Metre as a stylometric feature in Latin hexameter poetry

Ben Nagy

`benjamin.nagy@adelaide.edu.au`

August 6, 2019

I demonstrate that metre is a privileged indicator of authorial style in classical Latin hexameter poetry. Using only metrical features, I performed pairwise classification experiments between 5 first-century authors (10 comparisons) using four different machine-learning models. The results showed a two-label classification accuracy of at least 95% with samples as small as ten lines and no greater than eighty lines (up to around 500 words). These sample sizes are an order of magnitude smaller than those recommended for 'standard' BOW or n-gram approaches, and the reported accuracy is outstanding. Additionally, I explore the potential for novelty (forgery) detection, or 'one-class classification'. I analyse the disputed Aldine Additamentum (Sil. Ital. *Puni.* 8:144-225) and conclude (p=0.0012) that the metrical style differs significantly from that of the rest of the poem.

Keywords: Latin; digital humanities; machine learning; poetry; stylometry; authorship attribution

1 INTRODUCTION

While authorship attribution studies, and stylometry more generally, are becoming much more common, the bulk of the research has been applied to prose. In terms of work that considers poetry *qua* poetry, the pickings are rather slim. Kao and Jurafsky (2012) attempt to computationally assess aesthetics. Omer and Oakes (2017) apply the Arabic science of poetic metre (*Arud*) to authorship attribution, but the texts they analyse are all prose. Historically, some significant work has been done on Shakespeare but, for example, the recent computational results of Ilsemann (2018) have been sharply criticised by Barber (2019) as a misapplication of Burrows' Delta (Burrows (2002)). Tizhoosh, Sahba, and Dara (2008) and Chaudhuri, Dasgupta, Dexter, and Iyer (2018) both set about distinguishing poetry from prose, but offer no analysis of the poetry itself. One notable exception is Neidorf, Krieger, Yakubek, Chaudhuri, and Dexter (2019), which sets out to demonstrate the single-authorship of Beowulf

using a range of techniques, including a consideration of metre, albeit brief. This general lacuna has been noted by Plecháč, Bobenhausen, and Hammerich (2019), who have begun to treat metrical features as a first-class citizen of the stylometry universe, and are some of the few researchers applying the most recent multivariate and machine-learning techniques. One final paper that deserves attention is Forstall, Jacobson, and Scheirer (2011), which examines metre, specifically in Latin poetry. Unfortunately, the methods used by this team treated syllable quantities in Latin simplistically as an n-gram (a string of n symbols, where each symbol may only be ◡ or —), and they were unable to achieve useful separation between authors on this basis. The reason that poetry-specific approaches are important is that the best-known attribution methods for prose,[1] as aptly pointed out by Eder (2015), require fairly large sample sizes, which are often not available for verse.[2] Classical Latin poetry does not rhyme, so one of the key aims of this project, therefore, was to consider metrical features. Latin hexameter offers a particularly rich interplay of features, and it was hypothesised that the density of stylistic information contained in the poems' metre would allow for effective analysis with much less text.

2 RESEARCH QUESTIONS

I address two research questions in relation to the analysis of Latin hexameter poetry. The first, and most fundamental is this: "Are the metrical attributes of Latin hexameter poetry an effective feature for stylometric analysis"? I demonstrate, via two-label classification, that the answer is emphatically yes. As an application of these techniques, I then applied metrical analysis to the disputed authorship of the Aldine additamentum (Sil. Ital. *Puni.* 8:144-225), seeking to re-address the question: "Is this passage genuine?".

3 METHODS

The website Pedecerto (Colombi et al. (2007)) has, for some time, offered on-demand, and highly accurate, scansion, particularly optimised for hexameters and elegaic couplets. In partnership with the site Musisque Deoque (Mastandrea et al. (2007)), they also maintain an extensive digital library of Latin poetry. This scansion is presented via their webpage as an image, showing the syllable quantities and other metrical information (caesurae, elision, hiatus, diaresis etc). In March, 2019, however, in response to my enquiry, their technical team added a new feature which allowed the download of any of the MQDQ poems in XML format, including the full scansion.[3] This offers a valuable resource to Digital Humanities practitioners, and made much of this present work possible. In Fig. 1, we can see the wealth of information available for each hexameter line. As well as the pattern of dactyls and spondees in the first four feet

---

1 A good survey is available in Jockers and Witten (2010). The present consensus seems to favour analysis based on frequency of closed-class words (eg function words) or character n-grams. However Bag of Words (BOW) approaches are still popular, in particular as implemented in the R stylo `rolling_delta()` method (Eder, Rybicki, and Kestemont (2016)).

2 Eder (2015) suggests a 5000 word minimum for authorship studies on prose. Later (Eder, 2017, 4) he revises this to "as little as 2000 words in many cases", which is still a great deal too many to be viable for many questions concerning poetry.

3 http://www.pedecerto.eu/pagine/autori, Retrieved August 6, 2019

Figure 1: Raw XML output from MQDQ

```
<line name="952" metre="H" pattern="DDDS">
    <word sy="1A1b1c" wb="DI">Vitaque</word>
    <word sy="2A" wb="CM">cum</word>
    <word sy="2b2c3A" wb="CM">gemitu</word>
    <word sy="3b3c" wb="DI">fugit</word>
    <word sy="4A4T5A5b" wb="CF">indignata</word>
    <word sy="5c" wb="DI">sub</word>
    <word sy="6A6X">umbras.</word>
</line>
```

(DDDS) we can see the caesurae (CM or CF for strong and weak, respectively) and, most importantly, the full scansion of every syllable. Each syllable in the line is marked by two symbols, showing the foot in which it falls and a letter representing the position in that foot—so, for example, 1A1b1c corresponding to *Vitaque* describes a three syllable word, with the arsis of the first foot followed by two breves.

