**STUDIES IN ENGLISH MEDIEVAL
LANGUAGE AND LITERATURE**

Edited by Magdalena Bator

57

Rodrigo Pérez Llorido /
Carlos Prado-Alonso /
Paula Rodríguez-Puente (eds.)

*Of ye Olde Englisch
Langage and Textes:*
New Perspectives on Old
and Middle English
Language and Literature

PETER LANG

Jacob Thaisen

7. Classifying scripts, with particular reference to Anglicana and Secretary¹

Abstract This paper encourages quantitative analysis of quantitative data in palaeography by way of example. Scholars have long called for studies in the field to base their findings on quantitative data, and it is positive that more and more studies doing so have begun to appear. However, it is visual analysis of quantitative data which continues to dominate, although visual analysis risks overfitting and biased predictor selection. The paper addresses these methodological weaknesses. It does so by demonstrating how classes of co-occurring allographs can replace the theoretical, binary opposition between the varieties of Gothic Cursiva script known as Anglicana and Secretary. The basis for establishing the classes is patterns of co-occurrence of seventeen allographs encountered in English texts dating from the period 1350–1450, and TwoStep Cluster Analysis is the means employed for identifying the patterns. Employing TwoStep Cluster Analysis means the classes are empirically valid in the sense that if one were to divvy the dataset up into six classes, these six would represent the most distinct ones one could find. But the number six is randomly chosen and the classes are meant for illustration only, saving determination of the optimum number of classes for future scholarship. The means employed for determining the relative contribution of each allograph to each class is regression modelling with class membership as the dependent variable and the presence or absence of the seventeen allographs in texts as the independent variables. The paper includes a comparison of conditional inference trees and generalised logistic models and demonstrates how to determine the class membership of two test texts.

¹ The research reported in this paper is associated with the Middle English Scribal Texts programme jointly funded by the Research Council of Norway and the University of Stavanger, whose support is hereby acknowledged. The preparation of the paper was facilitated by a residential fellowship at the Netherlands Institute for Advanced Study in the Humanities and Social Sciences. The author thanks Anita Auer, Raymond Hickey, Erik Kwakkel, and Ursula Lenker for arranging for him to present the research at professional gatherings, as well as Joanna Nykiel, Melissa Green, and Merja Stenroos for providing constructive criticism on earlier drafts, and Torsten Hothorn for helpfully answering queries about the *partykit* package.

1 Introduction

A call has repeatedly been made for quantitative study of palaeographical variation. Studies which meet this call still number few but represent a step in the right direction. However, it is visual analysis of quantitative data which dominates among these studies and it is rare to encounter findings bolstered up by robust statistical tests. As a consequence, collinear variables may erroneously be considered orthogonal to each other, and variables which are not predictors may be treated as predictors. It may silently be assumed valid to generalise from a sample to the population from which the sample comes, or this assumption may be addressed in vague, general terms rather than properly tested. To put the key matter in less technical terms, textbooks agree on associating the allographs single-compartment *a*, open *g*, and open *w* with Secretary script but they do not address the possibility that the open *g* allograph could be a stronger predictor of the response 'Secretary script' than the open *w* allograph, that all texts which contain the single-compartment *a* allograph could also contain the open *w* allograph such that the two allographs are not independent of each other as predictors or that they interact with each other. Moreover, a substantial number of late Middle English texts contain allographs associated with Anglicana as well as allographs associated with Secretary, making it pertinent to subdivide the gradient between them. This paper, then, explores the possible quantitative analysis of quantitative palaeographical data, focusing on methodology. Readers already familiar with regression analysis may want to skip section 2, which is introductory.

2 Regression modelling

It has become a norm to probe the sociolinguistic variable by means of regression analysis. The concept of the sociolinguistic variable was introduced by Labov (1963, 1966, 1972) and is understood as a unit of linguistic structure which a language user may realise as one of a set of semantically equivalent variants. A particular configuration of variables results in a language user producing a given variant, and the investigation then focuses on determining which of the variables are predictors and what their relative contributions are to why the language user selects the given variant. It has been suggested that palaeographical variation is not fundamentally different from orthographic variation (McIntosh 1974, 1975), but Middle English palaeographical and orthographic variants have been shown not to co-vary without exception (Thaisen 2018). Moreover, at a more theoretical level, it is debatable whether allographs and orthographic variants can be considered units of linguistic structure. The principal theoretical

argument against any such identification is that it is consciously that language users select among available allographs and orthographic variants, unlike when they select among, say, allophones. Allographic variation may nevertheless be analysed by the same means.

There is technical literature comparing and contrasting regression algorithms and motivating them mathematically. Varbrul, the first software package used by sociolinguists to build regression models (Cedergren and Sankoff 1974), is an implementation, originally for a VAX computer and in the Fortran programming language, of a maximum likelihood algorithm for stepwise logistic regression; that is, it is able to estimate the relationship between a response and one or more predictors, where the predictors are usually measured on a continuous scale. The response is always measured on a discrete scale, but different kinds of response require different algorithms: the package is able to model a binary response with its *Ivarb* programme and a multinomial response with its *Mvarb* programme. The ensuing model assumes of the predictors that they are orthogonal to each other and are free of random effects such as the effect of a single informant having been observed multiple times within the training data. Later incarnations of the Varbrul package include Rbrul, Rvarb, and Goldvarb, but general-purpose software for statistical analysis like SPSS and R has today become available which also offers regression modelling functionalities.²

Still, the researcher may be constrained by the kinds of model the available tools are capable of building, at the same time as the nature of the training data determines what kind of model the researcher ought to build. For example, while the simple *Ivarb* was fully functional by 1985, the more complex *Mvarb* had not yet been implemented on the IBM PC, and it is only comparatively recently with the *lme4* package for R that software has come on the market for building as complex a model as a generalised logistic mixed-effects model. A mixed-effects model is a model which encompasses both fixed and random effects and a generalised model is a type of model developed for a response which, like the present data, does not have a normal distribution. Time has, in other words, seen developments in computer hardware and software which have enabled researchers to build more complex models.

