Digital Codicology on the Vatican Apostolic Library

Giuliano Giuffrida[0000-0002-8979-4614]

Vatican Apostolic Library, Cortile del Belvedere, Vatican City, Vatican State
g.giuffrida@vatlib.it

Abstract. 1650 Western manuscripts written between the year 800 AD and the year 1500 AD have been subjected to digital codicological analysis. The analysis involved approximately 600,000 high-resolution images (400 dpi), enabling the automatic extraction of 8 codicological parameters: Filling coefficient, percentage of ink on the entire page, percentage of ink in the written area, number of rows, line spacing, text height, page size, and page ratio. This study shows the evolution of these 8 parameters over the centuries, and finally a first neural network capable of automatically dating Western medieval manuscripts has been built.

Keywords: Digital Codicology, Image Analysis, Clustering, Neural Networks.

1 Introduction

Over the past two decades, numerous projects aimed at digitizing book heritage—including manuscripts and printed works—have been launched worldwide. This monumental undertaking has resulted in tens of thousands of digitized manuscripts that are freely accessible, mostly via the IIIF protocol[1]. This, in turn, has revived interest in the field of "quantitative codicology."

The term "quantitative codicology," first coined 40 years ago (see Bozzolo, Ornato 1980, 1982), has been revisited in several notable works, such as Ornato 2020 and Maniaci 2022. While the terms "quantitative" and "codicology" remain debated—failing to fully embrace the objectives and methodologies of this discipline—it is undeniable that one of its central goals is the study of "mise en page." The advent of digitization has significantly facilitated such studies (e.g., De Stefano et al. 2011) by enabling the semi-automatic extraction of key codicological parameters, including:

- The size of the folio: width and height
- The percentage of ink, i.e., the ratio between the number of pixels containing ink and the total number of pixels in the folio

[1] Cfr https://iiif.io.

- The filling coefficient, i.e., the ratio between the area covered by the written surface and the total area of the folio. Basically a number going from 0 (no text) to 1 (entire page covered by text),
- The number of lines of text and the number of columns
- The thickness of text lines and the spacing between various lines of text
- The number of characters present in the folio.

These parameters contribute to reconstructing various "writing recipes", namely, the methods employed by scribes and writing centers to prepare manuscript pages; see for example Cherubini (2004).

It is well-established that these parameters exhibit considerable variation, particularly within the Western tradition. For instance, transitions from early medieval scripts to Gothic writing styles brought about notable changes in nearly all the cited parameters (see, for example, The Oxford Handbook of Latin Palaeography, 2020; and also https://spotlight.vatlib.it/latin-paleography/feature/16-gothic). Consequently, studying the evolution of these parameters over time and space may offer insights into the history of manuscript book production. Furthermore, it opens the possibility of developing semi-automatic systems capable of identifying the origins—both temporal and geographical—of manuscripts or even individual manuscript pages.

This work represents an initial attempt to construct an automatic manuscript dating system. The next sections will detail the dataset used, the codicological parameters extracted, and the methods employed for parameter extraction. This will be followed by data analysis and the presentation of a preliminary neural network designed to estimate the production date of manuscripts—or individual manuscript pages—based on a set of codicological parameters.

2 Dataset

The initial dataset consists of digitized western manuscripts listed in the Vatican Apostolic Library's manuscript catalog. These manuscripts are either dated or have an estimated production date. To ensure consistency in data quality, the dataset was refined to include only digital images acquired at a resolution of 400 dpi, providing uniform digital data for subsequent analyses. Additionally, composite manuscripts[2] and those produced before 800 AD or after 1500 AD were excluded. The manuscripts produced before 800 AD were omitted due to their limited numbers, while the ones produced after 1500 AD were outside the scope of this study.

Following the extraction of codicological parameters, the dataset underwent further selection based on these criteria:

- Manuscripts must contain at least 30 pages with a stable filling coefficient. Pages with significant deviations, such as illuminated or heavily annotated ones, were excluded.

[2] Composite manuscripts have been excluded by evaluating metadata and analyzing the codicological parameters, however, it is possible that the filter may not have worked perfectly.

- Manuscripts with a filling coefficient below 0.2 were omitted, as they typically feature sparse text or are too deteriorated for reliable analysis.
- Manuscripts for which the line detection algorithm failed were excluded, as these often contained minimal or no text.

The reasons for the choice of these criteria will be further explained in the next section, where the techniques for extracting codicological parameters will be described.

After applying these filters, the final dataset consists of 1,650 manuscripts. Fig. 1 presents the distribution of these manuscripts by production year, as well as the uncertainties in their dating. Two important observations emerge:

- The age distribution of manuscripts is highly uneven and asymmetrical, with a significantly smaller number of older manuscripts compared to those written after the 13th century.
- While a few hundred manuscripts are precisely dated (zero uncertainty), the majority have only an estimated production century, resulting in a typical uncertainty of 50 years.

These characteristics significantly influence the subsequent analyses, as will be discussed later.

