "I'm sorry, I am only knowledgeable about the historical and geographical topics related to Norman Sicily. Please ask questions specific to this context."
- **Unknown Information:** If the chatbot encounters a question it cannot answer, it will respond with a polite suggestion to consult other sources:

  "I do not have the information you're asking for. You might want to search online, consult professionals, or refer to relevant books for more accurate details."
- **Sensitive Topics:** For sensitive or inappropriate topics, the chatbot will decline to respond, maintaining ethical boundaries:

  "I cannot engage in discussions about this topic. Please feel free to ask about other historical aspects of Norman Sicily."
- **Multiple Topics in One Query:** If a user asks about multiple topics in a single query, the chatbot will attempt to provide a relevant answer but may ask the user to specify one topic at a time for more clarity:

  "It seems you're asking about multiple topics. Could you please ask about one subject at a time so I can assist you more effectively?"

These edge cases ensure the chatbot maintains a focused and ethical interaction with users, while also encouraging users to seek professional expertise for information outside its scope.

4.3.3 Evaluation Matrix

The evaluation of the chatbot's performance is an essential next step that has not yet been addressed. One promising method for this is the *model-graded evaluation method* outlined in the OpenAI Evals GitHub repository. This method allows for assessing the chatbot's accuracy and responsiveness through a series of queries and model outputs.

For an initial evaluation, a validation set of 10-20 query and answer pairs will be manually created. This set will cover a variety of topics related to Norman Sicily and include common questions as well as edge cases. Running the chatbot against this set will help identify strengths and areas for improvement.

Due to time constraints, implementing this evaluation may not be feasible in the current phase. However, it is a key direction for future development to improve the chatbot's accuracy and usability.

5 Discussions

5.1 Limitations

As with any research project, several challenges and limitations emerged throughout our study. One of the most significant issues is the accuracy and completeness of the data. Sufficient and accurate data is essential for producing valid and reliable results. However, as we investigated the data that we were provided, we found many missing data points across multiple columns in the CSV files. For example, when working with coordinates for the k-means clustering, there are around 20 sites that do not have a coordinate, so we had to delete those rows. This drastically limited the amount of data available to us. Similarly, when finding correlations in elevations, we also encountered many missing data points at different places in the database. To maximize our correlation result, we fill in the mean of the columns for the missing data. Although this method maintains consistency in our dataset, it also increases the potential inaccuracy of our results.

In addition to the problems presented in our datasets, LLM also has many limitations. Though LLM is pre-trained on immense amounts of data, the bot may still present hallucinations that cannot be prevented from using RAG.

Furthermore, our chatbot is designed to operate solely on the datasets we input. However, we cannot ensure that users will always input queries that are related to the datasets in the chatbot. If irrelevant or ambiguous queries are inputted, the chatbot might generate incorrect outputs and even lead to reinforcing the chances of hallucination of the bot. Another practical limitation is the computational costs of running the queries, especially when using advanced models like GPT-4o that run on a GPU. When using it for more complex tasks or with larger datasets, budgetary constraints may hinder the long-term deployment of the chatbot on public platforms without sufficient funding.

Furthermore, the lack of professional knowledge on the archaeology or history of Norman Sicily is a significant limitation. Due to the lack of expertise in the domain in our team, it becomes challenging to assess the accuracy of the chatbot output, which is essential given the black-box nature of the model. Without subject matter experts to validate the AI-generated output, we cannot ensure that chatbot responses are factually correct or aligned with historical realities.