3.1 *Feature Selection*

There are four general types of features that we have been able to extract from a single line of hexameter. The first, and most often studied, is the metrical pattern in the first four feet. A hexameter line is constructed from six feet, each of which can be either a spondee (— —) or a dactyl (— ⏑ ⏑). However by convention, the final two feet are almost always dactyl, spondee (— ⏑ ⏑ | — —).[4] This yields sixteen possible foot combinations. These combinations have been extensively studied by Classicists, the 'recent' standard works being Platnauer (1951) and a string of publications by Duckworth in the 1960s.[5] The next group of features are the *caesurae*. A caesura occurs whenever there is a word break in the middle of a foot. Caesurae can, thus, occur after the initial spondee, which I will refer to as a *strong caesura* or after the first breve in a dactyl, a *weak caesura*.[6] By convention, although any foot may (technically) contain a caesura, attention is restricted to the ones in the second, third and fourth feet. I should note that, in traditional analysis, a line of hexameter is considered to have a *principal* caesura, which entails a pause; often a sense-pause, although sometimes simply a pause for breath. The present analysis makes no attempt to determine which, if any, of the available caesurae are 'principal'. It is important to understand, however, that the general 'rule' was to place a strong caesura in the third foot, although every author deviated from this rule to a certain extent as a matter of stylistic preference. The third class of features requires

---

4 A spondee may appear in the fifth foot, but these are very rare. There are twenty-three such lines in the *Aeneid* of Vergil, comprising around 0.2%

5 cf, at the least, Duckworth 1965, 1967, 1969a 1969b. There are probably others.

6 An interested reader will find a range of other names in the literature; 'masculine' vs 'feminine', or one of several technical terms, derived from the Greek, that encode both the strength and the position—*pentethemimeral, hepthemimeral* etc.

a little explanation. The first syllable of each of the six feet is called an *ictus syllable*. This term, although Latin (stroke, thrust, blow) derives from ancient Greek hexameter. Due to the language difference, however, it was not the ictus syllables in Latin verse that carried the primary stress—instead the Latin words were stressed (or accented) as usual.[7] This interplay between Latin word accent and metrical ictus creates an inherent *conflict*, which formed an important part of the style of Latin hexameter. Perhaps counter-intuitively, the general trend was to create conflict throughout the first four feet (otherwise, presumably, the line sounded too 'sing-songy'), and in almost all cases the conflict resolves in the two final feet, where the ictus and the accent correspond. Once again, this has been well-studied in the traditional literature: consult in particular Knight (1931) and Duckworth (1969b, 17-24). In the following discussion I adopt the terminology of these studies, and refer to a foot in which the ictus and accent coincide as *homodyne*.[8] Finally, we measured *elision*. When Latin words ending in vowels or nasals precede words that begin with a vowel, an elision occurs, and the final syllable of the first word is not counted for metrical purposes. Opinions differ as to whether elided syllables are completely dropped when the poem is read aloud, whether they are slurred or perhaps lightly pronounced—once again, this is irrelevant for my purposes. For each line we simply recorded an elision count per line (not including *prodelision*, a slightly different feature)—our only non-binary indicator. Somewhat surprisingly, this feature is arguably the strongest single metrical classifier (Table 2).

While the metrical patterns and caesurae are trivial to extract from the MQDQ XML data, determining the ictus/accent conflicts required considerably more work. The general rule for word-accent in Latin is very simple, but there are a surprising number of exceptions and edge cases. In the end, it was necessary to develop new code to perform this analysis, which is included in the project github repository [CITE]. Whenever new analysis tools are developed, care must be taken to ensure that the results are accurate—it is useless to base statistical analysis on bad data. There are at least two points of comparison. The first is Knight's *Homodyne in the fourth foot of the Vergilian Hexameter*. Knight informs us that "[f]ourth-foot homodyne is by comparison rare in Vergil, but not very rare. The percentage of it to the total number of lines in the *Aeneid* is 35.95 per cent" (Knight, 1931, 186). Using my algorithm, fourth-foot homodyne was detected in 3507 lines (of 9840 in my edition), which is 35.65%. The difference in our percentages represents perhaps 30 lines on which we disagree. Knight also lists some percentages by individual books. For *Aen.* 1, he has 28.61% while I have 28.82%; *Aen.* 8, 39.97% to my 39.29%; *Aen.* 10, 32.59% vs 32.71%. While Knight does not share with us his methodology for scanning and tabulating the entire *Aeneid* in the 1930s (perhaps it was a long winter), he does mention that he performed "two computations, in which different texts were used and different conventions followed" and that his figures varied by only 0.45% (Knight, 1931, 186 n. 1). The second source for independently determined statistics of this kind is Duckworth (1969b, 19-20, plus final Table). Duckworth consistently

---

7 It is possible that the ictus syllables in Latin poetry were lightly accented. This question is immaterial in the present context.

8 Knight and Duckworth both focus exclusively on conflict in the fourth foot. As will be seen later, conflict preferences in *any* foot are important stylistic markers. Based on this analysis, conflict/harmony in the first foot (not the fourth) is actually the best place to look to distinguish different authors (see Table 1).

finds texts to have a slightly higher homodyne percentage than does Knight. Thus, for the entire *Aeneid*, Duckworth records 37.78% fourth-foot homodyne. This discrepancy is essentially due to disagreement among scholars regarding the correct accent of a few two syllable words (*ergo, illuc, . . .*), and whether certain conjunctions (*et, at, aut, sed, . . .*) bear any accent at all. Based on my research,[9] I have accepted the views of Allen (1965, 88-9) (which apparently accord closely with Knight's), but this would be a poor venue in which to set out the entire debate. In any event, it can be hoped that as long as the analysis is being applied consistently to all texts that the statistical effect of any systemic errors will be negligible.