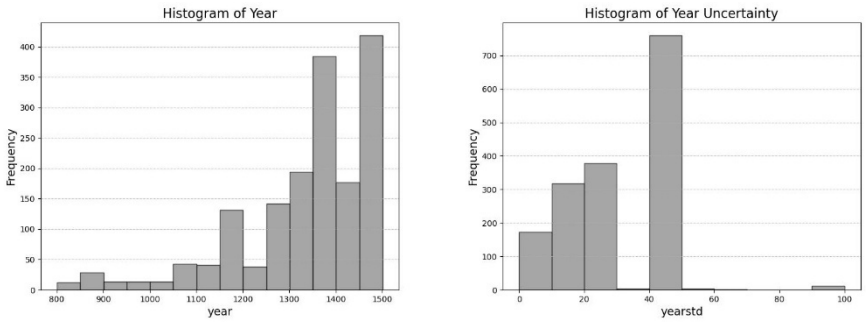

Fig. 1. Distribution of manuscripts' production years and their uncertainty.

3 Parameters Extraction

The dataset described in the previous paragraph counts almost 600000 images distributed among 1650 shelfmarks. A python software has been written to extract the following parameters:

1) The size of the folio: width and height
2) The percentage of ink on the page
3) The percentage of ink inside the written area
4) The filling coefficient
5) The number of lines of text and the number of columns
6) The thickness of text lines and the spacing between various lines of text.

The extraction of the first four parameters has been already described in Giuffrida (2024) and Giuffrida & Manoni (2025). Here the adopted techniques for these parameters will be shortly described, for a full description please refer to that works.

For what concern the size of the folio, a simple edge detection algorithm was typically able to successfully extract these parameters, even if irregularities of the parchment edges can bring some difficulty.

A bit more complex was the detection of the inner edge of the page, i.e. the separation between two facing pages, this measure has been realized using projections, see Fig. 2.

The percentage of ink on the page has been extracted analyzing pixel intensity histogram on which the parchment distribution could be interpolated with a analytical function; in this way the parchment pixel, i.e. the pixel that do not contains ink, can be removed from the histogram and what remains is only the ink. The complete details of this analysis are available on the aforementioned Giuffrida (2024).

The percentage of ink on the written area has been extracted using a common local threshold method: the method used for the full page couldn't be applied in this case due to the low amount of parchment pixels.

The filling coefficient has been evaluated by combining two approaches:

- Morphological operators (binarization, erosion and dilation) have been applied to the images transforming the written area into a feature that can be easily detected by contour algorithms. This algorithm has been developed during the master degree thesis in Physics of Massimiliano Foffi, as a part of a collaboration between the Vatican Library and the Department of Physics of University of Rome Tor Vergata (see Foffi, 2021).
- Analysis of vertical and horizontal projection of the page: through slope variations algorithm it is possible to detect the inner edge of the page and the position of the written area (see Fig.2).

The number of text columns has been evaluated analyzing the vertical projections: the presence of two columns creates a feature, visible on fig.2, that can be easily detected.

The most recent analysis, concerning the extraction of numbers of rows, text width and line spacing, will be described in detail here.

The analysis started from the horizontal projections already used for the filling coefficient evaluation: as it is evident from Fig. 3, since the lines of text produce an almost sinusoidal wave shape, the basic idea is to interpolate that portion of the projection via fast fourier transform, obtaining a continuous function that can be analyzed to extract the needed parameters.

In Fig. 3 the horizontal projection of the written area is shown for one of the image under analysis (folio 19r of Vat.lat.2088). The original projection is plotted in blue, while in green the interpolation via fft is available; the axis y 0 has been set to the mean value of the interpolated fft function, and in red is shown the intercept of the green function with the 0, i.e. the coordinates on which the green function cross the 0 level.

Looking at Fig. 4 it can be easily observed that the red lines delimitate the written text and the space between the lines of text. It is worth noticing, however, that the height of the text lines detected in this way is an average line height, not influenced by isolated character with long or short shafts. The algorithm elaborates the average text height and line spacing for each page, then an average value for the whole manuscript.