However, an increase in model complexity comes at a price. To fit a model means to test variables to establish whether they are predictors of the training

2 Loh (2014) sketches out developments in regression modelling in the past five decades with the needs of no specific discipline in mind, whereas Johnson (2009) compares the available software packages customarily used by linguists. Cf. Tagliamonte and Baayen (2012) and Roy (2014).

data and next to include or exclude them until a satisfactory combination of predictors has been found, that is to say a model which explains a sufficient percentage of the data. The reason why an increase in model complexity comes at a price is that the common manual method of fitting a model by eliminating variables stepwise until only predictors remain is more demanding on the researcher, especially since the result may be a function of the elimination order. If variables A and B are collinear, A may be found to be a predictor when B is eliminated whereas B may be found to be a predictor when A is eliminated. Moreover, two or more variables which are not independent predictors may together have a significant non-additive simultaneous effect on the response. This phenomenon is known as an interaction and is all the harder to detect when there are many variables.

Automated fitting may not always be a viable alternative to manual fitting. The reason is that what is selected as the best model may be dictated by the success criterion, may exclude interactions, and may be computationally taxing or even unsolvable, if the method is implemented at all. For example, a predictive accuracy of near 100 percent on the training data (that is, a model which explains near 100 percent of these data) and an unduly large number of predictors for a small number of observations usually indicates a model which exhaustively explains a sample of a population but will not successfully generalise to the population. The *bestglm* package for R avoids such overfitting. It does so by generating a separate model for every possible combination of the variables and selecting the best one based on an information theoretical measure of relative fit as the success criterion. However, the number of combinations may be prohibitive since it increases quadratically with the number of variables: with thirty variables, the package builds 2^{30} models, which adds up to more than a billion. Furthermore, the presence of a zero-variance predictor (a constant) relative to a specific value or level of the response will confound the computations.³ Algorithms may also halt if the number of random effects exceeds what is warranted by the training data (cf. Bates, Kliegl et al. 2015). Considerations of this kind sometimes favour the simpler model and the simpler algorithm: it may be advantageous to recast a multinomial response as a series of binomial ones for separate modelling. The present six classes are a multinomial response of this kind.

An important insight is that a model can be a perfectly satisfactory model even if it is not the best possible model according to some success criterion.

For example, the present article builds two different kinds of model for each class, both of which explain a reasonable percentage of the training data and will successfully generalise to the population represented by the training data. Which one of the two models will ultimately be selected as the best one is at the discretion of the researcher. Put simply, if three predictors explain 88 percent of the training data and four predictors explain 89 percent of them, it is far from unreasonable for a researcher to want to prefer the three-predictor model to the four-predictor one.

Recent implementation also holds true of the tree family of modelling methodologies as an alternative to logistic or linear ones. The *party*/*kit* package for R required for growing the member of this family that is the conditional inference tree (Hothorn, Hornik, and Zeileis 2006), discussed in what follows, came out in 2011 as a further development of the *party* package from 2005.⁴ These dates explain why a fairly recent paper probing the relationships between predictors of sibilant voicing in a variety of Spanish is correct to observe that 'this statistical method does not appear to have been widely used in phonetic research' (Strycharczuk, van 't Veer et al. 2014: 19). However, a literature search shows it increasingly being employed, with success, in various areas of linguistics, including phonetics/phonology. Examples dating from up to 2016 and not referred to elsewhere in the body of this paper are Ambbridge et al. (2012); Johnson and Barnes (2013); Jongmans et al. (2010); Moore and Carter (2015); Plug and Carter (2013); Szmyrcsanyi et al. (2016); Tagliamonte and Baayen (2012); and Tagliamonte et al. (2014).

The methodology estimates how variables work in tandem by partitioning the training data into maximally homogeneous sets and finding the variable with the strongest association with each set. The algorithm starts from the variable which is the strongest predictor as measured by the probability of a given value (continuous) or level (categorical) of it predicting the response. It binary partitions the training data based on this value or level before starting over separately for each such partition. Recursion stops when there are no more variables with a value or level which predicts the response with a probability of more than a set threshold, in the present case 5 percent, which is a conventional significance level in statistical analysis. The threshold makes the selection procedure robust against outliers and ensures a meaningful result with training data comprising

³ Tagliamonte and Baayen (2012: 22–28) discuss how to establish the usefulness of predictors using random forests.

⁴ Visit <http://cran.r-project.org/src/contrib/Archive/party/kit/> and <http://cran.r-project.org/src/contrib/Archive/party/> for the respective years of release for the *party*/*kit* and *party* packages.

a small number of observations and a large number of variables, which is the composition of the present training data. The ensuing model of the relationships between variables found to be predictors may be visualised as a tree forking at each point of partition.