3.2 *Two-Label Classifiers*

Before attempting to use metrical features for any sort of authorship attribution, it should first be established that these features are able to distinguish authors at all. It would be foolish to rely on a measure that cannot do what an expert reader can do by instinct—to tell the difference between the hexameters of, say, Vergil and those of Ovid. To this end, I collected and analysed Vergil's *Aeneid*; Ovid, *Metamorphoses*; Statius, *Thebaid*; Lucan, *Pharsalia* and Silius' *Punica*. These works are all composed in dactylic hexameters, and written as 'epics'. Each pair was then analysed as a two-label classification problem with four algorithms: Extremely Random Trees, Gaussian Naïve Bayes, Logistic Regression and Support Vector Machines. All tests were carried out using the Python package `scikit-learn` (Pedregosa et al. (2011)), and replication code is available in the github repository for this paper (Nagy 2019). The tests were conducted with an 80/20 train-test split, with a chunk-size of 81 lines (to model the later authorship problem). In every case except Vergil vs Silius and Vergil vs Statius the accuracies were virtually 100%. Because the most 'difficult' classification is Silius vs Vergil, the rest of this section is mostly concerned with that pair. In Fig. 2, we can see that analsying foot patterns (`Feet Only`) is only slightly more effective than considering nothing but the patterns of ictus/accent conflict (`Conflict Only`)—this underscores the value of conflict as an indicator of style. In fact, in terms of foot patterns, the importance of the first and fourth feet (dactyl or spondee) far outweighs the other two. It can be seen from Fig. 2 that the set `Caes+Confl+FirstFoot` (Caesurae, Conflict and First Foot) performs almost as well as the entire feature set.

Another benefit of working with pairwise classification is that it can provide some kind of intuition as to which metrical features are the most useful to consider when distinguishing between authors. Not all machine learning algorithms can produce understandable feature importances, but the Random Forest family of algorithms (we used `ExtraTrees` from scikit-learn) and Support Vector Machines (with a linear kernel) will. We compared each pairwise combination of the five works above (10 total comparisons) and averaged the feature importances according to `SVM` and `ExtraTrees`, with the results shown in Table 1. It is important to note that the two sets of values listed cannot be compared with each other, and it is not clear how the weights compare within

---

9 The best collection of the primary sources with respect to accent is Schoell (1876), although the truly dedicated may prefer to read the works unabridged in Keil (1857)

Figure 2: Effectiveness of Different Feature Subsets
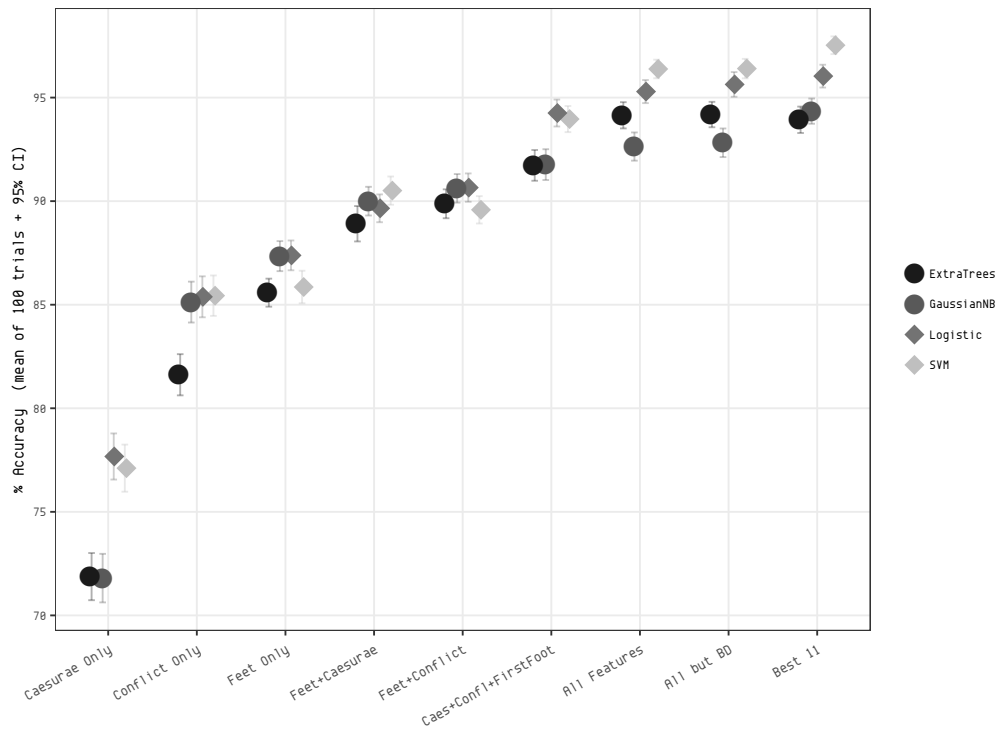
a set (ie a feature with a weight of ten may not be ten times as important as one with a weight of one). In terms of raw ranking, though, they provide an interesting guide.