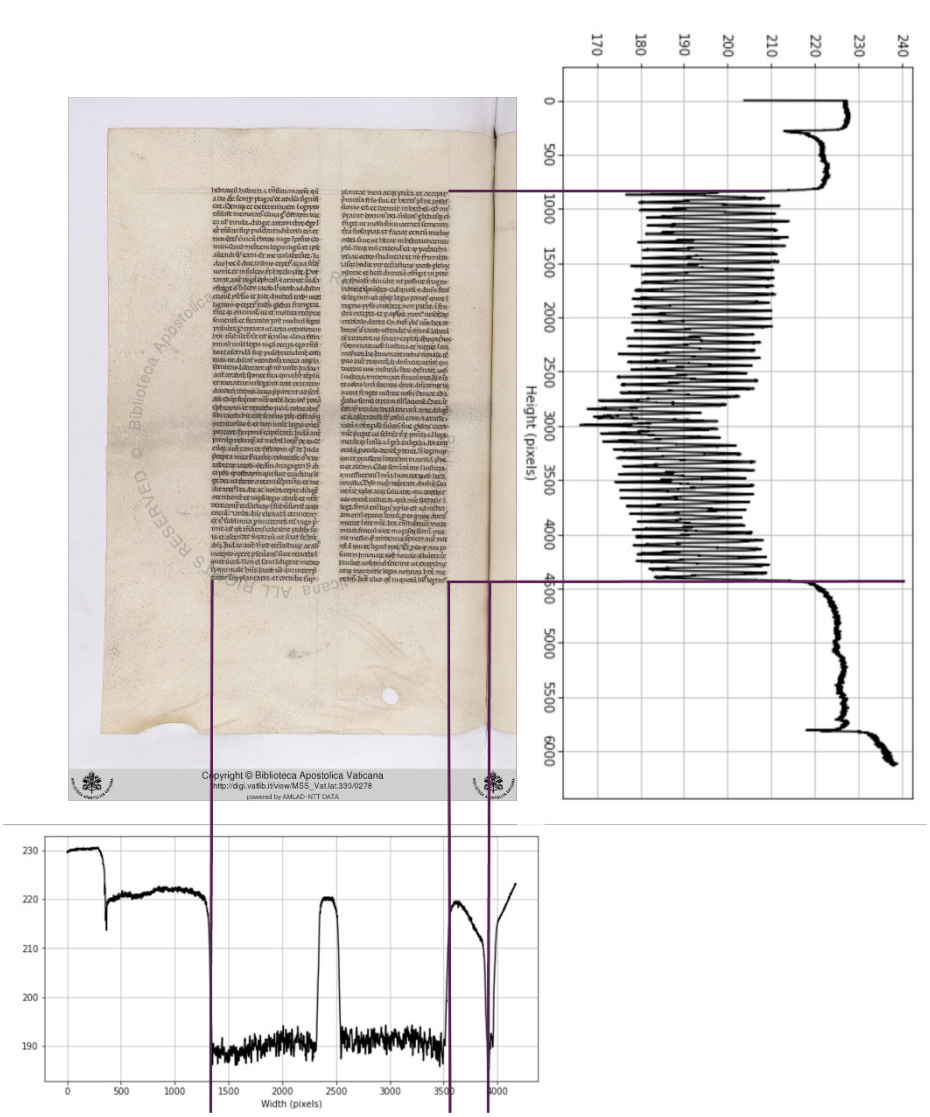

Fig. 2. Vat.lat.330, folio 137v along with vertical and horizontal projections. The red lines, automatically detected by the software, mark the written area edges and the inner edge of the page.

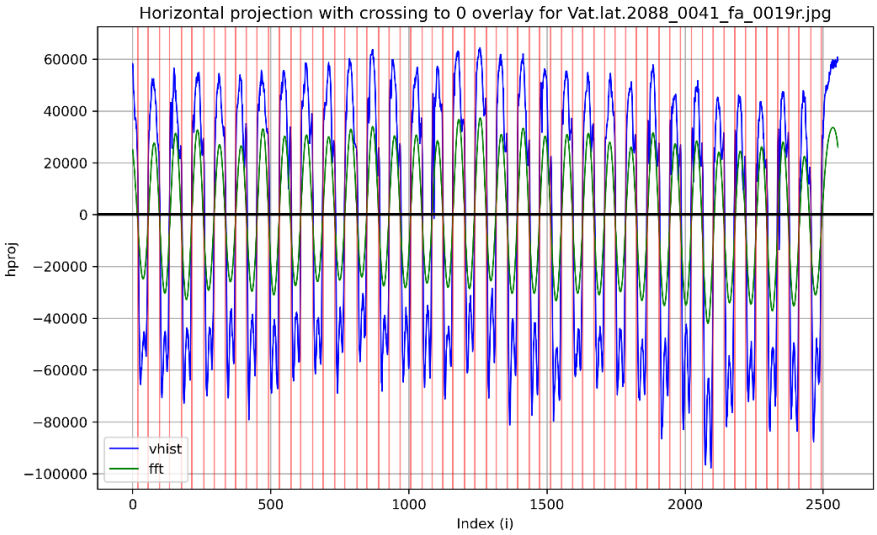

Fig. 3. Vat.lat.2088, folio 19r. Horizontal projection of the written area (blue line) along with a fft interpolation (green). In red the intercept between the fft interpolation and the 0.

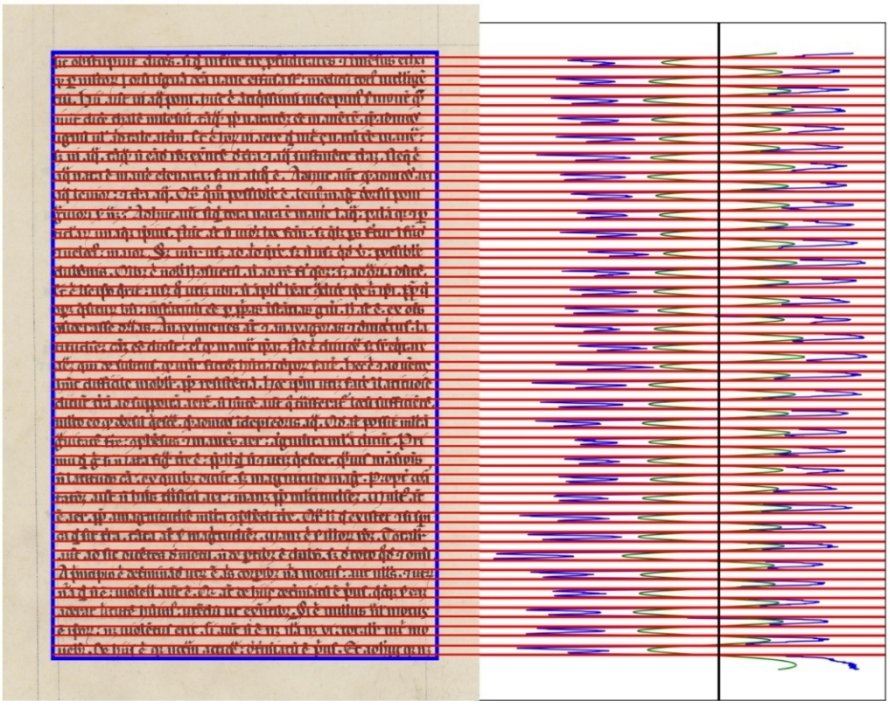

Fig. 4. Vat.lat.2088, written area of folio 19r. On the right the horizontal projection along with the y coordinates detected by the algorithm delimitating the lines of text and the line spacing.