Note that this methodology may lead to non-selection as a predictor of a variable which is strongly associated with the response. This happens when the algorithm selects the strongest among collinear variables, partitions the training data accordingly, and then finds the non-selected variable not to be the strongest predictor within the remaining data either. For example, in an imagined situation where all texts which contain the single-compartment a allograph also contain the open w allograph, the algorithm could select the single-compartment a allograph as the strongest predictor of the response 'Secretary script' for, say, 80 percent of the training data, thus disqualifying the open w allograph as a predictor for that 80 percent partition. It could, then, next, select some other allograph as the *strongest* predictor for the remaining 20 percent of the training data, thus disqualifying the open w allograph also as a predictor for that partition. In such a case, the open w allograph is a variable which visual analysis could fail to realise is collinear with the single-compartment a allograph. The comparative immunity of a conditional inference tree to collinearity is the specific property that motivates Owens et al. (2013) to adopt conditional inference trees as their methodology. Collinearity characterises the present data too.

Conditional inference trees circumvent two issues which cause trouble in logistic and linear modelling: They readily expose interactions between predictors, and the threshold safeguards against overfitting the model to the training data. The documentation for the *party* package promises 'that the right-sized tree is grown without additional (post-)pruning or cross-validation' (Hothorn and Zeileis 2015: 8), and Bernaish et al. (2014) confirm that the predictors of dative alternation in South Asian Englishes selected by their conditional inference tree do indeed not constitute an overfitted model. They do so by comparing the effect of the predictors selected by 2,000 separate trees, each grown on a random subset of their training data.⁵

⁵ The reliance on probability as the criterion for partitioning training data is what sets conditional inference trees apart from other non-parametric regression methodologies in the tree family. Their close relatives, Classification and Regression Trees, implemented in the *rpart* package for R (cf. Baayen 2008), rely on an information theoretical measure as the criterion for partitioning, which according to Hothorn, Hornik, and Zeileis (2006; cf. Loh, 2014: 332) unduly skews the selection of predictors

What follows describes the present training data and the corpus behind them before employing TwoStep Cluster Analysis to partition them into six classes. The training data record as variables the presence or absence of allographs within each text constitutive of the corpus. The classes represent how the texts group according to which allographs are present or absent in them. The paper then grows conditional inference trees to determine which allographs predict membership of these classes, and it translates them into generalised logistic models for comparison. The concluding remarks summarise and contextualise. It admittedly has overtones of circularity first to derive classes from training data by means of one algorithm, only next to model these same classes by applying another algorithm on these same training data. But the point is rather a different one: to show what one could do with classes if one already had them and for this purpose one needs a set of classes to work with. It is for other scholarship to suggest an empirically valid grouping variable; the most obvious way would be to work from a second set of training data independent of the first.

3 Corpus, variables, and methods

The corpus comprises the Middle English Grammar Corpus (MEG-C), version 2011.1 (Stenroos et al. 2011) as well as some forty associated texts, totalling 449 texts.⁶ This corpus samples the population of texts written in England and in English within the period c. 1350–1450 but it excludes texts deemed by the editors of the corpus to exhibit beginning orthographic standardisation. It was selected for convenience. Every text is catalogued and described by virtue of its appearing as a survey point in *A Linguistic Atlas of Late Mediaeval English* (McIntosh et al. 1986). A sample of every one of them was kindly made available to me in photographic reproduction courtesy of the Middle English Scribal Texts programme.

To collect training data from this corpus, I recorded the presence or absence by text of seventeen allographs, totalling 3,081 tokens. Allographs may consistently be relatively objectively distinguished by their shape. I did not pay any attention to the relative frequency of the allographs within the given text, their position in any given word, or small variations in their execution.

The allographs were:

in the direction of (1) variables with missing values or levels, and (2) variables with a small range of values or levels.

⁶ Transcriptions are downloadable from the University of Stavanger: <http://www.uis.no/mest>.

Single-compartment a: the single-lobed allograph of <a> with a pointed head

Double-compartment a: the double-lobed allograph of <a> whose upper lobe may extend above the x-height

Reverse e: the reversed and circular allograph of <e>

Pointed e: the allograph of <e> which has an extended tongue to the right

Closed g: the double-compartment figure 8 allograph of <g>

Open g: the single-compartment allograph of <g> with a small tail and a pointed or 'horned' head whose right limb may be higher than the left limb

Long r: the forked allograph of <r> that extends below the baseline

Short r: the right-shouldered allograph of <r> with no descender

V-shaped r: the v-shaped allograph of <r> with no descender

Sigma s: the sigma-shaped allograph of <s>

Kidney s: the word-final allograph of <s> which resembles a small 'B' or a 'c' and a '3' run together

ll3 w: the allograph of <w> which is composed of two looped 'l' and '3'

Round w: the allograph of <w> which resembles a circle enclosing a '2' or '3'

Open w: the simplified allograph of <w> which resembles two 'v'

Two-stroke x: the allograph of <x> which is executed as two separate strokes and possibly has a bar through the centre

Bottom x: the allograph of <x> which is executed as a single stroke to form a loop at the bottom

Right x: the allograph of <x> which is executed as a single stroke to form a loop to the right

I arranged these training data into a matrix with the 449 texts in the rows and the seventeen allographic variables in the columns.

The palaeographical literature associates these seventeen allographs with the two varieties of Gothic Cursiva script labelled *Anglicana* and *Secretary* by Parkes (1969). 'Associates' is the most accurate wording since that literature offers an inexact definition of either variety as a specific configuration of allographs, as has been mentioned. It enumerates all the allographs theoretically separating the two varieties,⁷ but is perhaps unclear on what the classification of a text would be if only some of the allographs were observed in it. Moreover, the literature frequently notes how the maximum contrast is rarely observed in actual texts. What abounds are hybrids (Petti 1977: 15; Roberts 2005: 161–2). A random text will

⁷ There is a slight disagreement among palaeographers. For example, Parkes (1969) and Petti (1977) do not include right x among the allographs with separate *Anglicana* and *Secretary* but Roberts (2005) does.

likely contain allographs associated with *Anglicana* alongside allographs associated with *Secretary*, and no clear criteria appear to have been devised for placing such hybrid configurations on a gradient from the one variety to the other. Since the hybridisation extends to allographs of the same grapheme, it is not possible to treat a grapheme as a categorical variable with its allographs as its levels.