3.3 *Chunking*

After each line is converted to a feature vector, it can be considered as a single observation. However, because of the inherent variability of the lines, the two-label classifiers were all found to perform poorly on the raw data, with accuracies under 60% (not much better than chance). This seems to be intuitively reasonable—a human reader would also find it more difficult to determine the author of a single line than to examine a hundred lines at once. The feature observations were, therefore, 'chunked', by taking a set of lines and considering the set to be a single observation, with the mean of the feature values; a set of twenty lines of which 15 had a strong third-foot caesura would have an F3SC value of 0.75. The trade-off, of course, is that the number of observations is reduced; 10,000 lines of the Aeneid with a chunk size of 20 becomes just 500 observations. The question then becomes 'how many lines do we need in a chunk before we can make reliable decisions?'. The answer, of course, is 'it depends'. Comparing Silius to Ovid, a 10-line chunk is enough to yield more than 90% accuracy (see Fig. 3). Silius and Vergil are the most similar,[10] so by considering four different classification algorithms (Fig. 4) it seems reasonable

---

10 Silius, it seems, admired the work of Vergil with almost religious fervour, cf *Plin. Ep.* 3.7, 8

Table 1: Ranked Feature Importances, as determined by `ExtraTrees` and `SVM`

ExtraTrees		SVM	
Feature	Score	Feature	Score
SYN	24.08	F1S	8.25
F1S	18.48	SYN	7.60
F4S	8.93	F3WC	5.29
F3WC	7.27	F4S	4.71
F1C	5.85	F4SC	4.02
F3SC	5.33	F3S	3.13
F3C	5.09	F3C	3.09
F3S	3.85	F3SC	2.89
F4C	3.75	F2C	2.72
F2C	3.30	F1C	2.18
F4SC	3.18	BD	2.13
F2WC	2.47	F4C	2.13
BD	1.97	F2SC	2.09
F2S	1.91	F4WC	1.89
F4WC	1.82	F2S	1.87
F2SC	1.62	F2WC	1.75

to expect that anything over 60 lines should yield sensible results. It should also be noted that the order of the lines in the original works were randomised before being chunked. The intention of the randomisation process was to normalise 'non-atmospheric' variation. Latin poets use the metre of their verses for poetic effect, and so Vergil will use different patterns in a battle-scene than he does when describing funeral games—and yet (by hypothesis) his style is always 'Vergilian'. By mixing the battles with the games, the love scenes with the death scenes, these surface variations are evened out, making it easier to see the deeper authorial style.

It is here that the low-dimensionality of our feature set works to our advantage. Instead of considering hundreds of words or character n-grams, we have just 16 metrical features. When the Aldine Additamentum is taken as a single observation, and a chunk size of 81 lines is used for Silius' *Punica* (the same length as the Additamentum) we have 150 observations. For algorithms that are sensitive to dimensionality it would be better to have more than 256 (16^2), but the dimensionality does not seem unreasonable. If a smaller set of features is required, experiments indicate that standard feature selection techniques can be used to trim the feature dimension down to 11 with no visible loss of classification accuracy (Fig. 5). Another viable approach, which we will see in the following section, is to create more training samples using repeated random sampling from the source lines.

3.4 *Detecting 'Scribal Interpolation' in Larger Works*

Having established that metrical features seem to be reliable indicators of authorial style, it now seems reasonable to apply them to an authorship problem.

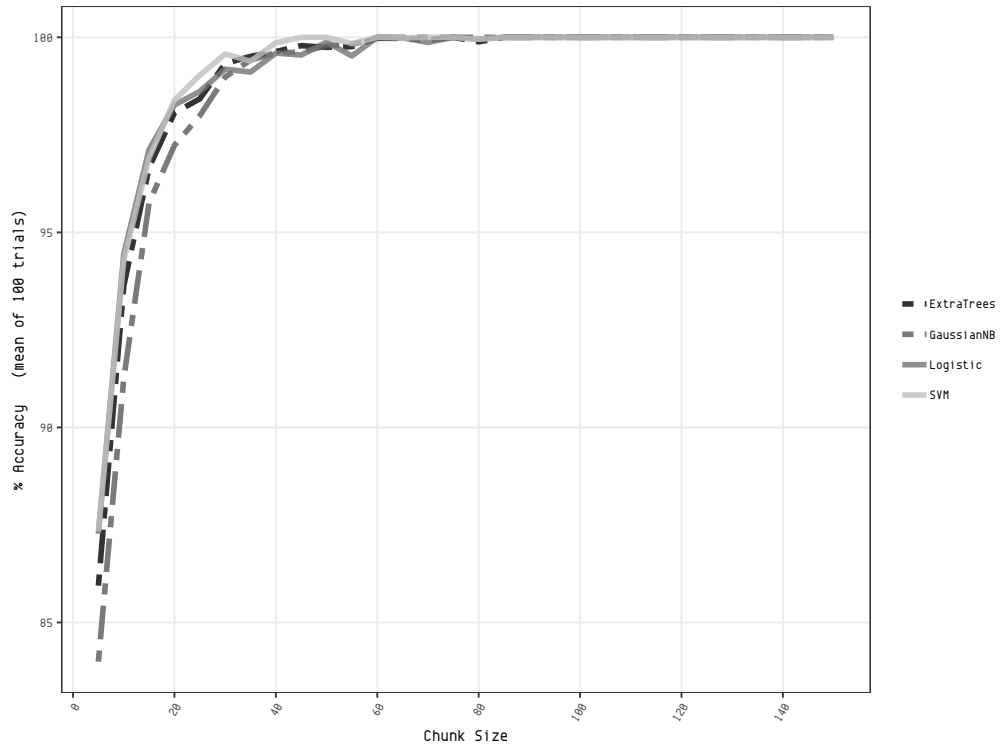Figure 3: Classification Accuracy by Chunk Size: Silius vs Ovid

