2012

University of North Carolina Wilmington

Master of Science in

Computer Science and Information Systems

Proceedings

https://csbapp.uncw.edu/mscsis

COMPUTATIONAL METHODS FOR DETERMINING THE SIMILARITY BETWEEN ANCIENT GREEK MANUSCRIPTS

Eddie Dunn

A Capstone Project Submitted to the
University of North Carolina at Wilmington in Partial Fulfillment
of the Requirements for the Degree of
Master of Science

Department of Computer Science
Department of Information Systems and Operations Management

University of North Carolina Wilmington

2012

Approved by

Advisory Committee

Dr. George Zervos	Dr. BryanReinicke

Dr. Curry Guinn, Chair

Accepted By

Dean, Graduate School

Table of Contents

Abstract

COMPUTATIONAL METHODS FOR DETERMINING THE SIMILARITY BETWEEN ANCIENT GREEK MANUSCRIPTS. Dunn, Eddie, 2012. Capstone Paper, University of North CarolinaWilmington.

The computational piece of this project utilizes data extracted from the work of B. Daniels and George Zervos. Their unpublished dissertations from Duke University present two separate critical apparatus of 89 and 45, respectively, (all different) manuscripts of the New Testament apocryphon/pseudepigrapha book now most commonly called Protoevangelium (Proto-gospel) of James. This paper will first present this early Christian work as very unique and important document whose dating and authorship has critical implications in biblical studies. This paper will then explore computational methods of preparing the textual data that might best elicit quantifiable, statistical variances/similarities in the documents. After the data preparation, several classification and exploratory data analysis techniques will be applied that provide verification for the current expert opinions on finding groups of documents that are more similar, and hence might have come from the same "family" of documents. Because there are other versions of the Protoevangelium of James that have not received much scrutiny, it is also hoped this line of inquiry might help better classify the dating and authorship of these documents. These understandings will give us further insight in the fascinating story of this document and by corollary our understanding of early Christian literature.

Chapter 1: Introduction

The Problem Domain

The idea of being able to distinguish the writing of different authors based on quantitative techniques alone has been present in the literature for quite some time. In the intervening 126 years since Mascol was using similar techniques under the moniker of stylometry in 1886 with, fittingly, the letters of Paul, many technological innovations and associated advancement in technique have made possible the present study. While the majority of the content contained herein will be firmly planted in the technical realm, the inherently multi-disciplinary nature of this inquiry makes this a very unique and interesting

opportunity to participate in several very important conversations and will not be overlooked. It is the goal of this researcher to provide enough background in both the technical and humanities' aspects of this work to make it accessible to readers from any related field.

Papyrology and Paleography

Papyrology is the study of ancient documents composed and copied on papyrus. The use of papyrus, made from the cross-laid strips of internal stem pith from a plant by the same name, dates back to the third century BCE in Egypt. This material gained wide usage throughout the Mediterranean region. It was cheap and easy to produce but the rolls of the material necessary for larger documents were prone to cracking and quick deterioration if stored in a humid environment. As a result we have very few of the plethora of these documents that once existed, and what we do have are often barely decipherable pieces. Of the best preserved examples we still have, almost all came from an arid environment that prevented deterioration. Papyrology as a scientific discipline gained popularity as large numbers of these documents were recovered by archeologists in the late nineteenth and early twentieth centuries. It is accepted to also include works on materials such as parchment and vellum (made from animal skin) in this category of study although they are not strictly *papyri* [3].

Paleography is a more general discipline that is defined as the study of ancient writing. Papyrology can be defined as a subset of Paleography and in this context this work is also of interest to the Paleography community..

Textual Criticism

Textual Criticism is the area that aspires to remove errors (whether intentional or unintentional) in an attempt at coming as close to creating the "original" or source documents(s) as possible. The hallmark of this type of work is a critical apparatus to show *variant* readings alongside a primary text (also called a base text). The traditional approach of laying out a primary source from the preparer's best estimate is an inherently subjective undertaking and as a result more recent scholars have adopted the practice of using an existing copy as the base text.

Authorship Attribution/Text Classification

Information Retrieval (IR)

The data preparation techniques employed in this research are firmly planted in the discipline known as Information Retrieval, an area that is heavily leveraged in the data sciences. Using techniques found in the mathematical discipline of linear algebra we will create a normalized vector of relevant features in multidimensional space for each document we are examining. This is generally referred to as the vector space model and is the focus of this work.

Clustering and Data Visualization

The field of machine learning has a sub discipline called classification that attempts to group like items in a set. Techniques from this field enjoy wide application across the computer/data science field.

Four techniques are used as part of this study:

1. K-means

K means attempts to find a solution to a set of data for a given (user supplied) number of classes such that the average distance among all members of the class is minimized.

2. Hierarchical clustering

   The technique takes a measure of dissimilarity and then in a top down approach builds a tree graph. Starting with each item in its own group the algorithm then finds the two most similar classes and merges them. The result is one less class than before. The process is repeated until all the documents have been consumed generating a tree structure in its wake. To visualize the clusters rectangles are painted on the dendrogram plot at the appropriate levels of the tree based on the number of clusters desired.

3. Correspondence Analysis

   This family of techniques has garnered a following in the ecological science community in analyzing species abundance information across geographically distinct sites. We will use a couple of members of this group, namely De-Trended Correspondence Analysis (DCA), Canonical Correspondence Analysis (CCA), and Non-Metric Multidimensional Scaling (NMDS).

The listed classifiers were chosen based their ability to show the similarity break down.

Chapter 2: Review of Literature and Analysis

Because this study crosses two very different disciplines there are essentially two separate literature reviews identified. First the biblical studies historical perspective is presented with respect to the document and its traditions. Attention is then focused on the placement of this study in the stylometric literature.

History of the analysis of the *Protoevangelium of James*

The Protoevangelium of James most likely dates back to the middle to later part of the 2nd Century CE. This document has been known by several names. In the earlier years of its life it was likely called Book of James as it is referred to in the writings of Origen who died in the middle of the 3rd Century CE [4]. As its most commonly known name in the literature today implies, proto-gospel means just that -- that it is a story before the gospels or life of Jesus. It seems to have been composed largely in reaction to the accusations by contemporary critics that were assaulting the burgeoning religion on the grounds that the parents of its messiah were commoners. The writer of this document portrays Joseph as a rich building contractor and Mary as herself being immaculately conceived and brought up (with her chastity protected) as a revered temple virgin perhaps in direct response to these accusations.

Our document is also interesting as the oldest manuscript we have is complete and it is significantly different than its closet contemporary versions as well as the majority of the later surviving examples. [11] It also enjoys the luxury of, while not being part of the canon, still being widely copied and distributed throughout the ancient world, especially in those eastern traditions with highly developed Mariological themes. Mariology is, as its name implies, the study of Mary the mother of Jesus. This term has a much more profound meaning to the traditions that evolved in the eastern world. In fact there are Eastern Orthodox feast days established based on information in this document. While all of the documents we will be performing computations on are in Greek, there are surviving copies of this document in many languages including Coptic, Syriac, Ethiopic,

Armenian, Georgian, and Slavonic. There is also a scholarly notion that an Arabic copy might have “influenced Qur’anic and later Islamic understandings of the place of Mary in the Christian tradition”[5], yet another way this text has impacted western religion.

For this study the focus is on two separate exhaustive collections of our gospel. These collections that are the basis for our dataset are specifically the unpublished dissertations from Duke University of Daniels(BD) and Zervos(GZ). These collections are both presented in their own critical apparatus. A critical apparatus in this context is an accepted way of showing how different copies of the same documents vary (called variant readings). There are over 167 extant Greek versions of the P J. Scholars have found that the earliest copy(from the Bodmer V collection) is very different than the base text used by Daniels and the base text used by GZ. For this reason GZ(and the editors of other, more recent, critical editions of other documents) chose to use the oldest and most complete version he had in his set of documents and his "Critical Edition". To further complicate matters neither BD nor GZ had the ability to include Bodmer V in their original apparatus. The Bodmer V collection is now available for research purposes. This present research work can be said to draw from its own "Critical Edition" of looking at both BD and GZ (as well as Bodmer V) altogether. We throw out the chosen base text of Daniels as it does not represent an actual, physical document.

In the course of the literature review the work of Timothy J Finney at the Vose Seminary in Australia was identified and found to have many similarities with the work contained herein. He uses many of the same analysis techniques described and he also employs the

use of the open-source statistics analysis application R. He specifically has worked much with a very hard problem not addressed in this research. [9] This is the problem called alignment in the NLP literature. The basic idea is that when one is performing machine translation you want to be able to impose markers at some at least quasi-regular interval(s) in each document so that if one is at point x in one document they will also be at point x in the parallel document(s). This is precisely the thing you want to do when generating critical apparatus but with the added complication of many more than two documents and potentially many more than two languages.

The main difference in his document similarity work and what is presented here is that the techniques described in this research are completely automated and work with strictly statistics of frequency count data and impose no inherent notion of variant readings or any alignment information other than what has been preserved in the encoding of the character grams. This technique, in theory, should work with little to no modification with any natural language! It is not the suggestion of this research that expert, human encoded information from presenting the data is without merit but it should be step two now that we have this arsenal of tools to examine the documents in their natural state. It would seem for the sake of future generations of research we should strive to accurately catalog photograph and create xml text transcriptions of single documents in and of themselves. If we strive for this then we make this data much more available to anyone wishing to undertake its study “from the ground up”. In the not too distant future we will have the ability to automate (or at the very least streamline greatly) the generation of

human digestible apparatus for study of variants. However now that it is feasible and even desirable to preserve each document separately we should move towards that end.

Computational methods in Authorship Attribution

While there were previous attempts dating back to the 19th Century at using statistical measures in attributing authorship, it was not until the publication of *Inference and Disputed Authorship: The Federalist* by Mosteller and Wallace in 1964 that this area of "non-traditional" authorship attribution study gained widespread attention.[1]. Previous work had attempted to use features such as average sentence length and rate of use of articles and pronouns. They found that while the rates of use in the case of some words such as "the" did not vary in a statistically significant manner from author to author, the use of what they refer to as connector words, such as "upon", can vary by as much as 3 standard deviations. Mosteller and Wallace used such features as word counts and rate of use of specific, non-article or pronoun words. By examining the distributions of individual words it was discovered that some word rates were best described by a Poisson distribution and others were better approximated with a negative binomial distribution. Bayesian inference was then applied using the probabilities calculated using the appropriate distribution. This was ultimately to come down on the side of supporting the historical notion that Madison was likely the author of the 12 then disputed Federalist papers. Their study also outlines a basic work flow of technique application that is still followed.

Chapter 3: Methodology and Challenges

Character Sets, OCR Applications, and Parsing algorithms

The first challenge encountered was homogenizing everyone working in the team in keeping the same UTF-8 character codes for each Greek letter (we assume the use of only lower case in this study) and keeping vigilant on the file formats necessary to later accurately read in the files. The data was originally provided in the form of rich text format documents generated from OCR (using Read Iris Pro) scans of the original hand typed manuscripts of both Daniels and Zervos. [2] [3] Early work utilized data extracted from these files directly. The initial idea was that we would be able to use the "raw" output from Read Iris (with slight modification) and be able to parse the data directly from these slightly/quickly modified files.

The process of reading in files that the Read Iris program outputted in rtf format combined with the additional complexity in the way of markup overhead generated by word processing applications resaving the files after edit prompted a very in depth look at the rtf format and several different java parsers as well as some custom attempts at coding rtf parsers in an ultimately vain attempt to be able to read and "auto-correct" to any useful degree the raw output from the OCR.

While we were able to achieve a great deal of success with this method, because of the many complications with using the raw rtf, it was decided at this juncture (thanks to the student resources available) to come up with our own html-like document format that was designed to be both easily parse-able by humans as well as programmatically while allowing the necessary flexibility to properly notate various meta-data aspects specific to

the discipline. This also allowed for a thorough, multiple-pass, multiple-person proofreading effort to, in the end, better ensure data accuracy and integrity.

Initially only the first chapter (then later, first three chapters) of the complete set of both Daniels and Zervos were considered because of the resources required to produce the source data in the quasi-xml format that was specified. It was later discovered that the Text Encoding Initiative has already defined xml standards for the description of these types of documents. It is this format that any final electronic compilation of our gospel should be presented in. In the intervening time from the initial work and the present effort the rest of the document was completed in a proofread, BD and GZ combined, single apparatus in Word .doc format. In the later stages of the project (and at considerable effort) the data for the entire document was processed directly from these files. For the analysis presented in this project the entire set of manuscripts from both BD and GZ as well as the Bodmer papyri was considered.