TwoStep Cluster Analysis (Bacher et al. 2004), implemented in the IBM SPSS Statistics software package, is a method for dividing categorical data up into maximally homogeneous classes. I used this method to establish six such classes, and added a column to the training data matrix with membership, by text, of these six classes to represent the response.

4 Results

The R software environment for statistical computing, version 3.1.2 (Pumpkin Helmet) grew six conditional inference trees by means of the *party* package and subsequently tested their accuracy. Fig. 1 presents the six trees. Each tree estimates the hierarchy of allographic predictors of membership of a given class. Variables absent from a node predict class membership with a probability of less than the 5 percent threshold given its mother nodes and are therefore disregarded as predictors; those variables number several more than do the predictors. Every node has a boxed identifier and gives the probability value associated with the binary partitioning of the training data at that node. The length of an edge is solely determined by aesthetics of layout, but the annotations which interrupt the edges have meaning: they indicate the levels of the predictors at which the training data are partitioned. The label 'x' associated with a terminal node gives the number of observations of the predictor in the training data. The bar charts which relate to terminal nodes have a right-hand scale which indicates the proportion of levels of the response given the particular configuration of predictors resulting in the particular terminal node, while their left-hand scale indicates levels of the response.

Near-categorical responses are in evidence. Specifically, it can be seen from the tree for Class 1 that the open g allograph is the strongest predictor of this class and that the highest proportion of membership in it is found in terminal node 11, where it is nearly categorical. The predictors must be configured as follows to lead to this response: the open g allograph must be present (node 1), the two-stroke x allograph must be absent (node 9), and the closed g allograph must be absent (node 10). The respective trees for three of the remaining five classes resemble the tree for Class 1. They do so in having too no more than a single terminal node where the response is categorical or near-categorical: node

Fig. 1: Conditional inference tree estimates of the relationship between allographic predictors of class membership

Tab. 1: Accuracy of conditional inference trees predicting membership of Classes 1–6

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Predicted						
Observed	0	1	0	1	0	1
1	392	19	361	24	366	16
2	36	5	59	2	65	15
3	63	12	73	3	64	
Key	0: Incorrect	1: Correct				

12 for Class 2, node 7 for Class 3, and node 7 for Class 5. The tree for Class 6 has two terminal nodes with high proportions of membership in the class, nodes 6 and 7. However, this non-resemblance to the tree for Class 1 is superficial, since the two nodes have the same mother, node 3 (the closed g allograph). Classes 1, 2, 3, 5, and 6 are clearly distinct.

Although found by the TwoStep Cluster Analysis, the conditional inference tree for Class 4 is impressionistically less solid as a definition of a maximally homogeneous class. What gives this impression is that the two terminal nodes where the response is near-categorical, nodes 12 and 15, share no other node than the root.

Tab. 1 shows the accuracy of the trees. The accuracy of a tree is measured as the extent to which the responses observed in the training data agree with the predicted responses, that is, the extent to which the allographs observed in a text lead the given tree to assign that text to the same class as the TwoStep Cluster Analysis. It is apparent that the predictive accuracy of the trees is high: it ranges from 98.2 percent (3+5=8 false predictions out of 449 observations) for Class 6 to 91.5 percent for Class 4. These statistics thus confirm the clear distinction between the classes suggested by the near-categorical responses, and also the impression of Class 4 as being a slightly less homogeneous class. This being the nature of the six classes present in the training data, they reasonably lead to an expectation that testing on other data will validate them.

Tab. 2 offers a separate generalised logistic model for each class with class membership as the (binary) response, the multinomial response with six levels having been recast as a series of six binomial ones for separate modelling, as promised. The table includes all variables for the sake of completeness, although presence or absence of the V-shaped r allograph, for example, is never a predictor in any of the models. A coefficient estimate expresses the slope of the regression line for a particular predictor. Positive coefficients increase the likelihood of observing the predicted response and negative coefficients decrease it. The

Tab. 2: Regression coefficients and probability values in generalised logistic regression

Predictor {default level}	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
(Intercept)	-1.0183.	-0.5351	-15.6706***	-12.3351***	-1.4233***	-2.8437***
Single-comp a {present}		5.5614***				
Double-comp a {present}		-3.0923***	8.6116**			
Pointed e {present}			4.3785***		-2.1463***	
Reverse e {present}	-0.9926.		-6.4764***			-4.2041***
Closed g {present}	-3.6757***	0.2190		5.4641***		2.0571**
Open g {present}	3.0175***					
Long r {present}			12.7027***			
Short r {present}		-4.7556***		0.8299*		
V-shaped r {present}						
Sigma s {present}			-3.3098***	4.3846***		
Kidney s {present}		-5.0200***		1.8152***		
Round w {present}		4.5723***				
Open w {present}	3.0741***	-4.2550***	-6.8118***	1.3963**		
ll3 w {present}		-0.5650				0.6460
Two-stroke x {present}						
Right x {present}				3.1363***		
Bottom x {present}						
Significance codes			*** 0.001	** 0.01	* 0.05	. 0.1

intercept expresses the predicted level of the response when all predictors are nil, and the star notation is shorthand for a probability value that is inversely related to the effect of a predictor on the response: the more stars, the stronger the predictor. A key is provided immediately below the table. Curly brackets surround the default value of a predictor. The *stats* package for R built these models with *logit* as the link function, which is the canonical link function for binomial data.