4 Analysis

In this paragraph, the variations of codicological parameters between the year 800 AD and 1500 AD will be analyzed. The software developed for this work and described in the previous paragraph extracts the codicological parameters for each analyzed image, i.e., for each folio of the manuscript. Subsequently, the obtained parameters are combined to generate a single representative value for each manuscript. At this stage, pages with less text, such as richly illuminated pages, are excluded. For this reason, the manuscripts have been filtered as described in paragraph 2: manuscripts with few text pages relative to the total number of pages or those with text located only on limited portions of the pages are not of interest to this study.

After obtaining representative parameter values for each volume, the manuscripts were grouped by century of production. For each century, the average parameter values and their dispersion (represented by the standard deviation) were calculated. Regarding text line height and line spacing, three average trends were identified: (1) average measurements per century (converted into millimeters) for manuscripts with page heights between 20–30 cm, (2) average measurements for those with page heights between 30–40 cm, and (3) normalized values for both parameters, calculated by dividing the measurements by page heights, including manuscripts of all sizes. In this way, the normalized values are presented as a fraction of the page height.

The obtained values and dispersions are shown in Fig. 5.

An in-depth analysis of the plots in Fig. 5 lies beyond the scope of this work. The trends concerning ink percentage, filling coefficient, and page dimensions have already been discussed in previous works: Giuffrida (2024) and Giuffrida & Manoni (2025). Regarding text line height and line spacing, the transition from early medieval to Gothic style stands out. This transition reveals a marked reduction in line spacing and a minor reduction in text height. This finding aligns with expectations, considering that Gothic characters are characterized by shortened shafts, which are undetectable by the implemented algorithm (see Section 3).

The average variations in parameters across centuries is in line with established knowledge about the history of writing (see, for example, AA.VV. 2020, Bozzolo et al. 1980–1982, Maniaci 1995, Ornato 2020). This correlation validates the work conducted so far and prompts the following question: Can the codicological parameters obtained here determine the production date of a manuscript, or at least constrain its production to a probable range of years?

To explore this question, it will be useful to analyze the combined evolution of the parameters over the centuries. Specifically, the goal is to assess whether manuscripts produced within close temporal proximity form recognizable clusters in an n-dimensional space, where n represents the number of relevant codicological parameters. Several tools can aid this analysis. This work employs two techniques: dimensional reduction via principal component analysis (PCA) paired with k-means clustering, and unsupervised classification using Self-Organizing Maps (SOM).

Principal component analysis reduces the number of variables in a dataset. Given N variables, PCA projects them onto a new plane where the new variables are arranged in descending order of variance or, in other words, informational content. This method

enables the representation of an N-dimensional point distribution in 2 or 3 dimensions while minimizing information loss.