Feature Selection

Mainly due to the limited understanding of the Greek language by the primary researcher, this research uses features that would likely be applicable across language domain boundary. This research uses a composite feature space consisting of uni-grams and bi-grams, as well as character sequence grams of length 2-5. This produced a full set feature space of approximately 85,000 unique tokens for the whole document across all manuscripts. Note there is some overlap with features (specifically whole words of length 2-5) in that they will be over-represented in the counts. However, it was determined that

the results were not significantly altered from this overlap and for the exploratory nature of this work, was sufficient.

Classification techniques

Making use of the statistics program R and its library of tools this research applied K-means and hierarchical clustering, as well as DCA, CCA, and NMDS correspondence/ordination analysis techniques from an ecology package in R called vegan. The maintainer of the package Jari Oksanen had some great tutorials that proved invaluable in understanding the vegan tools and plotting. The decision to use R, aside from its widespread use, was made to allow for time to explore more statistical techniques and also as grounds for ease of repeatability by subsequent analysis. [10]

Chapter 4: Research Experiment

Using ordination analysis and clustering techniques widely applied in the data sciences this research created a log normalized vector in multidimensional space of TF-IDF weighted composite feature space for each document examined. We derive a correlation matrix by taking each document vector and multiplying it by the vectors for all the other documents to come up with a two dimensional structure containing the pair wise, angular cosine similarities of each document with every other document. We can then take this result matrix and apply our set of data analysis tools. We also use the scaled, TF-IDF values for each manuscript for each feature directly by using R and applying Bray-Curtis (also known as Sørensen similarity index or Jaccard index) statistical distance measure from the R package vegan.

Description of Data

Data Pre-Processing

The final set of MS-word files consisted of a homogenized critical apparatus of both Daniels and Zervos manuscripts set against Daniel's base text. This data had to first be further parsed and processed to clean up human data entry irregularities and was then programmatically deconstructed into its constituent exemplars. Due to the late nature of the arrival of the full data set in the project life-cycle no extra care was taken to account for scribal modifications. This allows for a close approximation of what was originally on the papyrus without additional parsing. This will obviously have to be revisited if any serious attempt at a TEI encoded version of this collation is to take place.

Here is a modified sample of data to illustrate:

10:6

λαχετε μοι ωδε

81 92 001 003 005 006 102 103 107-109 111 113 114 116 118

203 204 207 208 210 213-215 218 219 302 306 407 411 501

512-514 601 606 609 612 617 619 701 707 801-803 805 902

[] 61 308

λαχεται μοι ωδε 002 105 112

+ και το αργυρον 628

The interpretation of this scheme would be as follows: The base text has "λαχετε μοι ωδε" for this part of Chapter 10 verse 6 and all the samples listed on the line below agree with this reading to the letter. Sample 61 and 308 lacked this reading. Sample 002, 105,

and 112 had the text "λαχεται μοι ωδε" instead of the base text reading "λαχετε μοι ωδε". 623 had, in addition to the base text, the text "και το αργυρον" . After this step had been completed there was a separate file for each sample that contained the text of the respective document.

Term Frequency (TF)

$$\mathrm{tf}(t,d) = \frac{\mathrm{f}(t,d)}{\max\{\mathrm{f}(w,d) : w \in d\}}$$

In generating our raw TF counts an associative array is created for each document sample that consists of each feature from the entire corpora and a count of the number of times that the particular feature appears, for each document. Here is an example:

D0 = { [ιστοριαις,6], [δωδεκα,9], [ηγκικεν,4] … }
D1 = { [ιστοριαις,0], [δωδεκα,5], [ηγκικεν,12] … }

This data is called the term frequency (TF). Before any weighting, smoothing, or normalization can be performed additional data is needed. This data is as follows: For each feature, In how many of the samples does it occur? This information is crucial in the IDF portion of the calculations. So if the above example was representative of the entire corpora, raw TF counts would look like this:

C = { [ιστοριαις,6], [δωδεκα,14], [ηγκικεν,16] … }

Next each term is divided by the highest frequency count of any one term across all of the corpus (to remove document length bias). The final TF weights in the example would be:

C = { [ιστοριαις,0.375], [δωδεκα,0.875], [ηγκικεν,1.0] … }

Inverse Document Frequency (IDF)

The weighting of each feature is adjusted by calculating the inverse document frequency (IDF). Inverse document frequency is a technique that will cause words that appear least frequently among the documents (but more than in just a single document) to cause much higher influence than words that appear in many or nearly all of the documents. This technique alone is extremely powerful and even with a bug in the early code that lead to the TF data to be way off, interestingly enough yielded nearly identical results as when the code was corrected.

$$\operatorname{idf}(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

Compute TF/IDF weights

We can now combine all of this into a unified weight matrix. Here is the formula

$$\operatorname{tfidf}(t, d, D) = \operatorname{tf}(t, d) \times \operatorname{idf}(t, D)$$

The prepared data can now be further examined by an arsenal of tools to make sense of the not-at-all human digestible output of the Information Retrieval step. [7]

Hardware/Memory

When starting to allocate memory to perform the java cosine matrix calculations it became apparent that having a 64bit OS with ample memory was ideal (at the time I discovered this I was working with a 32bit OS VM). The data file that is written out (showing up to 18 decimal places of floating point precision for the full set 131 X 85,122) was itself more than 71.5 MB. To speed up calculations I cache all data I might potentially need in memory ahead of time. It was noticed that the java portion of this process would consume up to 8 GB of memory while running and R with the entire final workspace loaded consumes over 4 GB of memory idle which makes it a requirement

that you have 64bit hardware and corresponding software stack to even load the final dataset used in calculations.

Enter R:

At this point we have all of the data we need to perform our calculations. Here we also move our tool from eclipse/Java to R by means of UTF-8 .csv text files. Java code was written to output .csv files with the following sets of data for each of the three groupings (more on the groupings forthcoming):

1) The raw frequency counts for each feature
2) The documents appeared counts for each feature
3) The TF-IDF values for each document and feature
    a. These are the values we calculated above.
4) The Euclidean distance/cosine similarity correlation matrix
    a. These are a result of calculations coded in java. This is mainly a control to make sure that the distance measures and ordination techniques in R are doing similar things.

After the input of the full set of documents it became clear that it would make the most sense to break down the total manuscripts considered into three groups to help attempt to address plot clutter and also just to show “old” manuscripts separately. The groups are as follows: Full Set of 131 manuscripts, Only Old consisting of the oldest 32 manuscripts (those before 1100 CE), and Really Old consisting of the oldest 9 manuscripts. After all the data was read into data frames (with labels) into an R workspace the following operations were performed and saved into the workspace.

```
library(vegan)
# Sørensen (Bray/Curtis) similarity index
fullSet.dist <- vegdist(fullSet)
# Detrended Correspondence (DCA) analysis
fullSet.deco <- decorana(fullSet.dist)
# NMDS operation
fullSet.mds  <- metaMDS(fullSet.dist)
# Canonical Correspondence Analysis (CCA)
fullSet.cca <- cca(fullSet.dist)
```

This was done not only for the entire document but also each chapter for each of the six sets of documents. The time to read in the data and perform the calculations to create the workspace used in final analysis was nearly 18 hours and the size of the final .RData file was 1.3 GB (with some extra testing data). That is a lot of data generated from a document where a single "complete" copy of is less than 15 KB!

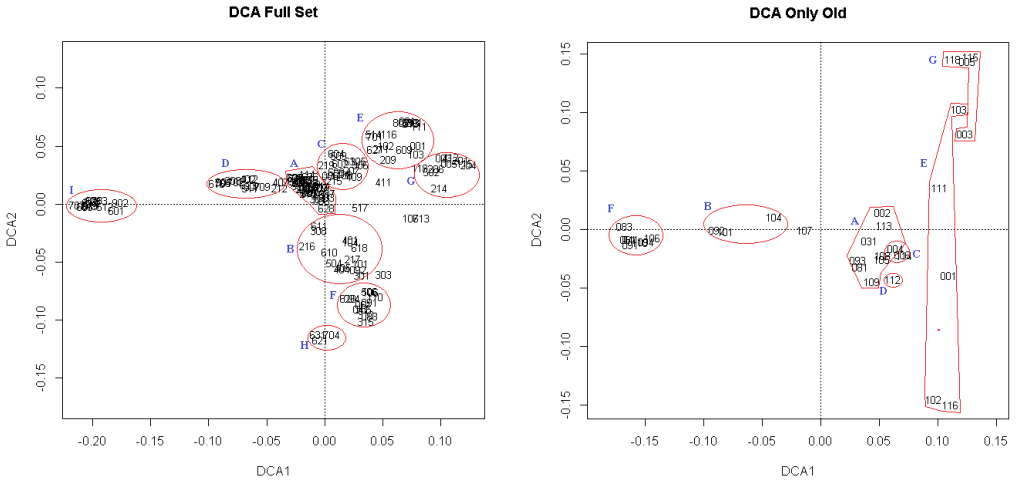
Results

In presenting the results of this study I will first start with characterizing some basic statistics about the feature space. I will show the results of the calculations keeping the set of manuscripts small where appropriate to illustrate technique and at the end of this section bring in what the experts to date have said about the families on the whole.

There are two sets of data that we are working with. One is the raw counts of each feature both on the entire corpus and each manuscript as a whole. The other set is the number of documents in which each feature was seen. The total number of features in the data set is 85,122. At first glance it is amazing at how similar these summaries end up being. In fact the mean is the only significant difference in the summary statistics even though the range is 2-131 in the case of the document appears count and 2-43,095 with the raw counts!

Correlation Analysis

As a next step in further analysis we will first look at the output of our ordination techniques. We use the vegan function ordiplot() to examine the results of DCA, CCA, and NMDS without the species scores. **Note:** When using these tools in this context it is helpful to mentally replace the term **site** with manuscript and **species** with feature. It is also very helpful to play around with xlim and ylim parameters of ordiplot() to zoom in on interesting areas of the plot projections. We will show the DCA plots as they seem to show the most, obvious clusters. Similar plots for NMDS and CCA are included as figures 1 and 2 respectively. The following is only intended as a visual, the full size plots can also be found in the figures section. I would like to point out Bodmer (031), near the center of cluster A on both the Full and Only Old group plots.



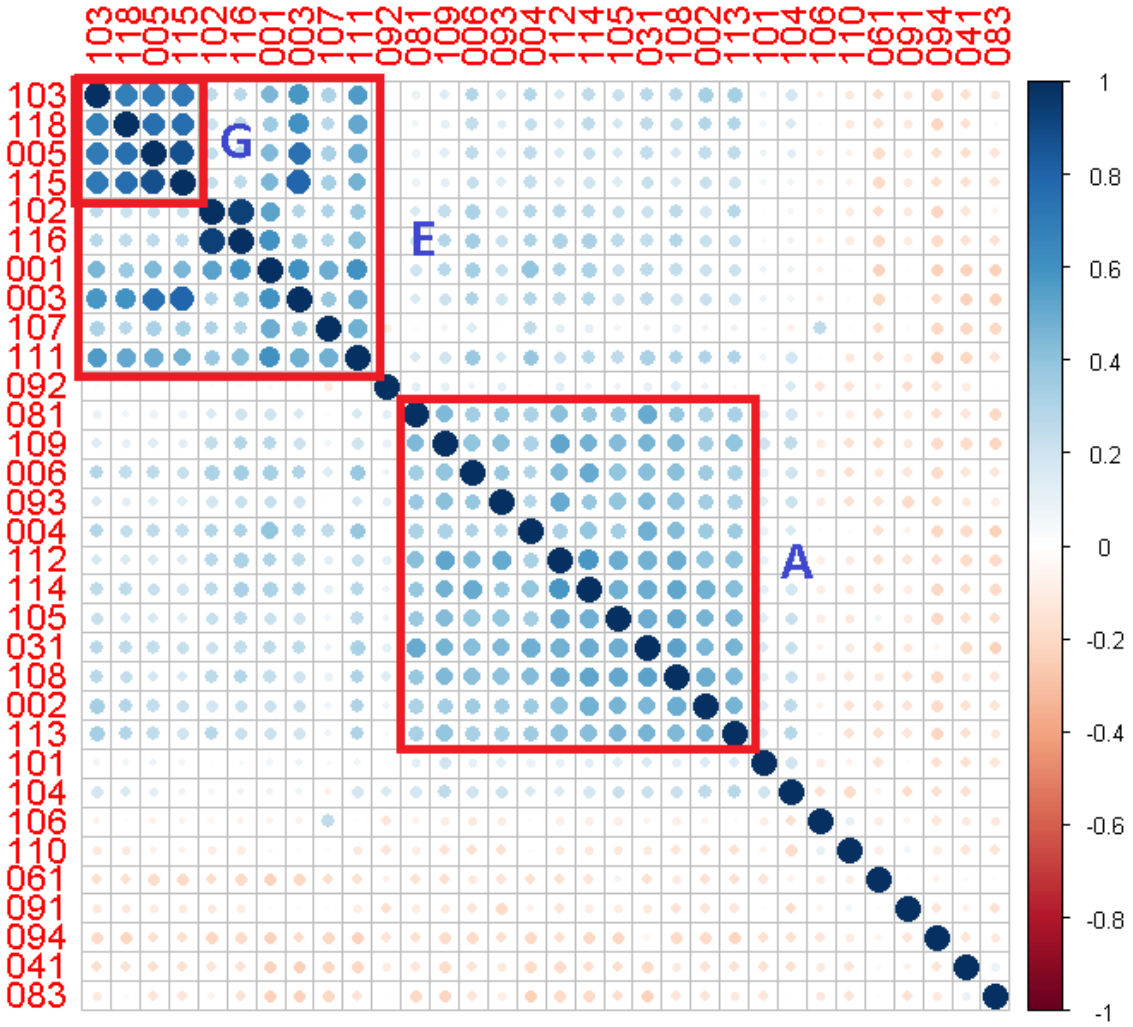
It is worth mentioning here that if one does not specify sites-only (with the parameter display=”sites”) when calling ordiplot() it displays the species information and the plot becomes cluttered with the text of all the features (the greek characters). The position of the feature on the plot corresponds to its contribution to the manuscript (site). R provides a function called identify() that allows the uncovering of the species one at a time with mouse clicks but this is not readily displayable here. It would seem to be very useful to