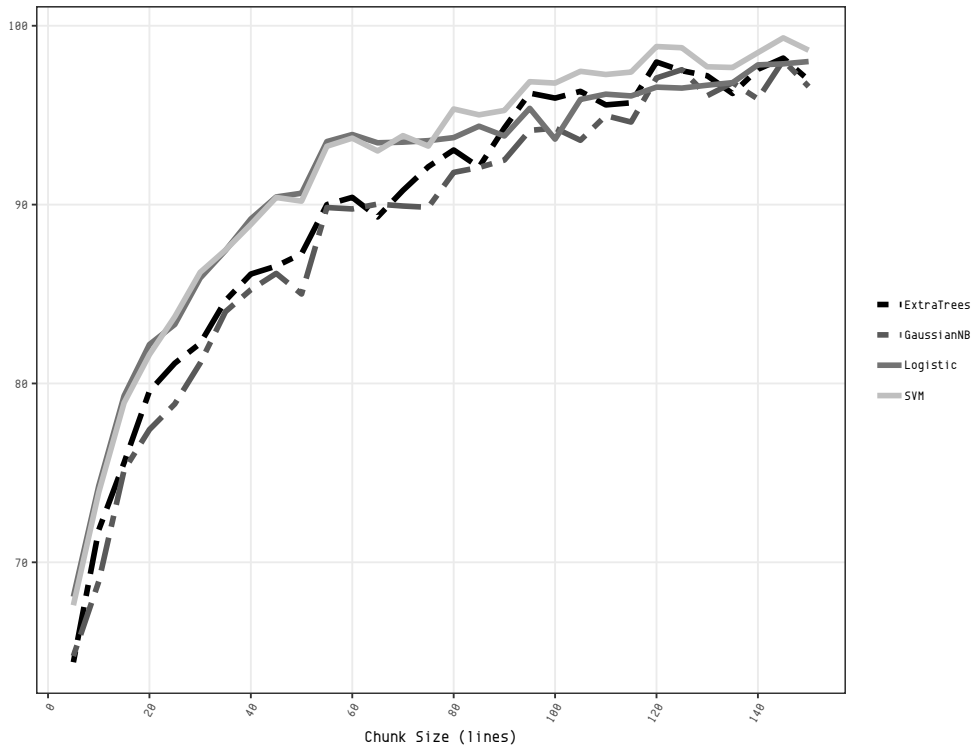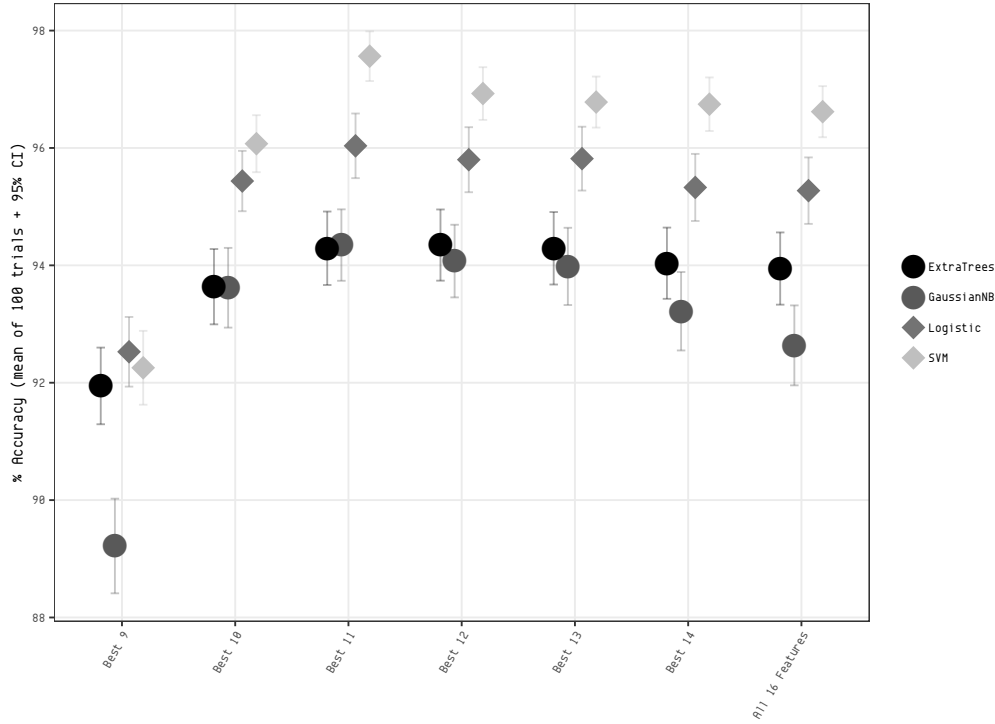Figure 4: Classification Accuracy by Chunk Size: Silius vs Vergil

Figure 5: Parsimonious Models: Selecting a Feature Subset (results based on Silius vs Vergil, 81-line chunks)

As mentioned, an interesting question is the authorship of the Aldine Additamentum.[11] At just over 500 words, the disputed text is too short to be a useful candidate for BOW or function-word approaches[12] but, at 81 lines, it fits comfortably in the range where all of the two-author classification experiments achieved over 95% accuracy. Broadly speaking, this problem fits best under the heading of 'novelty detection'. There are a variety of appropriate algorithms that we could have selected, including `OneClassSVM` or `IsolationForest` but the simplest and most explanatory seemed to be a multi-dimensional cluster distance. The general approach, therefore, is this: we arrange the samples in a 16-dimensional feature space, choose a suitable distance metric and assume some kind of Gaussian distribution. Then, by calculating the distance of each point from the central tendency, we can assume the distances to be distributed according to χ^2, and hence calculate a p-value. Because of the nature of the chunking, the features are already scaled in the range [0,1] (they are the mean of binary indicators), however as can be seen from Table 2, there are some problems with co-variance and correlation. Accordingly, the most sensible distance was seen to be the Mahalanobis distance, which 'normalises' the features by 'dividing' by the co-variance matrix. In mathematical terms, the

---

11 For an excellent summary of background to this issue I refer the reader to Lee (2017, 20-9), which includes a valiant attempt at univariate metrical analysis, as well as a survey of the existing opinions in secondary literature (Appendix 1, p. 144).