5.2 Future Work

However, limitations lead to the potential for future work. Since our research did not find any strong correlation between elevation and the existing columns we have, future scholars could focus on collecting additional sources of data that may offer a stronger correlation. These could include information related to climate, trade routes, topography, or demographics during the Norman period. Furthermore, in the chatbot aspect, we are currently updating and inputting the data manually, which is both time-consuming and inefficient. Therefore, in the future, we hope to improve the chatbot so it can extract information from the Norman Sicily project's webpage automatically and update the data in its database consistently. This will ensure the chatbot stays up-to-date with minimal human intervention. Moreover, GPT-4o is capable not only of textual analysis but also of image and audio analysis, opening up potential pathways for expansion. Therefore, we hope to incorporate an image analysis and recognition feature into the bot in the future. Additionally, we noticed that there is a lot of photo data stored in the database, which implies that images may be crucial to the exploration of Sicily's history. Thus, such features could provide valuable historical insights and enrich the exploration of Sicily's culture. Ultimately, we could improve the evaluation of the chatbot's performance by incorporating OpenAI evals. We could create a manual validation set with a variety of input-output pairs to assess the chatbot's accuracy, relevance, and overall quality of responses. This evaluation framework would enable us to measure the chatbot's effectiveness and continue refining it systematically.

6 Conclusion

In conclusion, this research paper investigates the spatial patterns in monasteries and fortifications in the Norman Sicily period. While our analysis did not reveal a strong correlation between elevations and other elements of the sites, we did find some notable patterns in the settlements, which may provide valuable insights for future studies of Sicilian history during the Norman era. We believe that many of the limitations and potential inaccuracies in our findings are largely due to the constraints of a relatively small dataset and limited financial resources. Moreover, we have successfully developed a chatbot designed to help researchers retrieve historical data more efficiently. Our approach of using RAG is believed to deal well with hallucination and inaccurate answers. So by incorporating Retrieval-Augmented Generation (RAG) with large language models (LLMs), we minimized the risk of the bot making inferences and ensured the accuracy and reliability of the chatbot's responses. We believe this could help other researchers who are looking to do similar developments to gain some useful insights.

Looking ahead, we aim to make this chatbot publicly accessible, especially by hosting it on the Norman Sicily Project's website, as it can be useful to other researchers and scholars studying in this field. Additionally, there are very few works being done around the Norman Sicily era in the past, so we hope our work will contribute to fostering academic interest in this unique and historically significant period of Italian and European history. Lastly, we are eager to see more work centered around this subject in the future.

7 Division of Labor

We divided the work as follows:

- Data Processing: Selina Zhang, Carmen Wang, Vivian Tang, Nischith Srikanth
- Elevation Pattern Case Study: Vivian Tang, Carmen Wang
- Settlement Pattern Case Study: Selina Zhang
- Chatbot: Nischith Srikanth, Carmen Wang
- Paper: Selina Zhang, Carmen Wang, Vivian Tang, Nischith Srikanth

- Poster: Vivian Tang

8 Acknowledgments

We thank Professor Dawn Marie Hayes and Mr. Joseph Hayes from Montclair State University for trusting us with this project, as well as our mentor, Ms. Haripriya Mehta for all the support along the way.

References

[1] Gary Rodriguez. The norman conquest of southern italy and sicily · the norman sicily project, Sep 2021.

[2] Dawn Marie Hayes. · the norman sicily project, May 2020.

[3] Konstantinos I. Roumeliotis and Nikolaos D. Tselikas. Chatgpt and open-ai models: A preliminary review. *Future Internet*, 15(6), 2023.

[4] Yiqui Shen, Laura Heacock, Jonathan Elias, Keith Hental, Beatrui Reig, George Shih, and Linda Moy. Chatgpt and other large language models are double-edged swords | radiology, Jan 2023.

[5] David Steybe, Philipp Poxleitner, Suad Aljohani, Bente Brokstad Herlofson, Ourania Nicolatou-Galitis, Vinod Patel, Stefano Fedele, Tae-Geon Kwon, Vittorio Fusco, Sarina E.C. Pichardo, Katharina Theresa Obermeier, Sven Otto, Alexander Rau, and Maximilian Frederik Russe. Evaluation of a context-aware chatbot using retrieval-augmented generation for answering clinical questions on medication-related osteonecrosis of the jaw. *Journal of Cranio-Maxillofacial Surgery*, 53(4):355–360, 2025.

[6] Jakub Swacha and Michał Gracel. Retrieval-augmented generation (rag) chatbots for education: A survey of applications. *Applied Sciences*, 15(8), 2025.

[7] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.