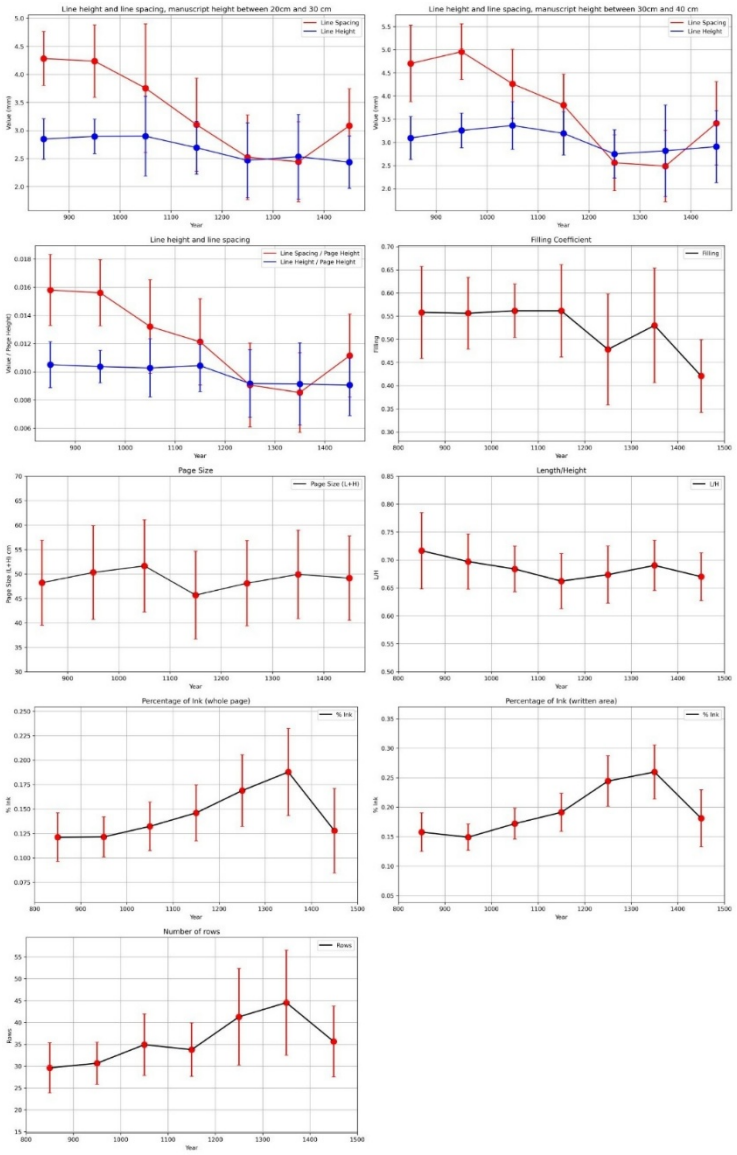

Fig. 5. Average value, per century, of the 8 codicological parameters along with their dispersion.

Clusters within this reduced-dimensional space can then be easily identified using a clustering algorithm. For this purpose, the k-means algorithm was chosen. This method divides data into k groups (where k is predefined) such that the sum of the squared

distances between data points and the center of each cluster is minimized. For further details on PCA and k-means clustering, please refer to Ding & He (2004) and Jolliffe & Cadima (2016).

PCA analysis was performed on eight codicological parameters: filling coefficient, percentage of ink in the entire page, percentage of ink in the written area, number of rows, line spacing, text height, page size, and page ratio. The first two principal components were subsequently analyzed using k-means clustering. While various values of k were tested, k=3 proved to be the most insightful, as increasing the number of clusters did not improve the separation of manuscripts produced during different periods.

Fig. 6 illustrates the spatial distribution of the first two principal components for the three clusters identified by k-means analysis. It also includes a histogram displaying the distribution of manuscript production years within these clusters. It is important to note that neither the PCA representation nor the k-means clustering considered manuscript production year.

From this analysis, it is clear that k-means clustering does not enable precise differentiation of manuscripts by production period. However, certain patterns emerge. For example, Cluster 1 (orange in the plot) predominantly contains manuscripts produced after 1400, while Cluster 2 (green) is primarily composed of manuscripts from 1200 to 1400. Cluster 0 (blue) features a mix of manuscripts spanning various centuries.

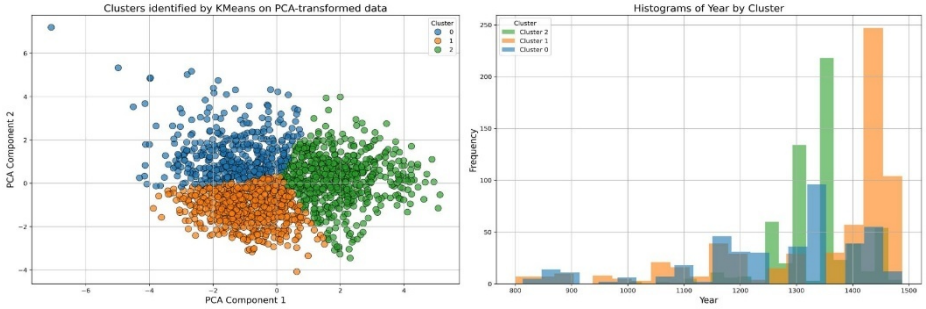

Fig. 6. Left: First vs Second principal component coming from the 8 codicological parameters along with 3 K-mean clusters.
Right: Histogram of year distribution of the 3 clusters.

Based on the results of this preliminary analysis, it is worthwhile exploring how manuscripts shift within the plane of the first two principal components when grouped into four time periods:

- Manuscripts produced between the years 800 AD and 1000 AD
- Manuscripts produced between the years 1000 AD and 1200 AD
- Manuscripts produced between the years 1200 AD and 1400 AD
- Manuscripts produced between the years 1400 AD and 1500 AD.

These four groups are illustrated in Fig. 7.

The plots in Fig. 7 confirm the findings from the k-means analysis: manuscripts produced between 1200 and 1400 occupy a distinct area in the principal component plane

compared to manuscripts produced later, albeit with some overlaps. However, manuscripts produced before 1200 are located within regions also covered by manuscripts from all time periods.

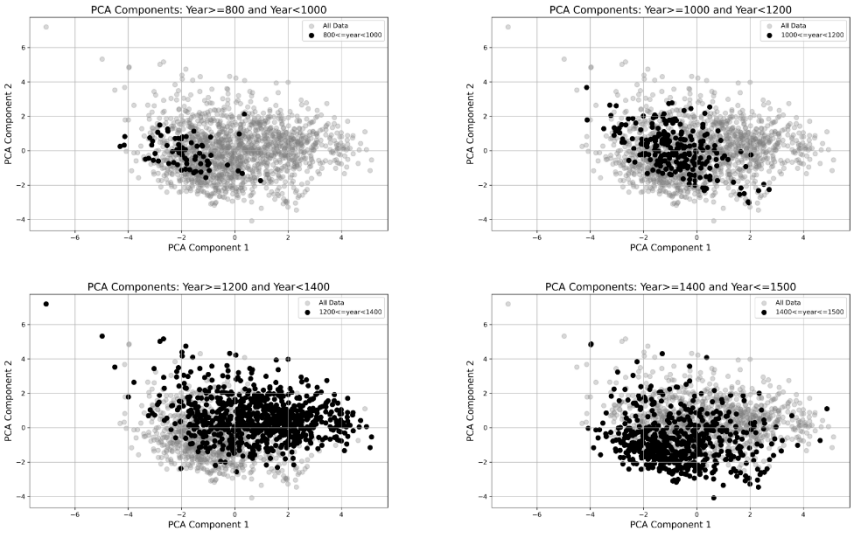

Fig. 7. First vs Second principal component for 4 subgroups selected in years.