12 As mentioned, this question of "how much text is enough?" is explored effectively in Eder (2015) and Eder (2017)

Table 2: Some highly correlated features in Silius' *Punica*

Feature 1	Feature 2	Pearson's r
F3S	F3WC	-0.518
F2C	F2WC	-0.681
F3C	F3SC	0.816
F3C	F3WC	-0.855
F4C	F4SC	0.875
F3SC	F3WC	-0.802

Mahalanobis distance of an observation x from the sample mean μ is defined to be $\sqrt{(x-\mu)^\mathsf{T} C^{-1}(x-\mu)}$ (where C is the covariance matrix of the target distribution $\mathbf{X}$). In the following discussion, we will mainly see M^2, the square of the Mahalanobis distance, because it can be proven that M^2 follows the χ^2 distribution, allowing for the calculation of a p-value. For this experiment, the target distribution was created by repeated random sampling.[13] The lines of the Additamentum were removed from the corpus, and then 10,000 sets of 81 lines were randomly selected with replacement. Based on experiment, this sample is large enough that there was negligible variation in M^2 due to the random process. The M^2 distance of the Additamentum from the target, 36.82, yields a corresponding p-value of 0.0013, low enough to be statistically significant. These results can be fully reproduced with the code at [CITE]. In an attempt to assess the reasonability of the final result, I made two other comparisons. When the Additamentum was compared to Vergil's *Aeneid*, the corresponding p-value was 0.0006. Next, I compared the mean vector (or *centroid*), representing an 'average' chunk of Vergil, to Silius, and found a much lower distance of 15.88 with a p-value of 0.39 (recall, though, that we can tell Vergil from Silius with more than 95% accuracy). In other words, the metrical style of the Additamentum differs much more from either Vergil or Silius than they differ from each other. These results are summarised in Table 3. On face value, these results support the theory that the Aldine Additamentum was not written by Silius—with the most likely alternative being that it was introduced by a humanist scholar in C15–16 to resolve a lengthy manuscript lacuna. An additional advantage of the Mahalanobis distance[14] is that the distance can be easily decomposed to show a per-feature contribution to the total distance.[15] This *feature contribution vector* seems to have good explanatory power, which can be extremely helpful in verifying results. For example, suppose that the contribution vector for a passage with a high M^2 distance indicates an unusual

---

13 An astute reader may wonder why, if repeated random sampling is a reasonable approach, it was not used during our classification experiments to offset the loss of sample size caused by line chunking. The problem is that if we had trained on a repeated random sample and then performed classification on 'real' (sequential) chunks, much of the training data would re-appear (although in random places) in the test data. This would artificially inflate our accuracy and inevitably lead to over-fitting.

14 This may be a novel observation; certainly I have not come across it elsewhere in the literature.

15 The core of the Mahalanobis distance is $(x-\mu)^\mathsf{T} C^{-1}(x-\mu)$, and it is easy to see that this reduces to a vector dot product ($x^\mathsf{T} y = k \in \mathbb{R}$). All that needs to be done is to multiply xy point-wise and record the product vector. That vector is the feature contribution vector, and the sum of its entries is the squared Mahalanobis distance M^2.

Table 3: Feature contribution vectors and Mahalanobis distance (M^2) when comparing the Additamentum to Silius' *Punica* and Vergil's *Aeneid*. For comparison, distance from the centroid of the *Aeneid* to the *Punica*.

| Feature	Addit. vs Punica	Addit. vs Aeneid	Aeneid vs Punica
F2WC	12.79	13.53	0.09
F1C	7.43	3.48	0.21
BD	7.06	6.26	-0.03
F4SC	2.88	0.18	1.80
F2SC	2.31	3.45	0.16
F3C	2.15	8.55	5.17
F3SC	2.09	0.35	0.35
F3WC	1.09	0.00	0.82
F4S	0.83	0.90	-0.00
F4WC	0.46	1.10	0.02
SYN	0.40	0.17	1.52
F3S	0.29	0.04	0.05
F1S	0.15	1.25	4.14
F2S	-0.21	0.64	0.21
F4C	-1.44	0.04	-0.62
F2C	-1.45	-0.59	2.00
Raw Distance	36.82	39.36	15.88
p-value	0.0013	0.0006	0.39

concentration of dactylic first feet. A poetic analysis might be able explain that divergence on stylistic grounds, considering the content of the passage. This kind of explanatory feature increases the *transparency* of computational methods. The relevant feature contribution vectors for this section are shown in Table 3.

3.5   *Interrogating the Authorship Results*

While these results are certainly very suggestive, it should be noted that they are not necessarily conclusive. Metre is only one aspect of poetic style, and it is certainly true that metre was consciously used by classical poets to impart an additional dimension to the verse. A worthwhile question to ask is whether some genuine passages might be this far from the 'typical' style. As with any statistical determination, when considering enough samples this is certainly possible and, indeed, even likely. To illustrate this, I did the following: I first removed the Additamentum from the text of the Punica, and then took a rolling window of 81-line chunks throughout the text, advancing each time by 27 lines (so that the chunks overlap). At each step, the distancve was calculated in the same way as for the Additamentum—removing that chunk from the rest of the poem, and then taking the M^2 distance based on a target distribution of 10,000 random 81-line samples. In plain terms, I tried to determine how many continuous sections of the undisputed text were at least as unusual as the Additamentum. These results are shown in Fig. 6. There were 6 chunks which

were at least as unusual, or about 1.3%. On examining them individually (albeit briefly), there seem to be plausible poetic reasons for the metrical distinctiveness of these chunks,[16] but nevertheless, the results are enough to mandate caution. Although this evidence certainly supports the theory that the Additamentum is non-genuine, more work will be required to make the analysis conclusive. In particular, it seems sensible to conduct further tests using orthogonal feature sets, such as character n-grams, function words, rare lexicon or other approaches that have been similarly well-established in attribution studies.