automate the display of only as many of the most distinguishing features(species) as one can without significantly crowding the plot. The groupings shown were identified by a visual inspection conducted by the primary researcher. Larger versions of the DCA plots and a detail list of the actual manuscripts for each group can be found in the Figures 3,4 and 5.

Corrplot

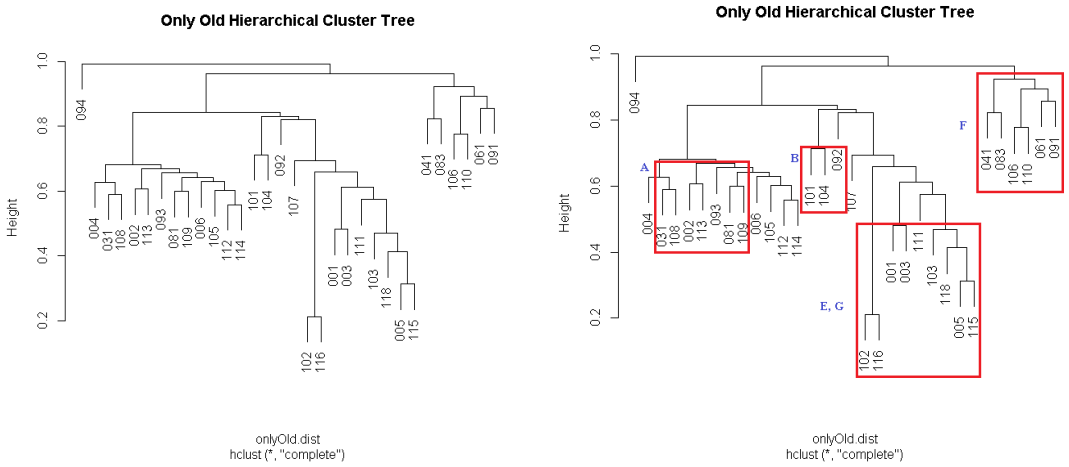
Another visualization technique is the corrplot() function from the R package of the same name. It provides a very pretty visualization of correlation matrices. This technique shows the pairs of manuscripts with the most similarity with an increasingly larger and bluer circle and conversely those with less similarity larger and redder in color. No color or a faint circle indicates neither similarity nor dissimilarity.

A corrplot of the full manuscript set with grouping is in Figure 7 but here is the corrplot() output for only the old manuscripts with the groupings highlighted:



Hierarchical Clustering

Now that we a few data visualizations have been presented let's move on to the cluster analysis. The plot of the dendrogram (tree) of the only old set can be seen below. From this it can be determined that among the oldest manuscripts it appears that there are five of the nine families identified from the full set plot present of which Bodmer (031) is a member of one of those families.



K-means

Attention is now turned to the K-means algorithm. This one in particular is not well suited to exploratory analysis. This is mainly due to the difficult nature of visualizing the results of varying the number of clusters. This is particularly true when one tries to digest the entire set of manuscripts at once. With the hierarchical technique it is easy to see how the clusters fit together at the various values of k. Non-the-less here is K-means for the Only Old set of manuscripts for 11 clusters (Please note letters **do not** correspond to family letters identified thus far):

	A	B	C	D	E	F	G	H	I	J	K
1	m002	m003	m001	m092	m106	m102	m091	m083	m061	m041	m094
2	m004	m005	m107	m101	m110	m116					
3	m006	m103	m111	m104							
4	m031	m115									
5	m081	m118									
6	m093										
7	m105										
8	m108										
9	m109										
10	m112										
11	m113										
12	m114										

Notice that the results are very similar to the results obtained from the hierarchal analysis (also notice Bodmer 031 is in the largest group) but without re-running the algorithm and

comparing the result of each attempt it is impossible to see how the clusters fold together as you decrease k or conversely how they break apart as k increases. The upside is that this was a very quick and easy method to apply and using the package RCaller and Apache POI package used in the data import it was trivial to automate the generation of Excel files for any of the datasets with the results for the entire document and each chapter separately for any number of clusters.

Full Document Results

We now turn our attention to what we can draw from the literature. Because Tichendorf's documents were a small percentage of the documents we have in our study and the questionable thoroughness of deStryker as highlighted by GZ we will examine what is in GZ's dissertation for confirmation of our results. We will refer to the Bray-Curtis hierarchical cluster chart of the entire document set contained in the appendix in the analysis in this section. A more complete and updated critical apparatus of the gospel is being collated at the time of the draft of this study and in large part this study is meant to support that effort. GZ actually uses the two sections highlighted by BD as being the best for seeing the family heritage. These two sections are the "I, Joseph" (IJ) passage that appears from 18:3-19:6 and "Anna's Lament" (AL) which is up to six stanzas of lamentations found in Chapter 3. He cites here that the order of the stanzas and the presence/absence of a given stanza provide additional family information. I feel it necessary that the techniques described will not be able to differentiate things such as the order of the stanzas unless more of a notion of alignment other than chapter/verse is used. However this does not seem to affect the technique's ability to group documents accurately.

The first group of documents is described as the largest by both BD and GZ consisting of 003, 005, 103, 115, 118, 201, 204, 206, 214, 502, and 609. The hierarchical plot shows all of these manuscripts as being in the same cluster. This corresponds to groups E and G from the DCA groupings. GZ suggests that 612 and 409 might also be close and we do place them in the same group in the hierarchical as well as being in Group C on the DCA groupings. However it is not near the rest of the mentioned manuscripts. This is interesting and should be examined with the knowledge of the Greek language.

Another family widely agreed upon is the one made up from 112, 208, 212, 402, 407, 511, 616, 702, 705, 709, and 901. All with the exception of 702 and 709 came from the St. Panteleimonus monastary in Athens so we will call this group by the same name. It is interesting to note that it is mentioned that there are two sub-groups in this family consisting of 511, 702, 709 that follow 212 and 616, 705, 901 that follow 208,402. The hierarchical plot confirms all of this information. Also notice this group corresponds with Group D in our DCA groupings.

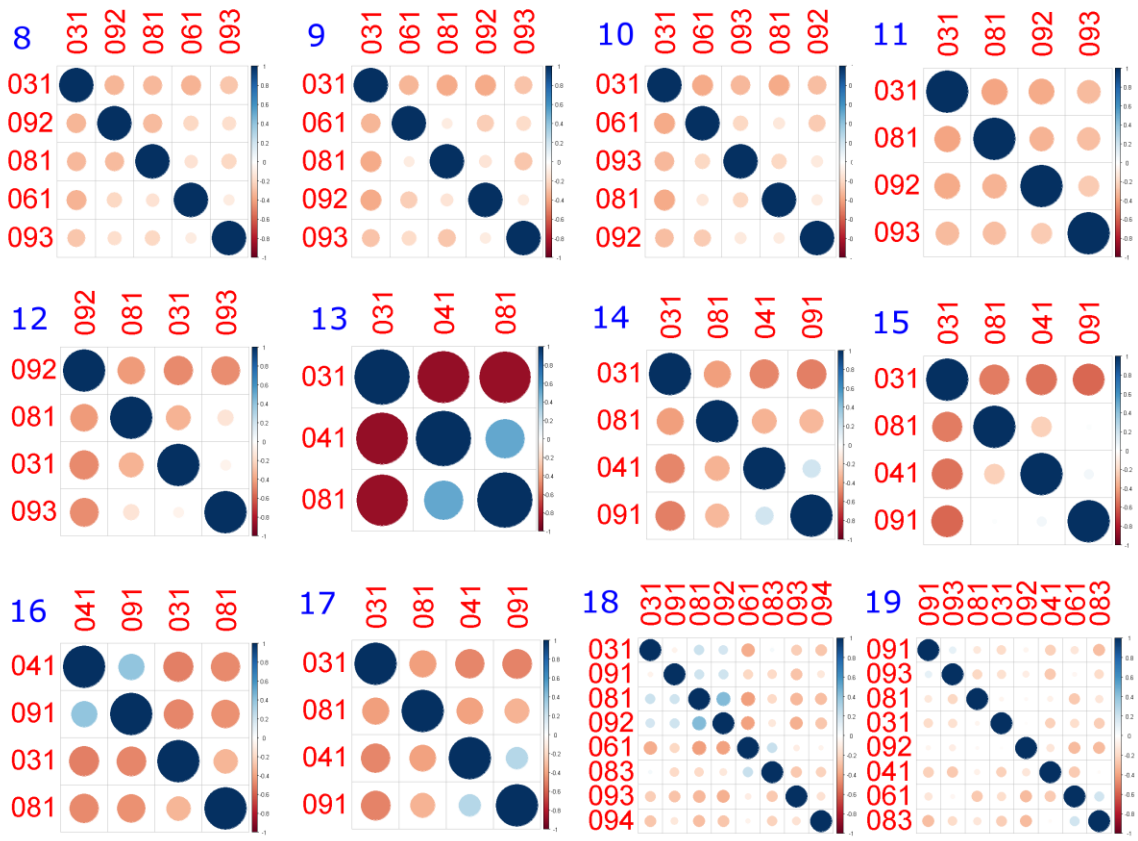
The group 601, 606 is also highlighted. 601 has in its sub-group 512, 615, 619 and 606 with sub-group 617, 703, 707, 803, 805, and 902. Our plot also confirms these observances. This is DCA group I. In looking at the tree it seems proper to this researcher to place the 601, 606 group and its associated documents into the Panteleimon family. It is also worth noting that 621, 631, and 704 (Group H) are mentioned as being in this group as well. It was found that while these three documents were indeed found to

be similar, they were placed a good distance from the rest of its other neighbors and should be examined by experts for further analysis.

Next we examine the group from the monastery of Vatopedi on Mt. Athos. This group consists of 111, 218, 501, 513, 801, 802. Again this is all confirmed in both the hierarchical and DCA plots where it shows up as group E. The group that is now being called the Jerusalem group is also mentioned and consists of 202, 508, 603, 622, and 708. These are lumped in with our DCA Group A. As an aside it is mentioned the similarity of 509 and 604 and also 210 and 220. This information is also confirmed in our plots.