The models are simple main effects models; that is, they exclude interactions, even though interactions are in evidence in the training data. (It is common in logistic and linear regression analysis for models to do so.) For example, presence of the reverse e allograph does not independently predict Class 1 while presence of the reverse e and open w allographs together robustly predict it. Nodes 4 and 6 in the conditional inference tree for Class 1 in Fig. 1 corroborate this interaction. Presence of the short r, kidney s, and open w allographs provide another example as predictors of Class 2. Compare nodes 9, 10, and 11 in the tree for Class 2 in Fig. 1. Although the models exclude interactions, they have an acceptable accuracy as tested by the chi-squared value on the residual deviance and the degrees of freedom. The probabilities associated with these values are 1, 1, 1, 0.6153, 0.9787, and 1 for the respective models for Classes 1–6. None of these probabilities indicates any statistically significant lack of fit, at the same time as they too suggest Class 4 (at 0.6153) is less homogeneous.

These models can be used to assign class membership to configurations of allographs found in other texts from the period than those included in the corpus. The probability that a text belongs in Class *n* can be calculated according to Equation 1:

$$p(\text{Class } n) = 1/1 + e^{-X\beta} \quad [\text{Equation 1}]$$

Where *e* is the base of the natural logarithm, *n* the identifier of the class, and *Xβ* the sum of the coefficients given in Tab. 2.

Two test texts will illustrate. Derbyshire Record Office, Matlock D 187/1/212 is an indenture which is not among the texts which have contributed training data. It has the double-compartment a, reverse e, closed g, long r, sigma s, and ll3 w allographs. *Xβ* for this text is -5.6866 for Class 1, calculated as -1.0183 (the intercept) -0.9926 (reverse e) -3.6757 (closed g), since the remaining variables are not predictors of Class 1 and so do not contribute to *Xβ*. Inserting this value into Equation 1 gives $p(\text{Class } 1) = 0.0034$. The values of *Xβ* for the other five classes are, respectively, -3.9734, -29.5479, -2.4864, -1.4233, and -4.3447, corresponding to $p = 0.0185$, <0.0001, 0.0768, 0.1941, and 0.0128. The logistic models thus select Class 5 as the most likely membership for the Matlock text. Constructing the models afresh to include the zero-variance predictors results

in their selecting Class 5 with greater certainty: $X\beta = -1.0183$ (the intercept) $+20.4980$ (reverse e) $= 19.4797$, corresponding to $p = 0.9999$. The conditional inference trees too select Class 5 as the most likely membership for this text. The configuration of allographs results in terminal node 7 in the tree for Class 5 given in Fig. 1, and the proportion of membership in the class is above 80 percent for this node.

The other test text is Durham University Library, Dean & Chapter Muniments, Locellus 25.138. This text is another document, specifically a bill, and it too has not contributed allographs to the present training data. The configuration of allographs observed in it comprises single-compartment a, double-compartment a, reverse e, closed g, long i, sigma s, and ll3 w, which together select Class 2 as the most likely class ($X\beta = 1.5880$, $p = 0.8303$). As before, the conditional inference trees agree with the logistic models about the likely membership: The configuration selects terminal node 12 in the tree for Class 2 in Fig. 1, and the proportion of membership in the class is above 90 percent for this node.

It was the conditional inference trees presented in Fig. 1 which gave predictors for the generalised logistic models presented in Tab. 2. Two observations deserve comment in this regard. First, if the risk of overfitting a model increases with its complexity, the conditional inference methodology is preferable: it attractively favours the less complex model while maintaining an acceptable accuracy. It is possible to build generalised logistic main-effects models which have a higher quality than the models presented in Tab. 2 do, as measured by the value of AIC (Akaike information criterion) or BIC (Bayesian information criterion). The higher quality is in the cases of all six classes achieved through adding further predictors to the existing ones. Such further predictors make a marginal contribution to predicting a class in those models. The predictors that are included in Tab. 2 but were rejected by the logistic modelling methodology are similarly marginal. Examples total up the closed g and ll3 w allographs as predictors of Class 2 and the ll3 w allograph as a predictor of Class 6.

Second, the logistic models exclude zero-variance predictors (constants), whereas the tree-structured models include them. The absence of variation in the levels of these predictors in relation to the response makes them the strongest possible predictors of a class, but it is also the property of them which causes a modelling algorithm like *besiglm* to fail. Where it may require visual inspection to prune zero-variance predictors from training data for a logistic model, they are readily visible in a conditional inference tree. Four of the five predictors are zero-variance ones in the extreme case of Class 5, where proportions of membership in the class are found only below node 6 (the pointed e allograph) in the tree presented in Fig. 1.

5 Concluding discussion

The TwoStep Cluster Analysis has distinguished six maximally homogeneous classes of co-occurring allographs within the present training data. The classes represent a solid partitioning of the training data. The tree-structured regression analyses have successfully modelled the classes and have objectively determined the contribution of each allographic variable to each class without bias in the identification of variables as predictors. The tree-structured models appear to disfavour complexity in comparison to logistic (and linear) modelling methodologies in the sense that the latter more often lead to a variable whose contribution to the response is marginal being identified as a predictor. Importantly, as regression models with sufficient accuracy, the trees will generalise to the population from which the training data were taken. In other words, the models do not exclusively model the training data, for the population of texts written in Gothic Cursiva script, in English, and in England in the late mediaeval period will fall into the same six classes. These classes will be predicted by the same allographs, and the proportions of membership in each class will be similar.