Figure 6: Comparison: The Mahalanobis distance of undisputed lines, taking a rolling window of 81-line chunks. The Mahalanobis distance of the Additamentum is shown as a horizontal dotted line.
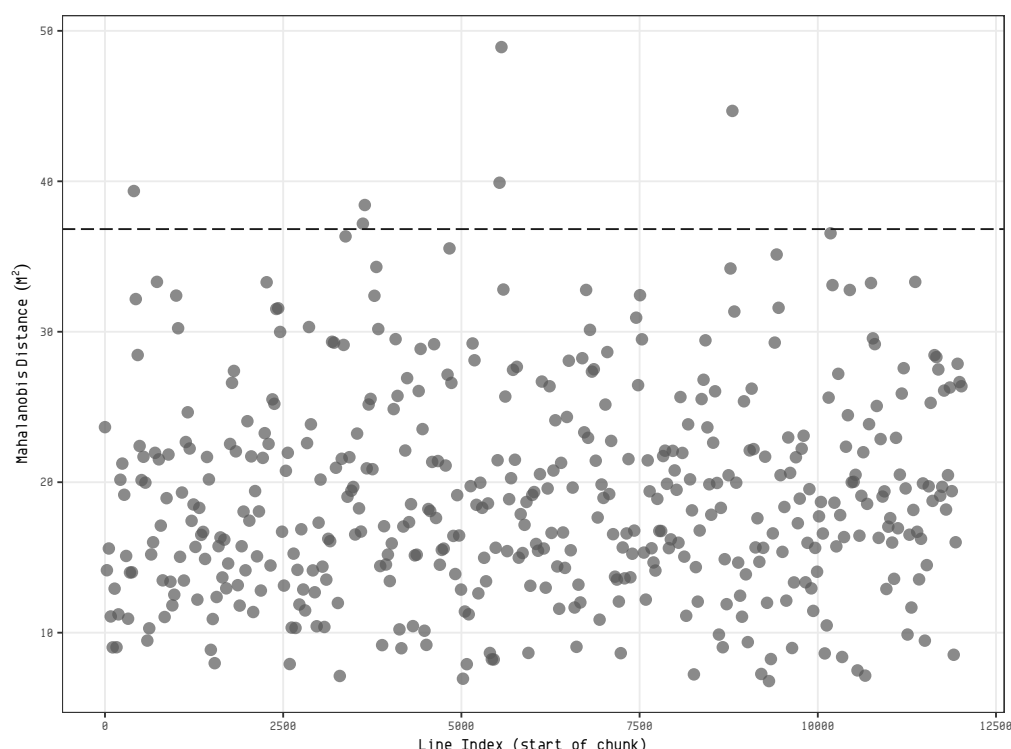
4 CONCLUSION

The clearest conclusion from this research is that metrical features are an extremely useful classifier for classical Latin hexameter poetry, and there seems little reason to doubt that they would be also effective for elegiac couplets. I note that this result is contra Forstall et al. (2011, 293), and must conclude that the difference lies in both the additional sophistication of the feature extraction (this study extracted more metrical features) and in the way the features were treated (it seems that n-gram frequencies on syllable quantity do not capture enough of the ambient stylistic information). In terms of further extensions,

---

16 To explain this in full would require a lengthy discursion. However, as a brief example, the most extreme outlier, *Puni.* 8.554–635, contains about double the average number of proper nouns, which require some metrical gymnastics to accommodate. Disregarding those lines brings the fragment back into the 'fairly normal' zone. In addition, that section (surprisingly!) contains a number of short sections of deliberate rhyme which, once again, distort the statistics.

I am very optimistic about the prospects for the less common Latin metres, and would be reasonably surprised if the techniques were not also applicable to classical Greek. While single-line observations do not seem to be useful, chunked observations can return good results with chunks as small as ten lines in cases where styles are very different (Fig. 3). Metrical features have low dimensionality, which may be useful in cases where limited training material is available. For Latin, a subset of 11 features yielded very similar accuracy to all 16 features when used for classification. When designing experiments with metrical features, practitioners should bear in mind that some features are strongly correlated (Table 2). In our classification tests, this does not seem to impede high accuracy.

In the authorship attribution test, the Mahalanobis distance was used as the metric, which is designed to correct for the issues of correlation and co-variance. As was shown, the Mahalanobis distance has two other desirable features. First, it can be used directly to calculate a p-value, which provides a well understood difference indicator. Second, during the calculation of the Mahalanobis distance, it is possible to extract a *feature contribution vector*, which seems to have useful explanatory properties (Table 3)—this allows us to 'bridge the gap' between opaque computational methods and traditional philology. To the best of my knowledge, this feature contribution vector has not been discussed in the stylometry literature. Based on this technique I analysed the disputed Aldine Additamentum in the Punica by Silius Italicus. Here there were some important differences in methodology. The most noteworthy is that a target distribution was created through random sampling with replacement. These large sample distributions provided distance measurements that minimised variation due the random process. The results appear to support those scholars who claim that the work is not genuine. With a p-value of 0.0012, the Additamentum differs, metrically, from the style of the rest of the poem in a statistically significant way. For comparison, I examined the rest of the poem using a rolling window and found only 6 passages (of 446) with higher M^2 distances. Although this result is certainly tendentious, more research should be conducted before it could be considered to be a substantial claim.