To provide further insight, a different algorithm was employed: Self-Organizing Maps (SOM). SOMs (refer to Kohonen, T., 2001) are a type of Artificial Neural Network designed for unsupervised clustering. Using datasets characterized by N parameters, SOM distributes data into a two-dimensional grid, or "map," of size NxM, where N and M are determined as inputs.

Several grid configurations were tested, finding that the optimal choice is a 2x2 grid, dividing the data into four clusters. These groups are depicted in Fig. 8. Interestingly, SOM analysis produced a distribution strikingly similar to the results from PCA paired with k-means clustering, despite being an entirely independent approach. Manuscripts produced between 1200 and 1400 and those produced after 1400 are mostly distributed in separate grids: the former span three grids (blue color), whereas the latter are predominantly concentrated in one grid (green color). Conversely, manuscripts produced between 800 and 1200 are scattered across all four grids.

These two independent approaches to manuscript classification, based on eight codicological parameters, convey a consistent message: manuscripts produced before 1200 are difficult to distinguish based solely on these parameters, but there is greater potential for identifying manuscripts produced between 1200 and 1500.

This brings us back to the central question: Can the codicological parameters obtained so far be used to determine the production date of a manuscript, or at least constrain it to a likely range of years? The answer is yes, at least partially. However, it should be noted that results for manuscripts produced before 1200 are likely to remain unreliable.

The next section will detail the development of an initial automatic manuscript dating system and its subsequent evaluation.

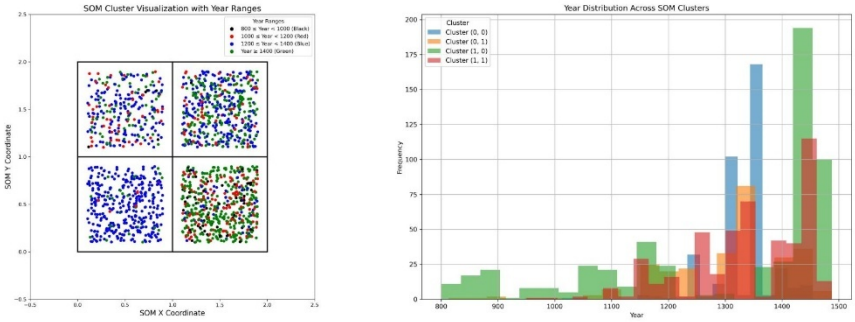

Fig. 8. Left: the 4 groups identified by the SOM algorithm, in 4 colors the ages subgroups. **Right**: Histogram of the years distribution for the 4 SOM clusters.

5 ANN

Before proceeding, it is helpful to summarize the findings so far. The eight codicological parameters (filling coefficient, percentage of ink on the entire page, percentage of ink in the written area, number of rows, line spacing, text height, page size, and page ratio) extracted from 1,650 Western manuscripts exhibit consistent evolutions over the centuries. Multidimensional analyses have further confirmed these trends, providing a basis for developing an automatic system to predict the production date of Western manuscripts based on these parameters. A classic feed-forward neural network was selected for this purpose, with the dataset divided into three groups:

1. **Training dataset:** This dataset trains the network by determining weights that minimize deviations between the expected output and the predicted value. Typically, 70–80% of the dataset is allocated for training (see Nguyen, H. T., et al., 2021; Bichri, H., Chergui, A., & Hain, M., 2024, and references therein). In this case, 1155 manuscripts were used.
2. **Validation dataset:** This dataset improves the network's accuracy during training while minimizing overfitting, ensuring the network performs well on unseen data. The validation dataset includes 247 manuscripts.
3. **Testing dataset:** This dataset is excluded from training and serves to evaluate the network's performance independently. It consists of 248 manuscripts.

The neural network consists of eight input nodes (corresponding to the eight normalized codicological parameters) and one output node (the normalized production year). The number of hidden layers and nodes per layer were optimized through testing with the dataset. The final network design features two hidden layers with 16 and 34 nodes, respectively. A schematic representation is provided in Fig. 9.

The network's behavior across the three datasets was analyzed using residuals, i.e. the difference between the actual production year and the year predicted by the network. Fig. 10 shows trends for the three datasets.

From the analysis of the three plots, the main problem of this network becomes evident: the model misinterprets the general trend of the data, giving much more weight to the data between years 1300 and 1400. Unfortunately, such a result was predictable just by observing the histogram in Fig. 1: most available manuscripts were produced within this range of data. Clearly, the network is much more influenced by these manuscripts than by those produced before 1300 or after 1500.

