Expert Systems With Applications 200 (2022) 116903

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Multiple regression techniques for modelling dates of first performances of Shakespeare-era plays[☆]

Pablo Moscato [a,*], Hugh Craig [b], Gabriel Egan [c], Mohammad Nazmul Haque [a], Kevin Huang [d], Julia Sloan [d], Jonathon Corrales de Oliveira [d]

[a] School of Information and Physical Sciences, The University of Newcastle, University Drive, Callaghan, NSW 2308, Australia
[b] School of Humanities, Creative Industries and Social Sciences, The University of Newcastle, University Drive, Callaghan, NSW 2308, Australia
[c] School of Humanities, De Montfort University, The Gateway, Leicester, LE1 9BH, UK
[d] California Institute of Technology, 1200 E California Blvd, Pasadena, CA 91125, United States

ARTICLE INFO

Keywords:
Shakespeare-era plays
Continued fraction regression
Dating of plays
Play's genre
Memetic algorithm

ABSTRACT

The creation of new computational methods to provide fresh insights on literary styles is a hot topic of research. There are particular challenges when the number of samples is small in comparison with the number of variables. One problem of interest to literary historians is the date of the first performance of a play of Shakespeare's time. Currently this must usually be guessed with reference to multiple indirect external sources, or to some aspect of the content or style of the play. This paper highlights a dating technique with a wider potential, using this particular problem as a case study.

In this contribution, we introduce a novel dataset of Shakespeare-era plays (181 plays from the period 1585–1610), annotated by the best-guess dates for them from a standard reference work as metadata. We introduce a memetic algorithm-based Continued Fraction Regression (CFR) which delivered models using a small number of variables, leading to an interpretable model and reduced dimensionality, applied for the first time here in a problem of computational stylistics.

Our independent variables are the probabilities of occurrences of individual words in each one of the plays. We studied the performance of 11 widely used regression methods to predict the dates of the plays at an 80/20 training/test split. An in-depth analysis of the most commonly occurring 20 words in the CFR models in 100 independent runs helps explain the trends in linguistic and stylistic terms. The use of the CFR has helped us to reveal an interesting mathematical model that links the variation in the use of the words through time, which helps to provide estimates of the dates of plays of the Shakespeare-era. We check for genre effects as a possible confounding variable.

1. Introduction and motivation for the study

The motivation of our study is to understand the performance of widely used machine learning methods for multivariate regression problems for the area of literary chronology. We are also motivated to evaluate the generalisation ability of some of these methods for other works that could potentially be older than the literary work of earliest first appearance in our dataset. Equally, a work may have had its first appearance after the newest work in our training set. We hoped that some mathematical models would be able to generalise well not only those "within the range of first occurrences" in our training set, but also outside of it. We are also interested in the important problem of identifying if such information can be extracted from the probability of occurrence of particular words of the language. We have previous experience using word probability on several occasions

[☆] Funding: This work was supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (projects DP160101527, DP200102364). P.M. acknowledges a generous donation from the Maitland Cancer Appeal. This work has been supported by the University of Newcastle and Caltech Summer Undergraduate Research Fellowships (SURF) program. In particular, SURF Fellows J. Sloan and K. Huang acknowledge the support of Samuel P. and Frances Krown and Arthur R. Adams, respectively, for their generous donor support to their activities.

[*] Correspondence to: Room: ES 230, Engineering Science - D.W. George Building, The University of Newcastle, University Drive, Callaghan, NSW 2308, Australia.

E-mail addresses: Pablo.Moscato@newcastle.edu.au (P. Moscato), Hugh.Craig@newcastle.edu.au (H. Craig), gegan@dmu.ac.uk (G. Egan), Mohammad.Haque@newcastle.edu.au (M.N. Haque), khuang@caltech.edu (K. Huang), jsloan@caltech.edu (J. Sloan), jonco@caltech.edu (J. Corrales de Oliveira).

https://doi.org/10.1016/j.eswa.2022.116903
Received 18 April 2021; Received in revised form 9 December 2021; Accepted 12 March 2022
Available online 22 March 2022
0957-4174/© 2022 Elsevier Ltd. All rights reserved.

regarding Shakespeare-era authors' work (Arefin et al., 2015, 2014; Marsden et al., 2013; Naeni et al., 2016; Rosso et al., 2009; Zaher et al., 2015) and the recognised importance of literary chronology of English Renaissance plays. Therefore, we developed a unique annotated dataset which we are now contributing here for the first time to study the performance of several regression methods.

We first start by explaining the relevance of uncovering the literary chronology of English Renaissance plays using objective quantitative approaches. In 1778, a century and a half after Shakespeare's death in 1616, the scholar Edmond Malone published the first attempt to give dates to Shakespeare's plays and to place them in chronological order (Malone, 1778). Malone relied on allusions to the plays in documents surviving from Shakespeare's time and on evidence from the early printed editions. He admitted to many doubts and uncertainties about his suggested dates and ordering, and the debate has continued unabated since.

Shakespeare's plays and those of his contemporaries were performed and printed in an era when little attention was paid to recording dates for posterity. The focus of the theatre was commercial and theatrical, rather than literary or archival. Since stage performance was paramount, and audiences in the theatre paid almost all the bills, with printed versions and income from commissions to perform at court accounting for only a fraction of revenue, the drama was in large part an ephemeral art form. Many plays, perhaps the majority, have been lost and the documentation for those that survive is incomplete.