Table 4: Feature abbreviations used

Feat.	Description
F1S	First Foot, Spondee (1 if Spondee, 0 if Dactyl)
F2S	Second Foot, Spondee
F3S	Third Foot, Spondee
F4S	Fourth Foot, Spondee
F1C	First Foot, Conflict (1 if there is an ictus/accent conflict, 0 for harmony)
F2C	Second Foot, Conflict
F3C	Third Foot, Conflict
F4C	Fourth Foot, Conflict
BD	Bucolic Diaeresis (1 if the fourth foot ends at a word break: — ⏑ ⏑ ‖ or — — ‖)
F2SC	Second Foot, Strong Caesura (1 if there is a word break after the initial spondee of the second foot: — ‖)
F3SC	Third Foot, Strong Caesura
F4SC	Fourth Foot, Strong Caesura
F2WC	Second Foot, Weak Caesura (1 if there is a word break after the first breve. May only exist in dactylic feet: — ⏑ ‖)
F3WC	Third Foot, Weak Caesura
F4WC	Fourth Foot, Weak Caesura
SYN	Synalepha (elision) (A count of elisions, some lines may have more than one. Does not include prodelision.)

REFERENCES

Allen, W. S. (1965). *Vox Latina*. Cambridge: CUP.

Barber, R. (2019). Marlowe and overreaching: A misuse of stylometry. *Digital Scholarship Humanities*, *34*(1), 1–12. Retrieved from `http://academic.oup.com/dsh/article/34/1/1/5072008`

Burrows, J. (2002). 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, *17*(3), 267–287. Retrieved from `https://academic.oup.com/dsh/article-lookup/doi/10.1093/llc/17.3.267`

Chaudhuri, P., Dasgupta, T., Dexter, J. P., & Iyer, K. (2018). A small set of stylometric features differentiates Latin prose and verse. *Digital Scholarship Humanities*. Retrieved from `https://academic.oup.com/dsh/advance-article/doi/10.1093/llc/fqy070/5258951`

Colombi, E., et al. (2007). *Pedecerto.* Retrieved from `http://www.pedecerto.eu/pagine/autori`

Duckworth, G. E. (1965). Horace's Hexameters and the Date of the Ars Poetica. *Transactions and Proceedings of the American Philological Association*, *96*, 73–95. Retrieved from `https://www.jstor.org/stable/283719?origin=crossref`

Duckworth, G. E. (1967). Five Centuries of Latin Hexameter Poetry: Silver Age and Late Empire. *Transactions and Proceedings of the American Philological Association*, *98*, 77. Retrieved from `https://www.jstor.org/stable/2935870?origin=crossref`

Duckworth, G. E. (1969a). Maphaeus Vegius and Vergil's Aeneid: A Metrical Comparison. *Classical Philology*, *64*(1), 1–6. Retrieved from `https://www.jstor.org/stable/268004`

Duckworth, G. E. (1969b). *Vergil and classical hexameter poetry: a study in metrical variety*. Ann Arbor: University of Michigan Press.

Eder, M. (2015). Does size matter? Authorship attribution, small samples, big problem. *Lit Linguist Computing*, *30*(2), 167–182. Retrieved from `https://academic.oup.com/dsh/article/30/2/167/390738`

Eder, M. (2017). Short Samples in Authorship Attribution: A New Approach. In *DH.*

Eder, M., Rybicki, J., & Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *R Journal*, *8*(1), 107–121. Retrieved from `https://journal.r-project.org/archive/2016/RJ-2016-007/index.html`

Forstall, C. W., Jacobson, S. L., & Scheirer, W. J. (2011). Evidence of intertextuality: investigating Paul the Deacon's Angustae Vitae. *Literary and Linguistic Computing*, *26*(3), 285–296. Retrieved from `https://academic.oup.com/dsh/article-lookup/doi/10.1093/llc/fqr029`

Ilsemann, H. (2018). Christopher Marlowe: Hype and Hoax. *Digital Scholarship Humanities*, *33*(4), 788–820. Retrieved from `http://academic.oup.com/dsh/article/33/4/788/4951676`

Jockers, M. L., & Witten, D. M. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, *25*(2), 215–223. Retrieved from `https://academic.oup.com/dsh/article-lookup/doi/10.1093/llc/fqq001`

Kao, J., & Jurafsky, D. (2012). A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature* (pp. 8–17). Montréal, Canada: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/W12-2502`

Keil, H. (1857). *Grammatici Latini ex recensione Henrici Keilii*. Leipzig: Teubner.

Knight, W. F. J. (1931). Homodyne in the fourth foot of the Vergilian hexameter. *The Classical Quarterly, 25*(3-4), 184–194.

Lee, J. M. (2017). Silius Italicus' Punica 8. 1-241: a commentary (MPhil. Thesis). Retrieved from `http://hdl.handle.net/2440/113421`

Mastandrea, P., et al. (2007). *Musisque deoque: A digital archive of latin poetry.* Retrieved from `http://mizar.unive.it/mqdq/public/index`

Neidorf, L., Krieger, M. S., Yakubek, M., Chaudhuri, P., & Dexter, J. P. (2019). Large-scale quantitative profiling of the Old English verse tradition. *Nat Hum Behav, 3*(6), 560–567. Retrieved from `http://www.nature.com/articles/s41562-019-0570-1`

Omer, A. I. A., & Oakes, M. P. (2017). Arud, the Metrical System of Arabic Poetry, as a Feature Set for Authorship Attribution. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)* (pp. 431–436).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Platnauer, M. (1951). *Latin elegiac verse*. Cambridge: CUP.

Plecháč, P., Bobenhausen, K., & Hammerich, B. (2019). Versification and authorship attribution. A pilot study on Czech, German, Spanish, and English poetry. *SMP, 5*(2), 29–54. Retrieved from `http://ojs.utlib.ee/index.php/smp/article/view/smp.2018.5.2.02`

Schoell, F. (1876). De Accentu Linguae Latinae. In F. Ritschelius (Ed.), *Acta Societatis Philologae Lipsiensis*, VI (pp. 1–231). Leipzig: Teubner.

Tizhoosh, H. R., Sahba, F., & Dara, R. (2008). *Poetic features for poem recognition: A comparative.* Retrieved from `https://www.researchgate.net/publication/228375375_Poetic_Features_for_Poem_Recognition_A_Comparative_Study`