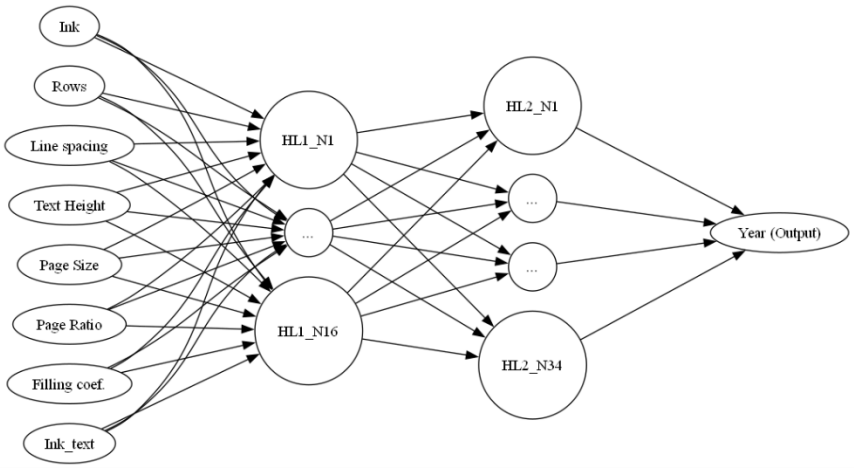

Fig. 9. Schematic representation of the ANN design, the network contains two hidden layers, the first one having 16 nodes and the second one 34 nodes.

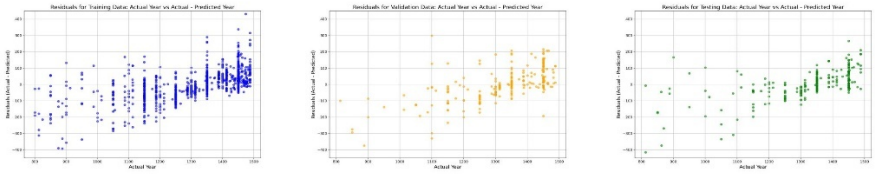

Fig. 10. Year Vs Residuals for the three datasets. From left to right: training, validation, testing.

Additionally, it is worth recalling what was reconstructed from the multidimensional study of codicological parameters in the previous paragraph: manuscripts produced before 1200 are difficult to distinguish from those produced after.

To further investigate, the testing dataset (unseen during training) was evaluated for three age ranges: 800–1200, 1200–1400, and 1400–1500. Fig. 11 presents histograms of residuals for these ranges.

As expected, the first histogram peaks at negative values, while the second peaks at positive values; only the central histogram is roughly symmetrical around zero. However, it is worth examining the overall histogram of residuals for the whole testing dataset, shown in Fig. 12.

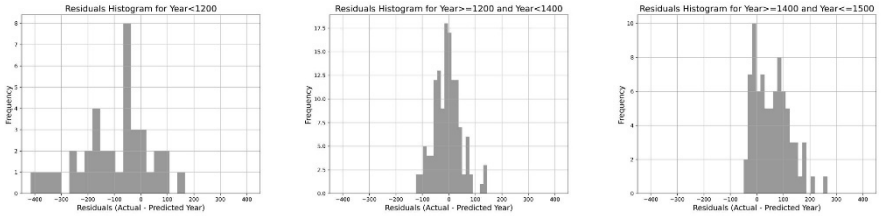

Fig. 11. Histograms of the residuals for the testing dataset in 3 range of years.

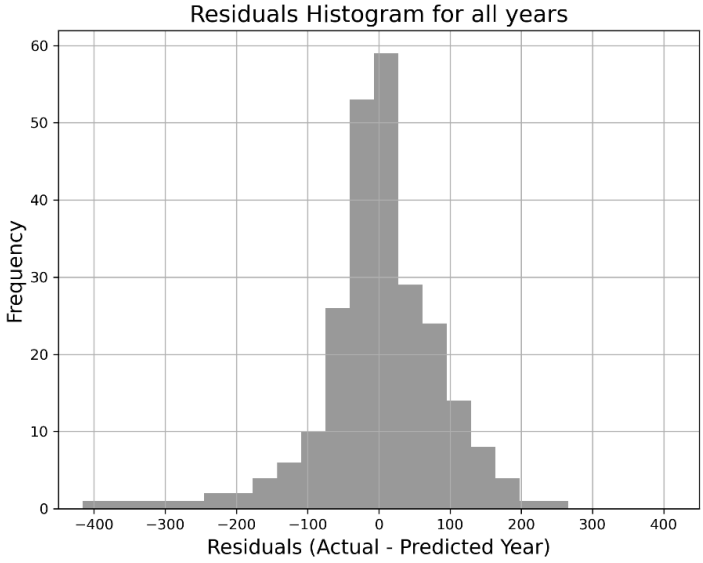

Fig. 12. Years residuals for the testing dataset.

Despite the network's limitations, several positive outcomes can be noted:

1. The average residual is 0.31, with a standard deviation of 89.4. This is comparable with the dating uncertainty for the manuscripts, being often around 50 (see Fig. 1). In the range of 1200–1400 (central histogram of Fig.11), the standard deviation drops further to 50.67, a result unlikely to be improved with the current dataset.
2. The network makes significant errors in the range 800-1200; however, it can mostly correctly determine whether a manuscript was produced before or after 1200. Considering what was highlighted in the previous paragraph, this result is already quite good on its own.

Finally, as an independent test, four manuscripts from the Bodleian Library's digital collection were analyzed. Due to unknown DPI settings, page size was excluded, and seven codicological parameters were extracted. This led to the creation of a new neural network using only seven parameters, its behavior is nevertheless very similar to the previously described network.

The table below shows the results obtained on the four manuscripts' pages. The results align with earlier conclusions: the network tends to overestimate production years for manuscripts created before 1300 and underestimate years for those produced after 1400. Nonetheless, it correctly distinguishes manuscripts produced before 1200 from those produced between 1300 and 1400.