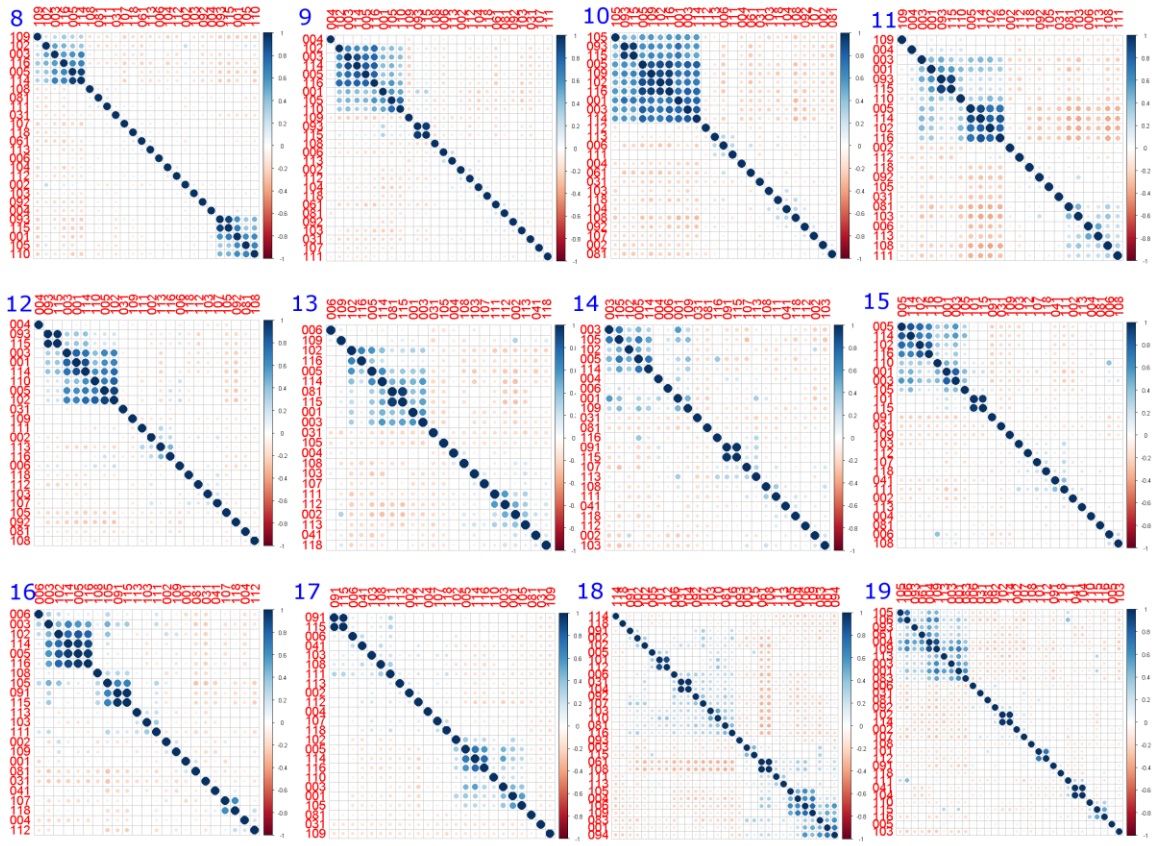
Chapter by Chapter Results

The results of the full document analysis while providing a great deal of information and striking results also shows some confusion with respect to some documents. This is especially clear with the DCA Only Old plot. There are two dynamics that contribute to this effect: The first is that the letter groupings were established from the full set which includes the very tightly grouped but also very different traditions that do not seem to be present in the earliest documents in our set. The second reason is that a more detailed examination of the chapters shows that there is a great deal of variation contained within each manuscript in certain sections versus others. This is where the corrplot really shines. It provides a way to see how each chapter breaks down and that indeed we have situations where in one chapter the scribe is using content from one tradition and then another in different sections. Here is the oldest nine documents chapter-by-chapter from chapters 8-19:



From this we can verify our initial statement that Bodmer (031) with the exception of chapter 18 is indeed different than any of its closet contemporary examples. Notice chapter 13 seems to be a point of wide variation.

Next let's look at this same view of the oldest 32 documents and see if there is any additional statements that can be made:



The most striking information that can be gleaned immediately from the addition of two centuries worth of documents is that we indeed see very tight families emerge. It is also significant that in nearly all instances these families are almost entirely comprised of the later documents.

Chapter 5: Conclusions and Future Work

It is apparent by the agreement among the results of these calculations and the expert, scholarly opinion that these results should garner some level of trust in validity of the results. In fact as is, this technique could be used to assist the study of groups of text transcribed documents in any language, from any period.

Future Work

This area is ripe with potential for future work. It would be extremely interesting to use these techniques in other areas of papyrology where there is a strong surviving manuscript tradition, specifically the canonical New Testament documents.

It would be interesting to apply this technique to the synoptic problem and perhaps to see where we might be able to observe any possible evidence for the primacy of PJ to the canonical accounts as GZ suggests.

For me personally as a researcher this has brought to the forefront in my mind another ripe area of research in data science/analytics and the need for tools to better utilize our rapidly expanding massively parallel computing infrastructures and associated data set size.

References

[1] Mosteller, F & Wallace, D.:. (1964). Inference and disputed authorship: The Federalist. Addison Wesley – Boston, MA

[2] Daniels, Boyd (1956) - The Greek Manuscript Tradition of Protoevangelium Jacobi - Unpublished PhD Dissertation, Duke University Durham, NC

[3] Zervos, George T. (1986) - Prolegomena to a Critical Edition of the Genesis Maria (Protoevangelium Jacobi) - Unpublished PhD Dissertation, Duke University Durham, NC

[4] Schaps, David (2011) – Handbook for Classical Research– Routledge New York, NY

[5] Bart Ehrman and Zlatko Plese (2011) - The apocryphal Gospels Texts and Translations - Oxford University Press - New York, NY

[6] Paul Foster (2009) - The Apocryphal Gospels A very short introduction - Oxford University Press - New York, NY

[7] Manning, Christopher D. and Schutze, Hinrich (1999) – Foundations of Statistical Natural Language Processing – MIT Press Cambridge, MA

[8] Zervos, George (1994) – Dating the Protoevangelium of James: The Justin Martyr Connection – SBLSP pp415-34

[9] Finney, Timothy J., How To Discover Text Groups
http://www.tfinney.net/Groups/index.xhtml

[10] Oksanen, Jari, Multivariate Analysis of Ecological Communities in R: vegan tutorial, http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf

[11] Testuz, Michel, Papyrus Bodmer V: Nativite de Marie, Geneva: Bibliotheca Bodmeriana, 1958

Appendixes

A. English Translation

BOOK OF JAMES, OR PROTEVANGELIUM
From "The Apocryphal New Testament"
M.R. James-Translation and Notes
Oxford: Clarendon Press, 1924

I. I In the histories of the twelve tribes of Israel it is written that there was one Ioacim, exceeding rich: and he offered his gifts twofold, saying: That which is of my superfluity shall be for the whole people, and that which is for my forgiveness shall be for tile Lord, for a propitiation unto me. 2 Now the great day of the Lord drew nigh and the children of Israel offered their gifts. And Reuben stood over against him saying: It is not lawful for thee to offer thy gifts first,-forasmuch as thou hast gotten no seed in Israel. 8 And Ioacim was sore grieved, and went unto the record of the twelve tribes of the people, saying: I will look upon the record of the twelve tribes of Israel, whether I only have not gotten seed in Israel. And he searched, and found concerning all the righteous that they had raised up seed in Israel. And he remembered the patriarch Abraham, how in the last days God gave him a son, even Isaac. 4 And Ioacim was sore grieved, and showed not himself to his wife, but betook himself into the wilderness, and pitched his tent there, and fasted forty days and forty nights, saying within himself: I will not go down either for meat or for drink until the Lord my God visit me, and my prayer shall be unto me meat and drink. II Now his wife Anna lamented with two lamentations, and bewailed herself with two bewailing's, saying: I will bewail my widowhood, and I will bewail my childlessness. 2 And the great day of the Lord drew nigh, and Judith her handmaid said unto! How long humblest thou thy soul? The great day of the Lord hath come, and it is not lawful for thee to mourn: but take this headband, which the mistress of my work gave me, and it is not lawful for me to put it on, forasmuch as I am a handmaid, and it hath a mark of royalty. And Anna said: Get thee from me. Lo! I have done nothing (or I will not do so) and the Lord hath greatly humbled me: peradventure one gave it to thee in subtlety, and thou art come to make me partaker in thy sin. And Judith said: How shall I curse thee, seeing the Lord hath shut up thy womb, to give thee no fruit in Israel ? 3 And Anna was sore grieved [and mourned with a great mourning because she was reproached by all the tribes of Israel. And coming to herself she said: What shall I do ? I will pray with weeping unto

the Lord my God that he visits me]. And she put off her mourning garments and cleansed (or adorned) her head and put on her bridal garments: and about the ninth hour she went down into the garden to walk there. And she saw a laurel-tree and sat down underneath it and besought the Lord saying: O God of our fathers, bless me, and hearken unto my prayer, as thou didst bless the womb of Sarah, and gavest her a son, even Isaac.
III. 1 And looking up to the heaven she espied a nest of sparrows in the laurel-tree, and made a lamentation within herself, saying: Woe unto me, who begat me? And what womb brought me forth for I am become a curse before the children of Israel, and I am reproached, and they have mocked me forth out of the temple of the Lord? 2 Woe unto me, unto what am I likened ? I am not likened unto the fowls of the heaven, for even the fowls of the heaven are fruitful before thee, O Lord. Woe unto me, unto what am I likened ? I am not likened unto the beasts of the earth, for even the beasts of the earth are fruitful before thee, O Lord. Woe unto me, unto what am I likened? I am not likened unto these waters, for even these waters are fruitful before thee, O Lord. 3 Woe unto me, unto what am I likened? I am not likened unto this earth, for even this earth bringeth forth her fruits in due season and blesseth thee, O Lord.

IV. 1 And behold an angel of the Lord appeared, saying unto her: Anna, Anna, the Lord hath hearkened unto thy prayer, and thou shalt conceive and bear, and thy seed shall be spoken of in the whole world. And Anna said: As the Lord my God liveth, if I bring forth either male or female, I will bring it for a gift unto the Lord my God, and it shall be ministering unto him all the days of its life. 2 And behold there came two messengers saying unto her: Behold Ioacim thy husband cometh with his flocks: for an angel of the Lord came down unto him saying: Ioacim, Ioacim, the Lord God hath hearkened unto thy prayer. Get thee down hence, for behold thy wife Anna hath conceived. 3 And Ioacim sat him down and called his herdsmen saying: Bring me hither ten lambs without blemish and without spot, and they shall be for the Lord my God; and bring me twelve tender calves, and they shall be for the priests and for the assembly of the elders; and an hundred kids for the whole people. 4 And behold Ioacim came with his flocks, and Anna stood at the gate and saw Ioacim coming, and ran and hung upon his neck, saying: Now know I that the Lord God hath greatly blessed me: for behold the widow is no more a widow, and she that was childless shall conceive. And Ioacim rested the first day in his house.

V. 1 And on the morrow he offered his gifts, saying in himself: If the Lord God be reconciled unto me, the plate that is upon the forehead of the priest will make it manifest unto me. And Ioacim offered his gifts and looked earnestly upon the plate of the priest when he went up unto the altar of tile Lord, and he saw no sin in himself. And Ioacim said: Now know I that the Lord is become propitious unto me and hath forgiven all my sins. And he went down from the temple of the Lord justified, and went unto his house. 2 And her months were fulfilled, and in the ninth month Anna brought forth. And she said unto the midwife: what have I brought forth? And she said: A female. And Anna said: My soul is magnified this day, and she laid herself down. And when the days were fulfilled, Anna purified herself and gave suck to the child and called her name Mary.
VI. 1 And day by day the child waxed strong, and when she was six months old her mother stood her upon the ground to try if she would stand; and she walked seven steps and returned unto her bosom. And she caught her up, saying: As the Lord my God liveth,

thou shalt walk no more upon this ground, until I bring thee into the temple of the Lord. And she made a sanctuary in her bed chamber and suffered nothing common or unclean to pass through it. And she called for the daughters of the Hebrews that were undefiled, and they carried her hither and thither. 2 And the first year of the child was fulfilled, and Ioacim made a great feast and bade the priests and the scribes and the assembly of the elders and the whole people of Israel. And Ioacim brought the child to the priests, and they blessed her, saying: 0 God of our fathers, bless this child and give her a name renowned for ever among all generations. And all the people said: So be it, so be it. Amen. And he brought her to the high priests, and they blessed her, saying: 0 God of the high places, look upon this child, and bless her with the last blessing which hath no successor. 3 And her mother caught her up into the sanctuary of her bed chamber and gave her suck. And Anna made a song unto the Lord God, saying: I will sing a hymn unto the Lord my God, because he hath visited me and taken away from me the reproach of mine enemies, and the Lord hath given me a fruit of his righteousness, single and manifold before him. Who shall declare unto the sons of Reuben that Anna giveth suck ? Hearken, hearken, ye twelve tribes of Israel, that Anna giveth suck. And she laid the child to rest in the bed chamber of her sanctuary, and went forth and ministered unto them. And when the feast was ended, they gat them down rejoicing, and glorifying the God of Israel.