A methodological point is the use of co-occurrence patterns for allographs across texts as the basis for defining classes. When the palaeographical literature describes allographs as predictors, it typically describes them as being orthogonal on each other (i.e. independent of each other) and to be of equal strength. Such a listing of such predictors in effect amounts to an 'unfitted' model. It is plainly erroneous to assume of them both that they are orthogonal to each other and that they contribute equally to a classification. Models which rest on such an assumption may for these reasons be unsuccessful as estimations of relationships among predictors and they are unlikely to successfully generalise from sample to population.

Two non-English examples provide cases in point. Kwakkel's notion of Gothic Weight and the Liefinck system. The eleventh to thirteenth centuries saw a transition from Carolingian minuscule to Gothic Textualis and Cursiva scripts throughout Europe. Out of some thirty categorical predictors with levels that contrast between Carolingian and Gothic scripts, a text receives two points if the examples of a predictor present in it have the level associated with Gothic, zero points if the examples have the level associated with Carolingian, and one point if examples of both levels occur.⁸ It can be seen that the total score for a text, its Gothic Weight, assumes of the predictors not only that they are predictors in the first place and not just variables, but also that they are both orthogonal to each

⁸ For a description of Gothic Weight, see <http://medievalbooks.nl/tag/manuscript/>.

other and contribute equally to predicting the position of a text on a gradient from Carolingian to Gothic. Kwakkel (2012) does not explicitly test the validity of any of these assumptions about his training data.

The Liefinck classification system was developed for the palaeography of books produced in the Low Countries in the thirteenth and fourteenth centuries (Gumbert 1976, 1988, 2000). It is based on three binary variables: (1) presence vs absence of looped ascenders on the graphemes , <h>, <k>, and <l>; (2) presence vs absence of descenders on <f> and the long allograph of <s>; and (3) single- vs double-compartment a. The eight possible combinations of these three binary variables define three higher classes and a number of lower ones. Although scholarship has found that most texts from the appropriate region and period, including documents, do fall squarely into these classes (Burgers 2007), other scholarship has called attention to how it is impressionistic identification and ranking of the variables as predictors that underlie the system (Gumbert 2000; cf. Derolez 2003: 20–24), in agreement with the present paper.

The classification into either *Anglicana* or *Secretary* is yet another case in point. The *Anglicana* variety of *Cursiva* script appears to have developed in England from Gothic *Textualis* in the mid-thirteenth century and to have spread from documents to literary works. The *Secretary* variety, by contrast, was probably imported from Italy by way of France in the third quarter of the fourteenth century and too first appeared as a document hand and only later also as a book hand. It was the Privy Seal and Signet offices which introduced it on English soil, a development which Chaplais (1971: 52) describes as 'the result of a break-away from English writing traditions rather than the outcome of a gradual evolution, implying a conscious development. The separate origins of the two varieties may be the reason why the palaeographical literature insists on a binary classification despite the lack of empirical support for it. What is commonly observed in texts are, as has been mentioned, not configurations of allographs corresponding to any binary classification but hybrids of various kinds. The six classes of 'hybrids' distinguished by the present TwoStep analysis appear to be solid, and the present regression models clearly show how the presence or absence of the allographs as variables does not contribute equally to the selection of any given class. One wonders whether the hybrids are what was positively taught in late mediaeval writing centres and whether there was not a gradual evolution after all, with the *Secretary* allographs arriving to England one by one.

Chronology, genre/text-type and geography do not exhaust the list of non-allographic variables with which a class could exhibit a significant correlation. Another variable of interest is immediate palaeographical context. This is because, impressionistically, there are changes over time in how certain allographs prefer,

or are exclusively encountered in, certain positions in the word. For example, as is well-known, late tokens of the reverse e allograph do not always precede a space, unlike early ones, and there may be restrictions on whether a token of the sigma s allograph can follow a space or precede an allograph which has an ascender.

The most appropriate model for training data which take account of palaeographical context as a variable would be a mixed-effects model, since more than one token of every allograph would be collected from every text. The *lme4* package for R offers facilities for building a generalised logistic mixed-effects model, and a tree-structured alternative is viable: the R package for growing RE-EM (Random Effects Estimation Method) regression trees has comparatively recently been revised to build on the conditional inference modelling methodology available in the *party* and *partykit* packages (Fu and Simonoff, 2015). The variable 'palaeographical context', then, illustrates how quickly the constraints of the available tools are being addressed.

MEG-C does not readily permit accurate determination of the date or place of production for a text with a view to adding such information to the training data. This is because MEG-C is synchronic by design (McIntosh et al. 1986), and because the reported dates of production are sometimes stated in the texts themselves but at other times rely on secondary sources, including scholarly argument from palaeography. Posited dates of the latter type may be exact to within a quarter- or half-century. Moreover, it is unclear from McIntosh et al. (1986) and the MEG-C *Catalogue of Sources* which texts are extralinguistically localised and which are fitted into a reconstructed continuum relative to other texts based on their linguistic properties.

With these provisos, two patterns are nonetheless discernible which encourage further study. One is that texts belonging to Classes 1 or 4 tend to be later than texts belonging to the other classes; in fact, the training data comprise no more than a mere three texts belonging to Classes 1 or 4 which date from before 1400. The other pattern is that there are hardly any observations of a text belonging to Classes 3 or 6 which are localised north of 500 northings. The mean is 292 northings for texts belonging to those two classes, as against one of 384 northings for the other four classes; but the corpus is skewed such that later texts tend to be of more northerly extraction.⁹

⁹ Thaisen (2017) relates the distribution of the seventeenth allographs to non-allographic predictors such as texts' localisation in northings.