Tab. 1. Dating of 4 pages coming from Bodleian Library's manuscripts. The column "Actual Year" contains the Date Statement as available in https://digital.bodleian.ox.ac.uk/, while the column "Predicetd Year" contains the output of the neural network

Mss	Folio	Actual Year	Predicted Year
MS. Junius 25	4r	8th century, end 9th century	1037
MS. Auct. D. 5. 5	5	12th century; 13th century, second half	1313
MS. Add. A. 263	3v	14th century	1384
MS. Add. A. 26	5r	15th century, second half	1379

6 Conclusions and Future Prospects

This study analyzed 1,650 Western manuscripts, dated between 800 AD and 1500 AD, from the digital corpus of the Vatican Apostolic Library. It summarized the methodologies used to extract eight codicological parameters from the 600,000 pages comprising these manuscripts. The research demonstrated variations in these parameters across centuries, employing multidimensional analysis techniques. Additionally, a preliminary neural network was developed to estimate manuscript production dates.

However, the creation of an automatic dating system encounters two significant challenges: the inherent uncertainty in manuscript dating, which averages around 50 years, and the limited availability of manuscripts from the early Middle Ages. Furthermore, the eight codicological parameters utilized thus far appear insufficient for precise and fully reliable dating.

To enhance the results, progress can be made by addressing multiple aspects simultaneously:

1. **Enriching the dataset**, prioritizing the manuscripts produced before 1200 AD and those dated or datable with the highest precision.
2. **Increasing the number of codicological parameters** by adding, for example, the number of characters on the page (or, even better, their density, known as the exploitation coefficient).

3. **Generating synthetic data** from the available dataset to enhance the efficiency of the neural network.

References

Bichri, H., Chergui, A., & Hain, M. (2024). Investigating the impact of train/test split ratio on the performance of pre-trained models with custom datasets. *International Journal of Advanced Computer Science and Applications*, *15*(2). https://doi.org/10.14569/IJACSA.2024.0150235

Bozzolo, C., & Ornato, E. (1980). *Pour une histoire du livre manuscrit au Moyen Âge. Trois essais de codicologie quantitative*. Éditions du Centre national de la recherche scientifique.

Bozzolo, C., Coq, D., Muzerelle, D., & Ornato, E. (1982). Noir et blanc. Premiers résultats d'une enquête sur la mise en page dans le livre médiéval. In *Il libro e il testo. Atti del convegno internazionale* (pp. 195–221). Università degli Studi di Urbino.

Cherubini, P. (2004). Una nuova ricetta in volgare per rigare la pagina (secolo XV). *Miscellanea Bibliothecae Apostolicae Vaticanae*, *11*, 241–258. https://doi.org/10.1400/227012

Coulson, F. T., & Babcock, R. G. (Eds.). (2020). *The Oxford handbook of Latin palaeography*. Oxford University Press.

De Stefano, C., Fontanella, F., Maniaci, M., & Scotto di Freca, A. (2011). A method for scribe distinction in medieval manuscripts using page layout features. In *Proceedings of the International Conference* (pp. 393–402). Springer. https://doi.org/10.1007/978-3-642-24085-0_41

Ding, C., & He, X. (2004). K-means clustering via principal component analysis. In *Proceedings of the 21st International Conference on Machine Learning* (pp. 29–36). ACM. https://doi.org/10.1145/1015330.1015408

Foffi, M. (2021). *Development of large scale analysis techniques for feature extraction in ancient manuscripts of the Vatican Apostolic Library* [Master's thesis, Tor Vergata University].

Giuffrida, G. (2024). A percentage of ink across the centuries: A first analysis of the digital corpus of the Vatican Library. In A. M. Piazzoni (Ed.), *The process for the digitization of manuscripts in the Vatican Library*. Biblioteca Apostolica Vaticana.

Giuffrida, G., & Manoni, P. (2025). Digital libraries and digital codicology: The exploitation of the Vatican Apostolic Library's FITS archive and new experimental approaches in AI and IIIF. In A. Campagnolo & E. Pierazzo (Eds.), *Approaches to digital codicology: Interdisciplinarity and intersections*. Brepols. (forthcoming)

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Kohonen, T. (2001). *Self-organizing maps* (3rd ed.). Springer. https://doi.org/10.1007/978-3-642-56927-2

Maniaci, M. (2022). Statistical codicology: Principles, directions, perspectives. In M. Maniaci (Ed.), *Trends in statistical codicology* (pp. 1–32). De Gruyter. https://doi.org/10.1515/9783110743838-001

Maniaci, M., & Ornato, E. (1995). Intorno al testo: Il ruolo dei margini nell'impaginazione dei manoscritti greci e latini. *Nuovi Annali della Scuola Speciale per Archivisti e Bibliotecari*, *9*, 175–194.

Nguyen, H. T., et al. (2021). Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering*, 2021, Article ID 4832864. https://doi.org/10.1155/2021/4832864

Ornato, E. (2020). The application of quantitative methods to the history of the book. In F. T. Coulson & R. G. Babcock (Eds.), *The Oxford handbook of Latin palaeography* (pp. 651–668). Oxford University Press.

Received: April 30, 2025
Reviewed: May 14, 2025
Finally Accepted: June 10, 2025