VII. 1 And unto the child her months were added: and the child became two years old. And Ioacim said: Let us bring her up to the temple of the Lord that we may pay the promise which we promised; lest the Lord require it of us (lit. send unto us), and our gift become unacceptable. And Anna said: Let us wait until the third year, that the child may not long after her father or mother. And Ioacim said: Let us wait. 2 And the child became three years old, and Ioacim said: Call for the daughters of the Hebrews that are undefiled, and let them take every one a lamp, and let them be burning, that the child turn not backward and her heart be taken captive away from the temple of the Lord. And they did so until they were gone up into the temple of the Lord. And the priest received her and kissed her and blessed her and said: The Lord hath magnified thy name among all generations: in thee in the latter days shall the Lord make manifest his redemption unto the children of Israel. And he made her to sit upon the third step of the altar. And the Lord put grace upon her and she danced with her feet and all tile house of Israel loved her.

VIII. 1 And her parents sat them down marveling, and praising the Lord God because tile child was not turned away backward. And Mary was in the temple of the Lord as a dove that is nurtured: and she received food from the hand of an angel. 2 And when she was twelve years old, there was a council of the priests, saying: Behold Mary is become twelve years old in the temple of the Lord. What then shall we do with her ? lest she pollute the sanctuary of the Lord. And they said unto the high priest: Thou standest over the altar of the Lord. Enter in and pray concerning her: And whatsoever the Lord shall reveal to thee, that let us do. 3 And the high priest took the vestment with the twelve bells and went in unto the Holy of Holies and prayed concerning her. And lo, an angel of tile Lord appeared saying unto him: Zacharias, Zacharias~ go forth and assemble them that are widowers of the people, and let them bring every man a rod, and to whomsoever the Lord shall show a sign, his wife shall she be. And the heralds went forth over all the

country round about Judaea, and the trumpet of the Lord sounded, and all men ran thereto.

IX. 1 And Joseph cast down his adze and ran to meet them, and when they were gathered together they went to the high priest and took their rods with them. And he took the rods of them all and went into the temple and prayed. And when he had finished the prayer he took the rods and went forth and gave them back to them: and there was no sign upon them. But Joseph received the last rod: and 1o, a dove came forth of the rod and flew upon the bead of Joseph. And the priest said unto Joseph: Unto thee hath it fallen to take the virgin of the Lord and keep her for thyself. 2 And Joseph refused, saying: I have sons, and I am an old man, but she is a girl: lest I became a laughing-stock to the children of Israel. And the priest said unto Joseph: Year the Lord thy God, and remember what things God did unto Dathan and Abiram and Korah, how the earth clave and they were swallowed up because of their gainsaying. And now fear thou, Joseph, lest it be so in thine house. And Joseph was afraid, and took her to keep her for himself. And Joseph said unto Mary: Lo, I have received thee out of the temple of the Lord: and now do I leave thee in my house, and I go away to build my buildings and I will come again unto thee. The Lord shall watch over thee.

X. 1 Now there was a council of the priests, and they said: Let us make a veil for the temple of the Lord. And the priest said: Call unto me pure virgins of the tribe of David. And the officers departed and sought and found seven virgins. And the priests called to mind the child Mary, that she was of the tribe of avid and was undefiled before God: and the officers went and fetched her. And they brought them into the temple of the Lord, and the priest said: Cast me lots, which of you shah weave the gold and the undefiled (the white) and tile fine linen and the silk and the hyacinthine, and the scarlet and the true purple. And the lot of the true purple and the scarlet fell unto Mary, and she took them and went unto her house. [And at that season Zacharias became dumb, and Samuel was in his stead until the time when Zacharias spake again.]But Mary took the scarlet and began to spin it.

XL 1 And she took the pitcher and went forth to fill it with water: and lo a voice saying: Hail, thou that art highly favored; the Lord is with thee: blessed art thou among women. And she looked about her upon the right hand and upon the left, to see whence this voice should be: and being filled with trembling she~ went to her house and set down the pitcher, and took the purple and sat down upon her seat and drew out the thread. 2 And behold an angel of the Lord stood before her saying: Fear not, Mary, for thou hast found grace before the Lord of all things, and thou shalt conceive of his word. And she, when she heard it, questioned in herself, saying: Shall I verily conceive of the living God, and bring forth after the manner of all women? And the angel of the Lord said: Not so, Mary, for a power of the Lord shall overshadow thee: wherefore also that holy thing which shall be born of thee shall be called the Son of the Highest. And thou shalt call his name Jesus: for he shall save his people from their sins. And Mary said: Behold the handmaid of the Lord is before him: be it unto me according to thy word.

XII 1 And she made the purple and the scarlet and brought them unto the priest. And the priest blessed her and said: Mary, the Lord God hath magnified thy name, and thou shalt be blessed among all generations of the earth. 2 And Mary rejoiced and went away unto Elizabeth her kinswoman: and she knocked at the door. And Elizabeth when she heard it cast down the scarlet (al. the wool) and ran to the door and opened it, and when she saw

Mary she blessed her and said: Whence is this to me that the mother of my Lord should come unto me? for behold that which is in me leaped and blessed thee. And Mary forgot the mysteries which Gabriel the archangel had told her, and she looked up unto the heaven and said: Who am I, Lord, that all the generations of the earth do bless me? 8 And she abode three months with Elizabeth, and day by day her womb grew: and Mary was afraid and departed unto her house and hid herself from the children of Israel. Now she was sixteen years old when these mysteries came to pass.

XIII. I Now it was the sixth month with her, and behold Joseph came from his building, and he entered into his house and found her great with child. And he smote his face, and cast himself down upon the ground on sackcloth and wept bitterly, saying: With what countenance shall I look unto the Lord my God? and what prayer shall I make concerning this maiden? for I received her out of the temple of the Lord my God a virgin, and have not kept her safe. Who is he that hath ensnared me ? Who hath done this evil in mine house and hath defiled the virgin? Is not the story of Adam repeated in me? for as at the hour of his giving thanks the serpent came and found Eve alone and deceived her, so hath it befallen me also? 2 And Joseph arose from off the sackcloth and called Mary and said unto her O thou that wast cared for by God, why hast thou done this? Thou hast forgotten the Lord thy God. Why hast thou humbled thy soul, thou that wast nourished up in the Holy of Holies and didst receive food at the hand of an angel? 3 But she wept bitterly, saying: I am pure and I know not a man. And Joseph said unto her: Whence then is that which is in thy womb? and she said: As the Lord my God liveth, I know not whence it is come unto me.

XIV. I And Joseph was sore afraid and ceased from speaking unto her (or left her alone), and pondered what he should do with her. And Joseph said: If I hide her sin, I shall be found fighting against the law of the Lord: and if I manifest her unto the children of Israel, I fear lest that which is in her be the seed of an angel, and I shall be found delivering up innocent blood to the judgment of death. What then shall I do? I will let her go from me privily. And the night came upon him. 2 And behold an angel of the Lord appeared unto him in a dream, saying: Fear not this child, for that which is in her is of the Holy Ghost, and she shall bear a son and thou shalt call his name Jesus, for he shall save his people from their sins. And Joseph arose from sleep and glorified the God of Israel which had shown this favor unto her: and he watched over her.

XV. I Now Annas the scribe came unto him and said to him: Wherefore didst thou not appear in our assembly? and Joseph said unto him: I was weary with the journey, and I rested the first day. And Annas turned him about and saw Mary great with child. 2 And he went hastily to the priest and said unto him: Joseph, to whom thou bearest witness [that he is righteous] hath sinned grievously. And the priest said: Wherein? And he said: The virgin whom he received out of the temple of the Lord, he hath defiled her, and married her by stealth (lit. stolen her marriage), and hath not declared it to the children of Israel. And the priest answered and said: Hath Joseph done this? And Annas the scribe said: Send officers, and thou shalt find the virgin great with child. And the officers went and found as he had said, and they brought her together with Joseph unto the place of judgment. 3 And the priest said: Mary, wherefore hast thou done this, and wherefore hast thou humbled thy soul and forgotten the Lord thy God, thou that wast nurtured in the Holy of Holies and didst receive food at the hand of an angel and didst hear the hymns and didst dance before the Lord, wherefore hast thou done this? But she wept bitterly,

saying: As the Lord my God liveth I am pure before him and I know not a man. 4 And the priest said unto Joseph: Wherefore hast thou done this? And Joseph said: As the Lord my God liveth I am pure as concerning her. And the priest said: Bear no false witness but speak the truth: thou hast married her by stealth and hast not declared it unto the children of Israel, and hast not bowed thine head under the mighty hand that thy seed should be blessed. And Joseph held his peace.

XVI 1 And the priest said: Restore the virgin whom thou didst receive out of the temple of the Lord. And Joseph was full of weeping. And the priest said: I will give you to drink of the water of the conviction of the Lord, and it will make manifest your sins before your eyes. 2 And the priest took thereof and made Joseph drink and sent him into the hill-country. And he returned whole. He made Mary also drink and sent her into the hill-country. And she returned whole. And all the people marveled, because sin appeared not in them. 3 And the priest said: If the Lord God hath not made your sin manifest, neither do I condemn you. And he let them go. And Joseph took Mary and departed unto his house rejoicing, and glorifying the God of Israel.

XVII. 1 Now there went out a decree from Augustus the king that all that were in Bethlehem of Judaea should be recorded. And Joseph said: I will record my sons: but this child, what shall I do with her? How shall I record her? As my wife? Nay, I am ashamed. Or as my daughter? but all the children of Israel know that she is not my daughter. This day of the Lord shall do as the Lord willeth. 2 And he saddled the she-ass, and set her upon it, and his son led it and Joseph followed after. And they drew near (unto Bethlehem) within three miles: and Joseph turned himself about and saw her of a sad countenance and said within himself: Peradventure that which is within her paineth her. And again Joseph turned himself about and saw her laughing, and said unto her: Mary, what aileth thee that I see thy face at one time laughing and at another time sad? And Mary said unto Joseph: It is because I behold two peoples with mine eyes, the one weeping and lamenting and the other rejoicing and exulting. 8 And they came to the midst of the way, and Mary said unto him: Take me down from the ass, for that which is within me presseth me, to come forth. And he took her down from the ass and said unto her: Whither shall I take thee to hide thy shame? for the place is desert.

XVIII. I And he found a cave there and brought her into it, and set his sons by her: and he went forth and sought for a midwife of the Hebrews in the country of Bethlehem. 2 Now I Joseph was walking and I walked not. And I looked up to the air and saw the air in amazement. And I looked up unto the pole of the heaven and saw it standing still, and the fowls of the heaven without motion. And I looked upon the earth and saw a dish set, and workmen lying by it, and their hands were in the dish: and they that were chewing chewed not, and they that were lifting the food lifted it not, and they that put it to their mouth put it not thereto, but the faces of all of them were looking upward. And behold there were sheep being driven, and they went not forward but stood still; and the shepherd lifted his hand to smite them with his staff, and his hand remained up. And I looked upon the stream of the river and saw the mouths of the kids upon the water and they drank not. And of a sudden all things moved onward in their course.

XIX. I And behold a woman coming down from the hill country, and she said to me: Man, whither goest thou? And I said: I seek a midwife of the Hebrews. And she answered and said unto me: Art thou of Israel? And I said unto her: Yea. And she said: And who is she that bringeth forth in the cave? And I said: She that is betrothed unto me. And she

said to me: Is she not thy wife? And I said to her: It is Mary that was nurtured up in the temple of the Lord: and I received her to wife by lot: and she is not my wife, but she hath conception by the Holy Ghost. And the midwife said unto him: Is this the truth? And Joseph said unto her: Come hither and see. And the midwife went with him. 2 And they stood in the place of the cave: and behold a bright cloud overshadowing the cave. And the midwife said: My soul is magnified this day, because mine eyes have seen marvelous things: for salvation is born unto Israel. And immediately the cloud withdrew itself out of the cave, and a great light appeared in the cave so that our eyes could not endure it. And by little and little that light withdrew itself until the young child appeared: and it went and took the breast of its mother Mary. And the midwife cried aloud and said: Great unto me to-day is this day, in that! Have seen this new sight. 3 And the midwife went forth of the cave and Salome met her. And she said to her: Salome, Salome, a new sight have I to tell thee. A virgin hath brought forth, which her nature alloweth not. And Salome said: As the Lord my God liveth, if I make not trial and prove her nature I will not believe that a virgin hath brought forth.