The present paper is not alone in seeking to lend greater precision to the quantitative study of palaeographical variables. The most notable example is probably yet another non-English one: the Mediaeval Palaeographical Scale, which is a support vector regression model. It is trained on inventories of allographs encountered in dated documents produced in four localities in the Low Countries in the period 1300–1550 (He et al. 2014, Samara 2015). Such work is, however, still the exception. It is true that palaeographical studies building on quantitative data are becoming more common, but it is also true that they tend to disappoint by relying on visual analysis of graphs and tables, manual assignment of weight to predictors, and untested assumptions of variables both being predictors in the first place and being independent of each other.

Last, to develop the principal point, palaeography may lack a suitable collection of dated texts which can serve as a reference corpus (Gumbert 2000). However, what constitutes a useful and meaningful reference corpus in relation to a study of production date or any other variable, including allographic ones as predictors of each other, is not a collection of texts in itself. It is a sufficiently accurate model trained on data extracted from it, since a sufficiently accurate model will generalise to other collections.

Quantitative analysis of quantitative data will outperform visual analysis of them but this is not unproblematic. For the present, logistic (or linear) regression models insisted on including marginal predictors and as such revealed themselves to be prone to overfitting and biased predictor selection. Conditional inference trees, on the other hand, circumvent these two issues. These trees are always as accurate as they can possibly be given the variables and the training data. Based on training data harvested from a corpus of texts, this paper has, accordingly, grown and validated tree-structured models estimating the relationships obtaining within the population of texts sampled by this corpus. While the subdivision into exactly six classes was random, the paper has indicated ways in which such subdivision along the gradient between the Anglicana and Secretary varieties of Gothic Cursiva script may enrich studies of them.

6 References

- Ambridge, Ben, Pine, Julian, Rowland, Caroline, & Chang, Franklin. 2012. The Roles of Verb Semantics, Entrenchment, and Morphophonology in the Retreat from Dative Argument-Structure Overgeneralization Errors. *Language* 88: 45–81.
- Baayen, Rolf Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.
- Bacher, Johann, Wenzig, Knut, & Vogler, Melanie (2004). SPSS TwoStep Cluster – a first evaluation. (2., corr. ed.) (Arbeits- und Diskussionspapiere / Universität Erlangen-Nürnberg, Sozialwissenschaftliches Institut, Lehrstuhl für Soziologie, 2004-2). Nürnberg: Universität Erlangen-Nürnberg, Wirtschafts- und Sozialwissenschaftliche Fakultät, Sozialwissenschaftliches Institut Lehrstuhl für Soziologie. <https://nbn-resolving.org/urn:nbn:de:0168-ss0ar-327153> (accessed 31 March 2020)
- Bates, Douglas, Kliegl, Reinhold, Vasishth, Shravan, & Baayen, Rolf Harald. 2015. Parsimonious Mixed Models. <http://arxiv.org/pdf/1506.04967v1.pdf> (accessed 1 November 2019).
- Bernaisch, Tobias, Gries, Stefan Th., & Mukherjee, Joybrato. 2014. The Dative Alternation in South Asian English(es). *English World-Wide* 35: 7–31.
- Burgers, Jan. 2007. Palaeography and Diplomats: The Script of Charters in the Netherlands during the Fourteenth and Fifteenth Centuries. *Quaerendo* 37: 1–23.
- Cedergren, Henrietta & Sankoff, David. 1974. Variable Rules: Performance as a Statistical Reflection of Competence. *Language* 50: 333–55.
- Chaplais, Pierre. 1971. *English Royal Documents: King John-Henry VI 1199–1461*. Oxford: Clarendon Press.
- Derolez, Albert. 2003. *The Palaeography of Gothic Manuscript Books: From the Twelfth to the Early Sixteenth Century*. Cambridge: Cambridge University Press.
- Fu, Wei & Simonoff, Jeffrey S. 2015. Unbiased Regression Trees for Longitudinal and Clustered Data. *Computational Statistics and Data Analysis* 88, C: 53–74.
- Gumbert, Johan Peter. 1976. A Proposal for a Cartesian Nomenclature. In Gumbert, Johan Peter & de Haan, Max J. M., eds. *Miniatures, Scripts, Collections*. Amsterdam: Van Gendt, 45–52.
- Gumbert, Johan Peter. 1988. *Manuscripts Datés Conservés Dans les Pays-Bas: Catalogue Paléographique des Manuscrits en Écriture Latine Portant des Indications de Date. T. 2: Les Manuscrits d'Origine Néerlandaise (XIVe-XVIIe Siècles) et Supplément au Tome Premier*. Leiden: E. J. Brill, 23–31.
- Gumbert, Johan Peter. 2000. Writing and Dating: Some General Remarks. *Scriptorium* 54: 5–8.
- He, Sheng, Samara, Petros, Burgers, Jan, & Schomaker, Lambert. 2014. Towards Style-based Dating of Historical Documents. 14th International Conference on Frontiers in Handwriting Recognition (ICFHR), Heraklion, September 2014, 265–70.
- Hothorn, Torsten, Hornik, Kurt, & Zeileis, Achim. 2006. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* 15: 651–74.

