Analyzing Historical Set Data: the Case of Medieval Biblical Prologues

Sébastien de Valeriola (Université libre de Bruxelles)

Céline Engelbeen (ICHEC Brussels management school)

Chiara Ruzzier (Université de Namur)

1. Medieval Biblical Prologues

In the Middle Ages, Bible manuscripts do not all have the same textual content: they contain a varying number of supplementary texts of different kinds. In particular, copyists insert short texts (called "prologues") at the beginning of most of the biblical Books (one before the Book of Psalms, one before the Gospel of Matthew, etc.) intended to introduce the topics covered in the Book. Most of the time, there exists several prologues for the same biblical Book (each attributed to different authors). They have been studied by historians, but only from a qualitative perspective [1]. The purpose of this paper is to study this phenomenon with quantitative methods, using data mining and visualization tools.

The reasons which pushed the copyists to provide such manuscripts with such prologues do not appear in an obvious way. The historiography observed that a specific group of 66 prologues is found in the majority of biblical manuscripts copied in Paris during the 13-14th centuries: the Parisian sequence [2]. Our first objective is to search for other coherent groups of prologues, either on the basis of geographical origin or other criteria.

Not all the biblical Books are similar in terms of the prologues that accompany them: in some cases, the same prologue is used in the majority of manuscripts; in others, several different prologues "compete" with each other. Our second objective is to profile the biblical Books according to this distinction.

2. Historical Set Data

The data on which we work can be seen as a list of sets (the manuscripts) each containing a certain number of objects (the prologues). We thus handle a binary 353×328 matrix M, whose each entry M_{ij} indicates whether the i-th manuscript contains the j-th prologue. Moreover, some information about these sets and objects is known: geographical origin and dating for the manuscripts, biblical Book and presumed author for the prologues.

Both our research questions are related to the search for groups of prologues with common characteristics (e.g. biblical Book) contained in manuscripts with common characteristics (e.g. geographic origin).

3. Methods and Results

3.1 Searching for Coherent Groups of Prologues

To answer our first research question, we started by considering factorial methods (like PCA). The results were not satisfactory, mainly because the first factorial axis was always very close to the sum of M's columns. This issue is directly related to the heterogeneity of the dataset and proved to be one of the big challenges of our study: some prologues are attested in many manuscripts, and others almost none.

We then turned to (hierarchical) clustering methods, using several different distances on the columns of M (see [3]). As expected from the issue raised just above, some distance choices lead to clusters whose content is directly dictated by the sum of M's columns. Other distance choices put on the same footing very rare and very frequent prologues, leading to uninteresting results.

We also investigated association rule mining tools. The relationships they allow to discover highlighted some interesting phenomena, e.g. a group of prologues almost exclusively used in Italian manuscripts. This group, like all the others identified in the same way, is unfortunately not large enough to be of interest in our overall study.

The very mixed success of these three families of methods led us to look at the data from all angles using visualization tools. A group of prologues has then emerged (that we named the "Medium sequence"), which can be characterized as the prologues that are moderately frequent (between 50 and 150 manuscripts) and are more often found in Italian than in French manuscripts. Moreover, they are mostly written by a particular group of authors. A specialist in biblical prologues has confirmed that this grouping makes sense from a historical perspective.

3.2 Profiling Biblical Books

To answer our second research question, we divided the columns of M in 58 groups, each containing the prologues associated with the same biblical Book. We then compared the "manuscript co-membership" structure of these prologues within each group, to establish the profile of the biblical Books. Exploring this kind of dependence structure is not an easy task. Venn diagrams do the job in simple cases, but become unreadable as soon as we look at groups of more than 3 or 4 prologues.

To explore these relationships, we have drawn three sets of complementary plots:

a. UpSet plots [4], which are an excellent tool to get an overall view of prologues intersections (the intersection of two prologues is the set of manuscripts that contain them both). However, they lose their interest if we want to study certain types of intersections in particular,
b. network visualizations, whose nodes are prologues of the biblical Book under study. Each pair of prologues is linked with an edge, whose weight is a metric representing the degree of incompleteness of the co-membership relation between the two prologues. These networks allow to see the 2 by 2 intersections of the sets, but they do not allow to study closely intersections of 3 (or more) sets,
c. scatter plots (manuscripts × prologues): a dot at (x, y) indicates that the y prologue is attested in the x manuscript. These plots make it possible to see if the intersections of the prologues are grouped or if they are mutually exclusive.

To profile the biblical Books, we extracted from the last two types of plots two metrics for each biblical book, giving robust estimates of its number and the diversity of its prologues:

- Two books can be considered as extremes: Canonical Epistles' prologues are numerous and rather diverse, Revelation's prologues are few in number but very diverse,
- There is a dense group of books with few and not very diversified prologues, close to the situation of a single prologue belonging to the Parisian sequence,
- There is a more sparse group of books with several prologues from the Parisian sequence, with small diversity,
- Some books have mixed profiles and are found floating between these groups, like Psalms and Habakkuk.

Bibliography

[1] De Bruyne, D., "Prefaces to the Latin Bible. Introductions by Pierre-Maurice Bogaert and Thomas O'Loughlin", in *Studia traditionis theologiae* 19, Turnhout, 2015.

[2] Ruzzier, C., "Entre Université et ordres mendiants. La production des bibles portatives latines au XIII^{e} siècle", in *Manuscripta Biblica* 8, p. 68-72, Berlin/Boston, 2022.

[3] Gower, J.C., and Legendre, P. "Metric and Euclidean properties of dissimilarity coefficients", in *Journal of Classification* 3, p. 5-48, 1986.

[4] Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. "UpSet: Visualization of Intersecting Sets", in *IEEE Transactions on Visualization and Computer Graphics* 20 (12), p. 1983-1992, 2014.