XX. 1 And the midwife went in and said unto Mary: Order thyself, for there is no small contention arisen concerning thee. Arid Salome made trial and cried out and said: Woe unto mine iniquity and mine unbelief, because I have tempted the living God, and lo, my hand falleth away from me in fire. And she bowed her knees unto the Lord, saying: O God of my fathers, remember that I am the seed of Abraham and Isaac and Jacob: make me not a public example unto the children of Israel, but restore me unto the poor, for thou knowest, Lord, that in thy name did I perform my cures, and did receive my hire of thee. 3 And lo, an angel of the Lord appeared, saying unto her: Salome, Salome, the Lord hath hearkened to thee: bring thine hand near unto the young child and take him up, and there shall be unto thee salvation and joy. 4 And Salome came near and took him up, saying: I will do him worship, for a great king is born unto Israel. And behold immediately Salome was healed: and she went forth of the cave justified. And Io, a voice saying: Salome, Salome, tell none of the marvels which thou hast seen, until the child enter into Jerusalem.

XXI 1 And behold, Joseph made him ready to go forth into Judaea. And there came a great tumult in Bethlehem of Judaea; for there came wise men, saying: Where is he that is born king of the Jews? for we have seen his star in the east and arc come to worship him. 2 And when Herod heard it he was troubled and sent officers unto the wise men. And he sent for the high priests and examined them, saying: How is it written concerning the Christ, where he is born? They say unto him: In Bethlehem of Judaea: for so it is written. And he let them go. And he examined the wise men, saying unto them: What sign saw ye concerning the king that is born? And the wise men said: We saw a very great star shining among those stars and dimming them so that the stars appeared not: and thereby knew we that a king was born unto Israel, and we came to worship him. And Herod said: Go and seek for him, and if ye find him, tell me, that I also may come and worship him. 3 And the wise men went forth. And lo, the star which they saw in the east went before them until they entered into the cave: and it stood over the head of the cave. And the wise men saw the young child with Mar~, his mother: and they brought out of their scrip gifts, gold-and frankincense and myrrh. 4 And being warned by the angel that they should not enter into Judaea, they went into their own country by another way.

XXII. 1 But when Herod perceived that he was mocked by the wise men, he was wroth, and sent murderers, saying unto them: Slay the children from two years old and under. 2 And when Mary heard that the children were being slain, she was afraid, and took the young child and wrapped in swaddling clothes and laid him in an ox-manger. 3 But Elizabeth when she heard that they sought for John, took him and went up into the hill-country and looked about her where she should hide him: and there was no hiding-place. And Elizabeth groaned and said with a loud voice: 0 mountain of God, receive thou a mother with a child. For Elizabeth was not able to go up. And immediately the mountain clave asunder and took her in. And there was a light shining always for them: for an angel of the Lord was with them, keeping watch over them.

XXIII. I Now Herod sought for John, and sent officers to Zacharias, saying: Where hast thou hidden thy son? And he answered and said unto them: I am a minister of God and attend continually upon the temple of the Lord: I know not where my son is. 2 And the officers departed and told Herod all these things. And Herod was wroth and said: His son is to be king over Israel. And he sent unto him again, saying: Say the truth: where is thy son? for thou knowest that thy blood is under my hand. And the officers departed and told him all these things. 3 And Zacharias said: I am a martyr of God if thou sheddest my blood: for my spirit the Lord shah receives, because thou sheddest innocent blood in the fore-court of the temple of the Lord. And about the dawning of the day Zacharias was slain. And the children of Israel knew not that he was slain.

XXIV. 1 But the priests entered in at the hour of the salutation, and the blessing of Zacharias met them not according to the manner. And the priests stood waiting for Zacharias, to salute him with the prayer, and to glorify the Most High. 2 But as he delayed to come, they were all afraid: and one of them took courage and entered in: and he saw beside the altar congealed blood: and a voice saying: Zacharias hath been slain, and his blood shall not be wiped out until his avenger come. And when he heard that word he was afraid, and went forth and told the priests. 3 And they took courage and went in and saw that which was done: and the panels of the temple did wail: and they rent their clothes from the top to the bottom. And his body they found not, but his blood they found turned into stone. And they feared, and went forth and told all the people that Zacharias was slain. And all tile tribes of the people heard it, and they mourned for him and lamented him three days and three nights. And after the three days the priests took counsel whom they should set in his stead: and the lot came up upon Symeon. Now he it was which was warned by the Holy Ghost that he should not see death until he should see the Christ in the flesh.

XXV. 1 Now I, James, which wrote this history in Jerusalem, when there arose a tumult when Herod died, withdrew myself into the wilderness until the tumult ceased in Jerusalem. Glorifying the Lord God which gave me the gift, and the wisdom to write this history. 2 And grace shall be with those that fear our Lord Jesus Christ: to whom be glory for ever and ever. Amen.

B. Sample GZ dissertation page scan

χακεL επηΕε την σκηνην αυτου

411 412 612

χαχεL επηΕεν την σκηνην αυτου 627

χακη επηΕεν την σκηνην αυτου 614

χαι επηΕε την σκηνην αυτου εχει 219 220 512 513 615 616 618 619
622 702 703 707 708 801-803 805 902

XaL επηΕε την χηνην αυτου εχεL 617

χαι επηΕεν την σκηνην αυτου εχεL 218 309 511

XaL επLΕεν την σχηνην αυτου εχεL 217 514 517 709

xaL επηΕεν εχεL την σκηνην αυτου 510 705

XaL επηΕεν εχεL την σκηνην [[χν]] (χν IN RAS.) αυτου 901

XaL επηΕεν εχει σκηνην 626

XaL επηΕεν εαυτω σχηνην εχεL 625

XaL επηΕεν εαυτον σχηνην εχεL 628

XaL ουχ επηΕε την σκηνην αυτου εχεL 623

C. Scan of page of Bodmer V manuscript

D. Data analysis results (detailed)

D. R Code

```
library(vegan)
library(cluster)
rlabels <- scan('fullSetManuscriptNameVector.txt',what=character(),sep=',',nlines=1)
clabels <- scan('fullSetFeatureVector.txt',what=character(),sep=',',nlines=1,encoding='UTF-8')
fullSet <- read.csv('fullSetIDFFeatureMatrix.txt', header=FALSE)
fullSet.cosine <- read.csv('fullSetCosineMatrix.txt', header=FALSE)
rownames(fullSet) <- rlabels
colnames(fullSet) <- clabels
rownames(fullSet.cosine) <- rlabels
colnames(fullSet.cosine) <- rlabels
fullSet.gc <- read.csv('c:/users/tmwsiy/workspace/greektext/output/fullSetGlobalCounts.txt', header=FALSE,encoding='UTF-8')
colnames(fullSet.gc) <- c('gram','count')
fullSet.idfgc <- read.csv('c:/users/tmwsiy/workspace/greektext/output/fullSetGlobalIDFCounts.txt', header=FALSE,encoding='UTF-8')
colnames(fullSet.idfgc) <- c('gram','count')

# K-means
out = pam(fullSet.cosine,center=%numCenters%,nstart=10000);
out$cluster;
out$withinss;

# Sørensen similarity index
fullSet.dist <- vegdist(fullSet)
# multidimensional scaling
fullSet.mds0 <- monoMDS(fullSet.dist)
fullSet.mds  <- metaMDS(fullSet)
# canonical correspondence analysis
fullSet.cca <- cca(fullSet)