- Hothorn, Torsten & Zeileis, Achim. 2015. A Toolkit for Recursive Partitioning, Version 1.0–5. <https://cran.r-project.org/web/packages/partykit/partykit.pdf> (accessed 16 April 2016).
- Johnson, Daniel Ezra. 2009. Getting Off the GoldVarb Standard: Introducing Rbrul for Mixed-Effects Variable Rule Analysis. *Language and Linguistics Compass* 3: 359–83.
- Johnson, Mary & Barnes, Sonia. 2013. *Haywavs, Haiga*: An Analysis of the Variation Observed in Mexican Spanish Using a Mixed Effects Model. In Carvalho, Ana M. & Beaudrie, Sara, eds. *Selected Proceedings of the 6th Workshop on Spanish Sociolinguistics*. Somerville: Cascadia Proceedings Project, 32–40.
- Jongmans, Petra, Wempe, Ton G., Tinteren, Harm van, Hilgers, Frans J. M., Pols, Louis C. W. & As-Brooks, Corina J. van. 2010. Acoustic Analysis of the Voiced-Voiceless Distinction in Dutch Tracheoesophageal Speech. *Journal of Speech, Language, and Hearing Research* 53: 284–97.
- Kwakkel, Erik. 2012. Biting, Kissing and the Treatment of Feet: The Transitional Script of the Long Twelfth Century. In Kwakkel, Erik, McKitterick, Rosamund, & Thomson, Rodney, eds. *Turning over a New Leaf: Change and Development in the Medieval Book*. Leiden: Leiden University Press, 79–126.
- Loh, Wei-Yin. 2014. Fifty Years of Classification and Regression Trees. *International Statistical Review* 82: 329–438.
- Labov, William. 1963. The Social Motivation for a Sound Change. *Word* 19: 273–309.
- Labov, William. 1966. *The Social Stratification of English in New York City*. Washington: Center for Applied Linguistics.
- Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.
- McIntosh, Angus. 1974. Towards an Inventory of Middle English Scribes, *Neuphilologische Mitteilungen* 75: 602–24.
- McIntosh, Angus. 1975. Scribal Profiles from Middle English Texts. *Neuphilologische Mitteilungen* 76: 218–35.
- McIntosh, Angus, Samuels, Michael L., Benskin, Michael, with the assistance of Margaret Laing & Keith Williamson 1986. *A Linguistic Atlas of Late Medieval English*. Aberdeen University Press, Aberdeen.
- Moore, Emma & Carter, Paul. 2015. Dialect Contact and Distinctiveness: The Social Meaning of Language Variation in an Island Community, *Journal of Sociolinguistics* 19: 3–36.
- Owens, Jonathan, Dodsworth, Robin, & Kohn, Mary. 2013. Subject Expression and Discourse Embeddedness in Emirati Arabic. *Language Variation and Change* 25: 255–85.

- Parke, Malcolm B. 1969. *English Cursive Book Hands, 1250–1500*. Oxford: Oxford University Press.
- Petti, Anthony G. 1977. *English Literary Hands from Chaucer to Dryden*. Cambridge: Harvard University Press.
- Plug, Leendert and Carter, Paul. 2013. Prosodic Marking, Pitch and Intensity in Spontaneous Lexical Self-Repair in Dutch. *Phonetica* 70: 155–81.
- Roberts, Jane. 2005. *A Guide to Scripts Used in English Writings up to 1500*. London: British Library Publications.
- Roy, Joseph. 2014. Sociolinguistic Statistics: The Intersection between Statistical Models, Empirical Data and Sociolinguistic Theory. *Proceedings of Methods XIV: Papers from the Fourteenth International Conference on Methods in Dialectology, London, August 2011*, 261–75.
- Samara, Petros. 2015. Towards a Medieval Palaeographical Scale (1300–1550). In Fees, Irmingard, Hotz, Benedikt & Schönfeld, Benjamin, eds. *Papsturkundenforschung zwischen internationaler Vernetzung und Digitalisierung: Neue Zugangsweisen zur europäischen Schriftgeschichte*, <http://hdl.handle.net/11858/00-001S-0000-0023-9A13-A>.
- Stenroos, Merja, Mäkinen, Martti, Horobin, Simon, & Smith, Jeremy (compilers). 2011. *The Middle English Grammar Corpus, version 2011.1*. University of Stavanger. <http://www.uis.no/mest> (accessed 1 November 2019).
- Strzcharczuk, Patrycja, Van 't Veer, Marijn, Bruil, Martine & Linke, Kathrin. 2014. Phonetic Evidence on Phonology-Morphosyntax Interactions: Sibilant Voicing in Quito Spanish. *Journal of Linguistics* 50: 403–52.
- Szmezcany, Benedikt, Biber, Douglas, Egberg, Jesse & Franco, Karlien. 2016. Toward More Accountability: Modeling Gentive Variation in Late Modern English. *Language Variation and Change* 28: 1–29.
- Tagliamonte, Sali A. & Baayen, Rolf Harald. 2012. Models, Forests and Trees of York English: was/were Variation as a Case Study for Statistical Practice. *Language Variation and Change* 24: 135–78.
- Tagliamonte, Sali A., Durham, Mercedes, & Smith, Jennifer. 2014. Grammaticalization at an Early Stage: Future be going to in Conservative British Dialects. *English Language and Linguistics* 18: 75–108.
- Thaisen, Jacob. 2017. Secretary Letter-shapes in County Durham. *Folia Linguistica Historica* 38: 263–280.
- Thaisen, Jacob. 2018. The Round Allograph of <r> in Late Middle English. *Studia Anglicana Posnaniensia* 53: 129–144.