```

Figures

Figure 1

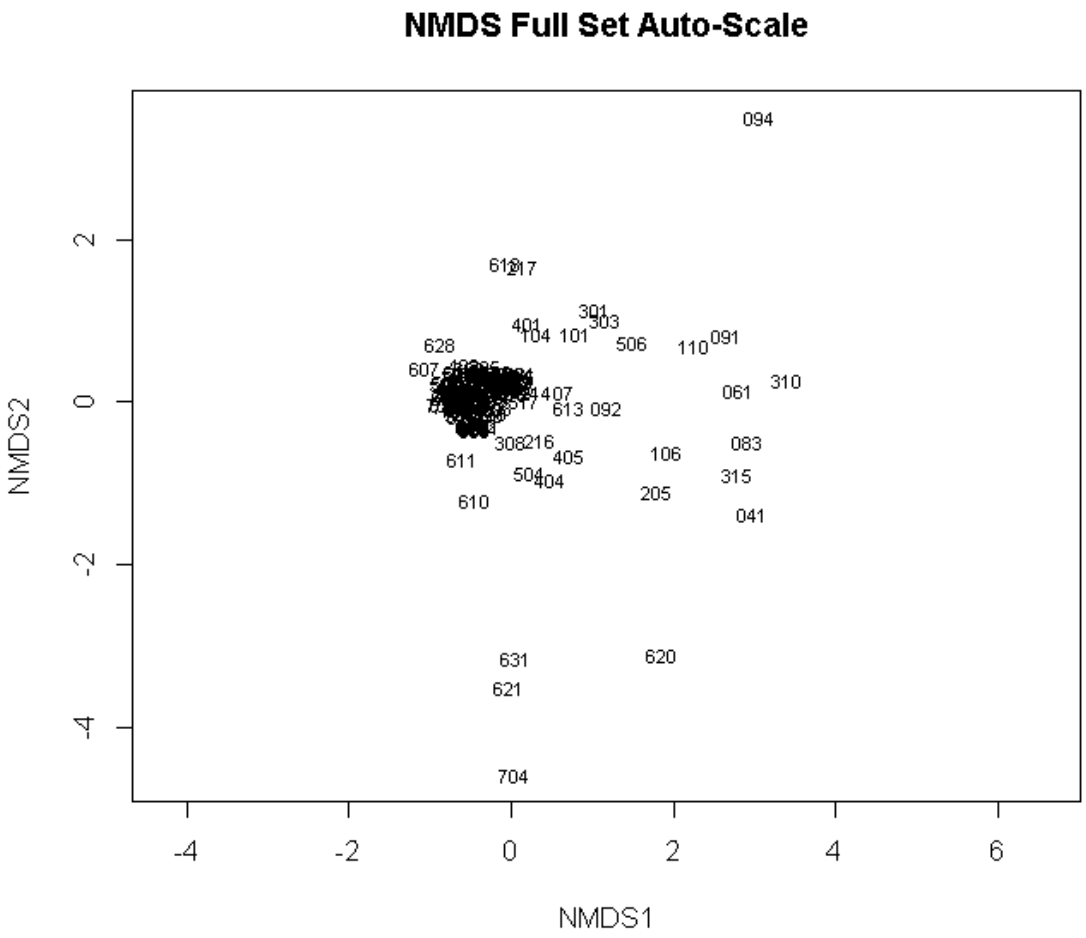


Figure 2

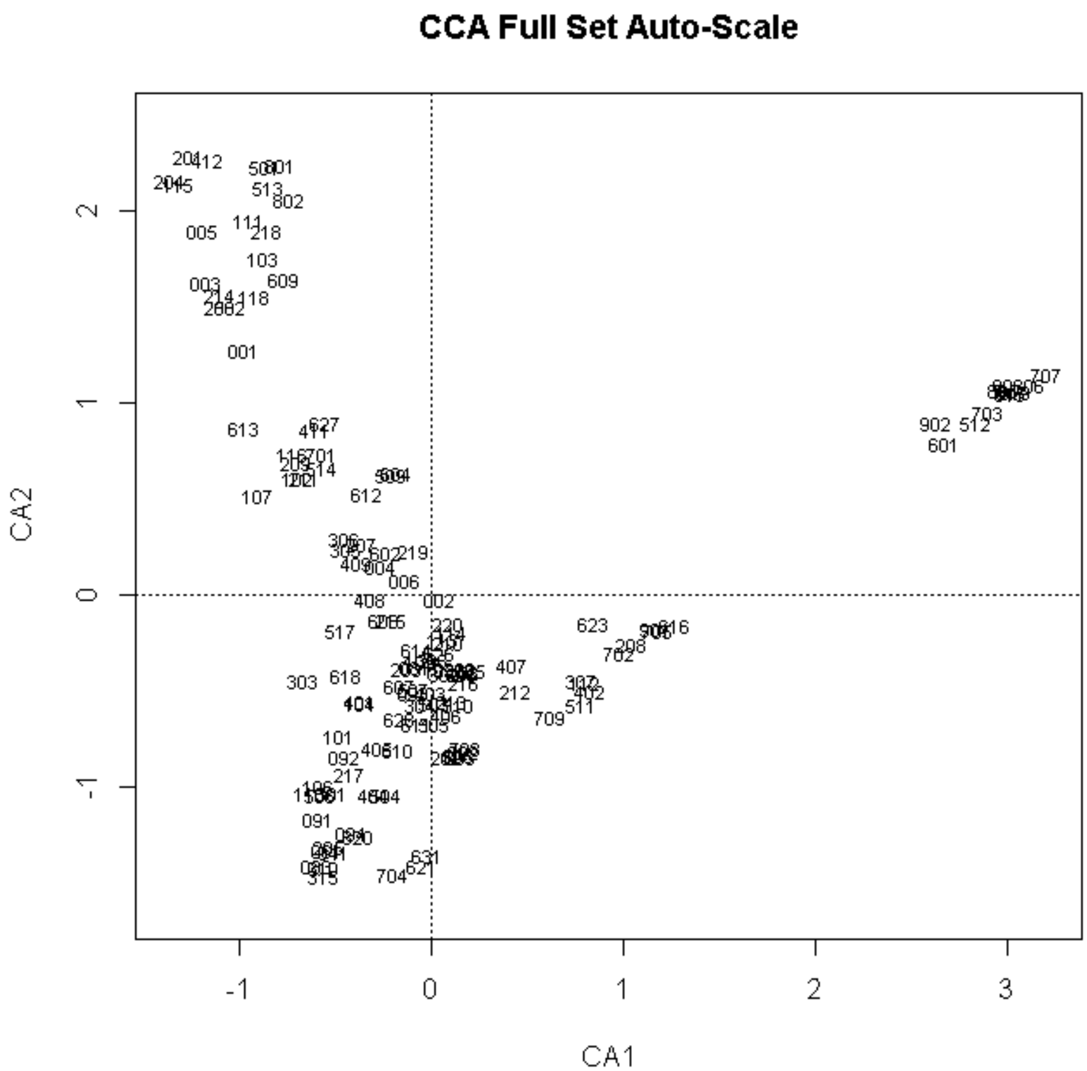


Figure 3

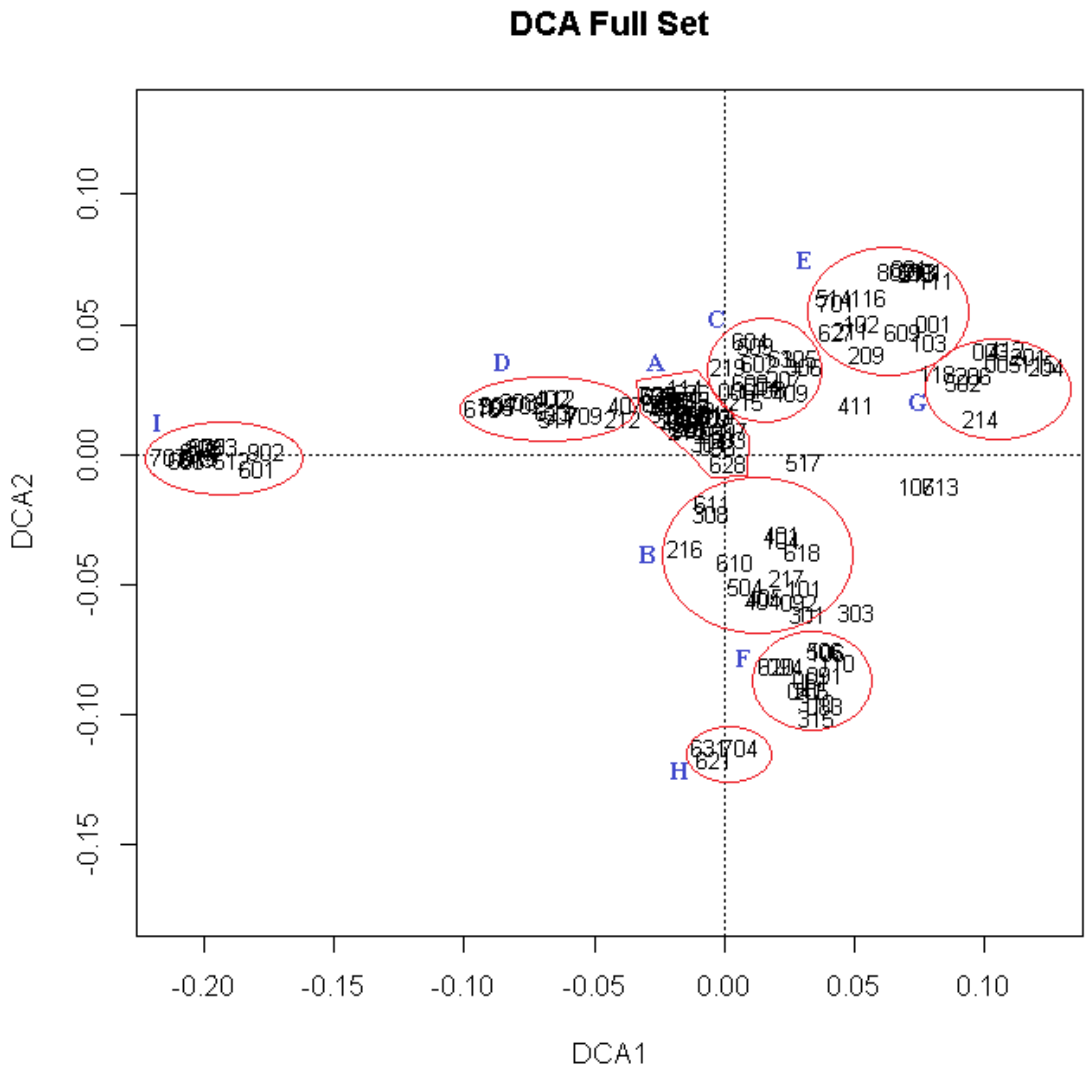


Figure 4

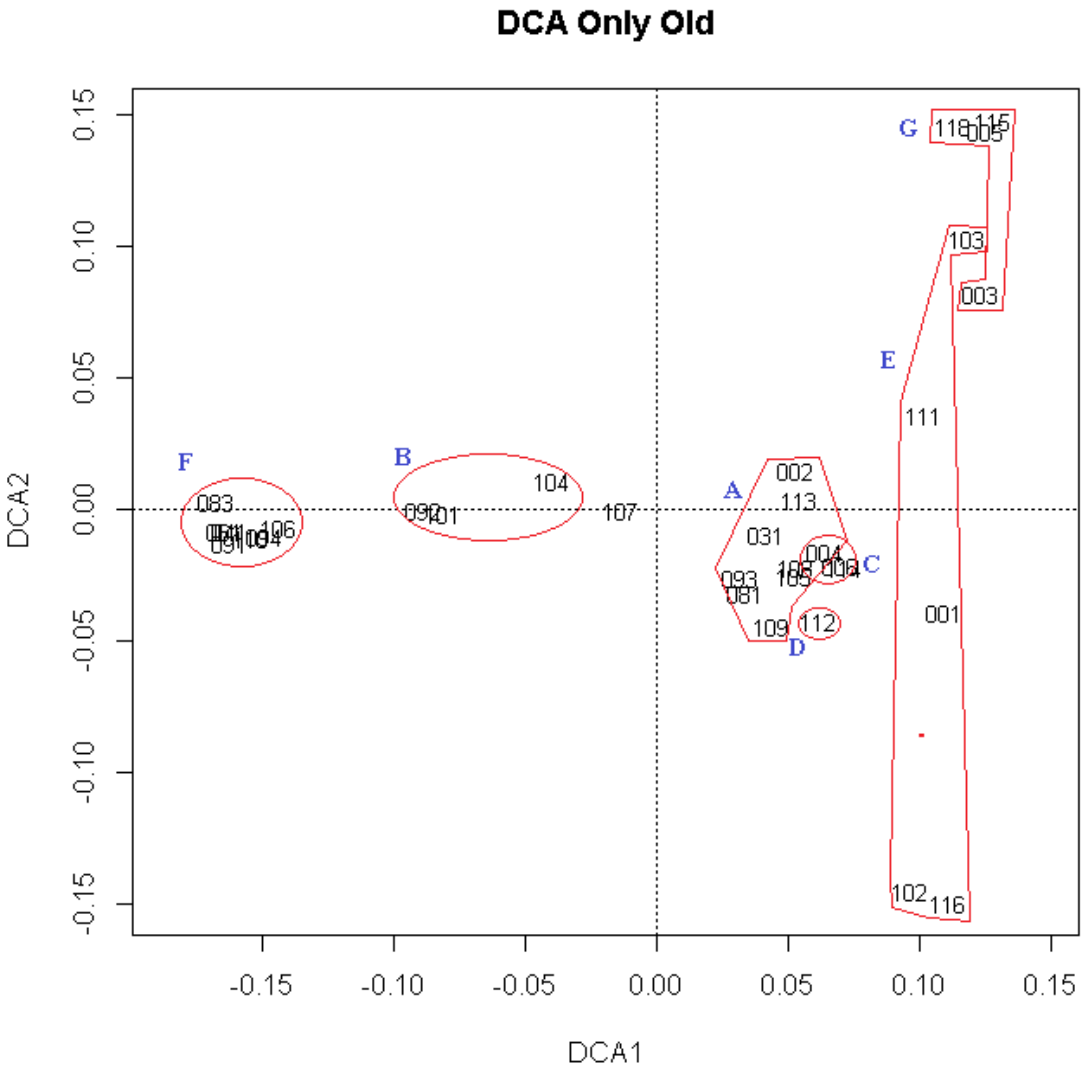

DCA Only Old
DCA2
DCA1
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
A
B
C
D
E
F
G
002
113
031
004
093
081
109
112
104
107
083
111
103
003
001
102
116
118
115
005

Figure 5

	A	B	C	D	E	F	G	H	I	J
1	m002	m092	m004	m112	m001	m041	m003	m621	m512	m107
2	m031	m101	m006	m208	m102	m061	m005	m631	m601	m303
3	m081	m104	m207	m212	m103	m083	m115	m704	m606	m411
4	m093	m216	m215	m307	m111	m091	m118		m615	m517
5	m105	m217	m219	m402	m116	m094	m201		m617	m613
6	m108	m301	m305	m407	m209	m106	m204		m619	
7	m109	m308	m306	m511	m211	m110	m206		m703	
8	m113	m401	m408	m616	m218	m205	m214		m707	
9	m114	m404	m409	m623	m501	m310	m412		m803	
10	m202	m405	m509	m702	m513	m315	m502		m805	
11	m203	m504	m602	m705	m514	m506			m902	
12	m210	m610	m604	m709	m609	m620				
13	m213	m611	m608	m901	m627					
14	m220	m618	m612		m701					
15	m302				m801					
16	m304				m802					
17	m309									
18	m403									
19	m406									
20	m503									
21	m505									
22	m507									
23	m508									
24	m510									
25	m603									
26	m605									
27	m607									
28	m614									
29	m622									
30	m625									
31	m626									
32	m628									
33	m708									

Figure 6

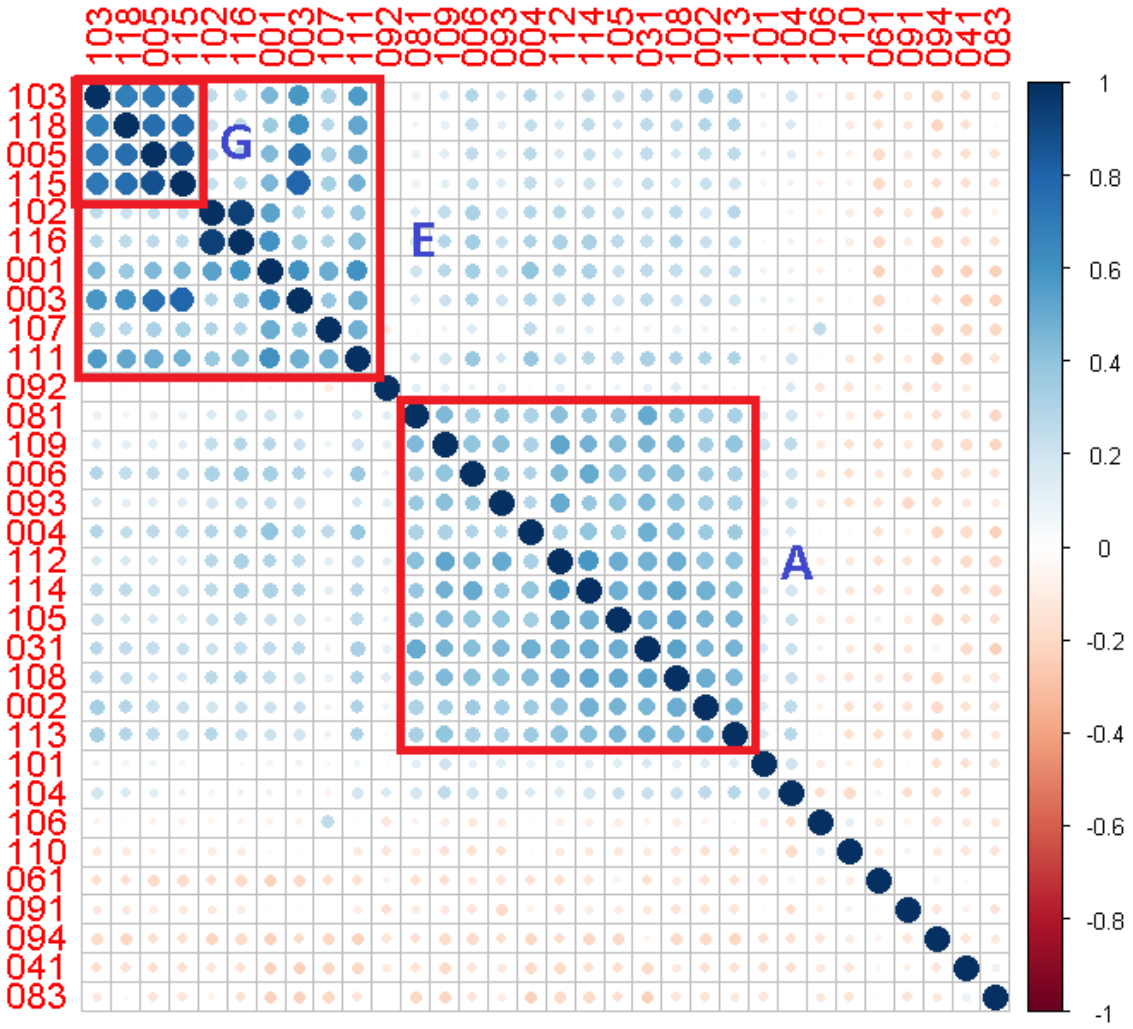

103
118
005
115
102
116
001
003
107
111
092
081
109
006
093
004
112
114
105
031
108
002
113
101
104
106
110
061
091
094
041
083
G
E
A
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Figure 7

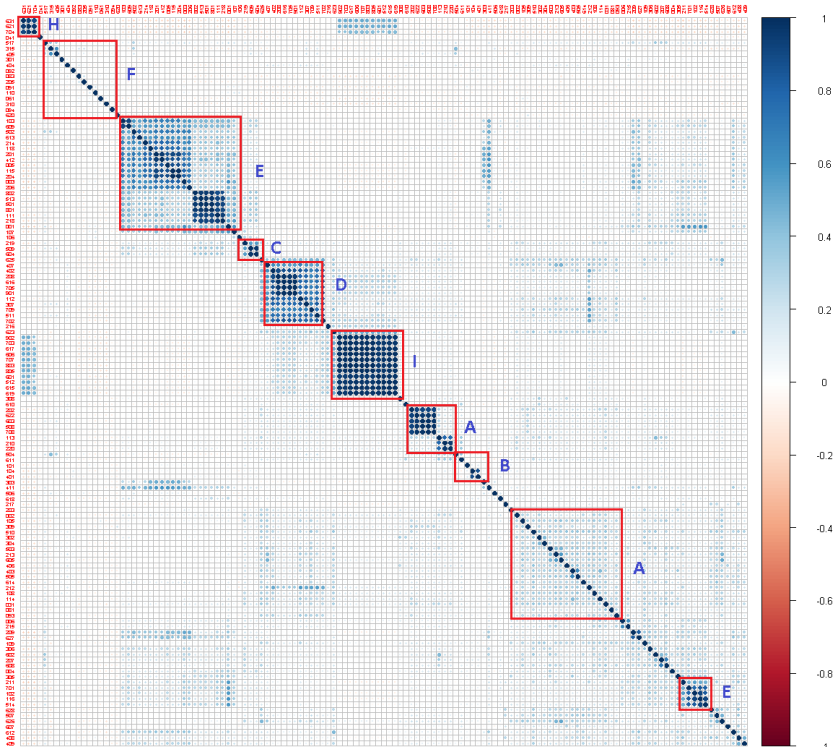


Figure 8

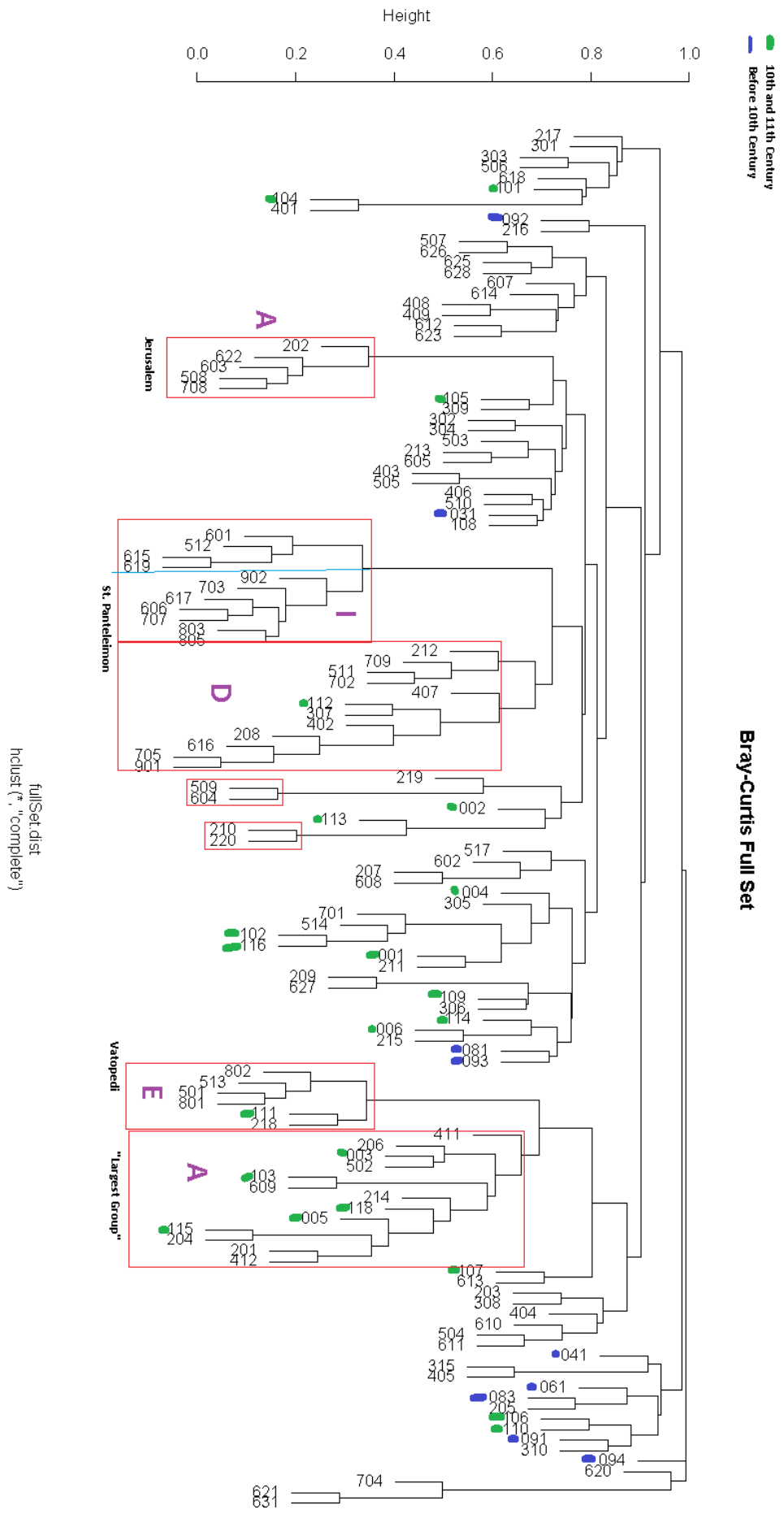
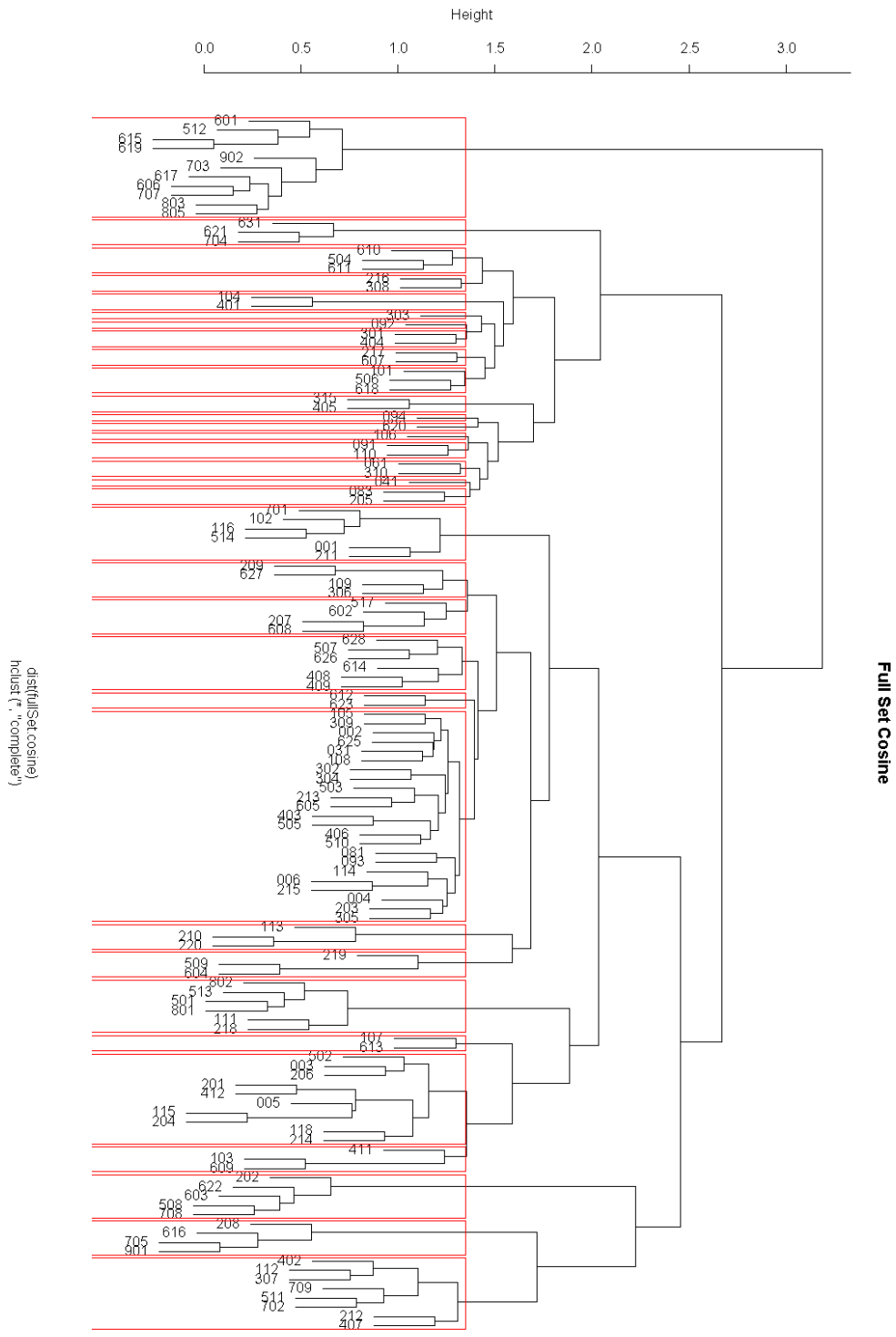

Height
0.0
0.2
0.4
0.6
0.8
1.0
10th and 11th Century
Before 10th Century
Bray-Curtis Full Set
fullSet.dist
hclust (*, "complete")
Jerusalem
St. Panteleimon
Vatopedi
"Largest Group"
A
D
I
E
A
217
301
303
506
618
101
104
401
092
216
507
626
625
628
607
614
408
409
612
623
202
622
603
508
708
105
309
302
304
503
213
605
403
505
406
510
031
108
601
512
615
619
902
703
617
606
707
803
805
212
709
511
702
407
112
307
402
208
616
705
901
509
604
219
002
210
220
113
517
602
207
608
004
305
701
514
102
116
001
211
209
627
109
306
114
006
215
081
093
802
513
501
801
111
218
411
206
003
502
103
609
214
118
005
115
204
201
412
107
613
703
308
404
610
504
611
041
315
405
061
083
205
106
110
091
310
094
620
704
621
631

Figure 6


Height
0.0
0.5
1.0
1.5
2.0
2.5
3.0
dist(fullSet cosine)
hclust (*, "complete")
Full Set Cosine