Over the years various new kinds of evidence about dating have been added to supplement what can be gleaned from the documentary record. In the latter part of the nineteenth century, Frederick G. Fleay argued that changes in Shakespeare's versification were a useful guide to chronology (Fleay, 1876, pp. 122–38) and this was taken up and extended by twentieth-century researchers (Gray, 1931; Langworthy, 1931; Oras, 1960; Wentersdorf, 1951) and continues in the twenty-first (Bruster & Smith, 2016). The editors of the 1987 Oxford Shakespeare used changes in the incidence of colloquialisms in the dialogue of the plays as an index of the order of the plays (Wells et al., 1987, pp. 69–144). MacDonald P. Jackson drew on the progressive decline in the length of speeches in Shakespeare as a marker of chronology (Jackson, 2007a; Taylor & Loughnane, 2017, Table 25.4).

The language of the plays more broadly has also been analysed for clues to dating. Eliot Slater introduced shared rare words as evidence of links between Shakespeare plays written at about the same time (Slater, 1975, 1988). Jackson has taken this up in various Shakespeare chronology studies, using larger text collections to calibrate rarity (Jackson, 1978, 2006, 2011, 2014, 2018). Waller (1966) looked at incoming and outgoing word forms like 'does' and 'doth' and Brainerd (1980) collected a set of words which appeared to vary in incidence with date in Shakespeare plays. Craig (2013) sought markers of change over time in very common and rarer words by comparing the language of sets of plays by Shakespeare's contemporaries, as well as by Shakespeare, from different eras.

Researchers have drawn on a range of different features of the plays for quantitative studies of chronology. Generally the focus has been exclusively on the works of Shakespeare rather than on the wider set of English plays of this period. This narrower set of samples has limited the opportunity for validating methods by withdrawing items and predicting their dates as if they were of unknown provenance. In our study we include works by a range of authors, follow a testing regime based on withheld samples, and draw on the frequencies of words used, providing a large feature set from which to select markers for our classifier.

For research scholars of these plays, the date of first performance is generally the most informative for a chronology. The impact of the work on audiences and other writers begins with the first performance. In a fast-moving, competitive commercial environment, it can be assumed that composition occurred close to the date of first performance. The date of composition may seem a more logical starting point, but it would have to take account of spans of time: the first creative impulse might be many years prior to its realisation, the work might be drafted and then put aside for years, and so on. The date of printing, though usually easy to ascertain, is not as useful as the date of first performance since it is often clearly widely separated from the date of the first performance, as with the eighteen plays associated with Shakespeare which were first printed in the Folio of 1623, seven years after Shakespeare's death in 1616.

The date of first performance can sometimes be fixed with certainty because a performance is mentioned, and mentioned as the first, in an official document, a reliable personal diary, or in a printed work. There are also some records kept by theatre managers which helpfully record the date of performances. In most cases, however, the best we can do is determine a date we can be reasonably certain is the earliest possible, another which we can be reasonably certain is the latest possible, and then a single year which can be hazarded as a best guess.

If we could devise a method to assign a date of first performance from internal evidence, using the evolution of style, for instance, as a continuum along which to place a given work, this would provide firmer foundations for the literary history of the drama of the time. The study in this paper is the first to offer a prediction of Shakespeare-era play dates based on internal evidence, validated by test samples, and extending beyond Shakespeare works.

In the present paper we return to the question of chronology, sample widely in Shakespeare-era plays, focus on language features, and aim to construct a state-of-the-art mathematical model that provides estimated dates of first performances of plays based on **the individual probability of word use in them**. We take advantage of advances in existing and novel multivariate regression to build an accurate model with a small set of probabilities of individual word uses, thus limiting dimensionality and simplifying the task of understanding the mechanism in linguistic and stylistic terms. We estimate the reliability of our model both for training and for test data.

We introduce a new memetic algorithm-based regression approach based on a continued fraction representation that produces analytic functions. We analyse its performance against other existing implementations of machine learning approaches for regression methods. The memetic algorithm has been chosen as a metaheuristic optimisation method to address the mixed non-linear optimisation problem with the joint need to select the best variables to create a mathematical model. We note however, other researchers may eventually choose other metaheuristics for the non-linear optimisation part such as the one proposed in Abualigah et al. (2021) and others surveyed in that work.

In particular, our contributions are as follows:

- We introduce in the literature a new dataset to predict the year of the first performances for the dramas during the Shakespearean era. The detail of the collection, curation and processing of the data is presented in this work.
- We introduce a feature selection method based on the Lasso regression where a piecewise linear regression model is fitted over the data. The top 50 features from multiple executions of the Lasso regression are selected as the feature subset (each iteration used a random subset of the data containing 80% of the samples).
- We extend the continued fraction regression method presented in Moscato et al. (2021) which represents mathematical expressions. We utilised two advantages of this method: feature subset selection and optimising the coefficients by means of an iterative approach.
- We test the proposed `iter-CFR` approach and the 10 state-of-the-art implementations of regression methods on out-of-domain works and show that `iter-CFR` performs very creditably in this application.
- We offer for the first time a method for prediction of Shakespeare-era play dates based on internal evidence, validated by test samples, and extending beyond Shakespeare to the works of his contemporaries.

Following this introductory section, we organise the rest of the paper as follows: Section 2 provides a detailed description and processing of the dataset followed by the feature selection approach. We also illustrate on the proposed Iterative Continued Fraction Regression (iter-CFR) used in this contribution. Section 3 presents the experimental settings for the proposed iter-CFR and other state-of-the-art methods, and the results of various methods along with statistical tests. In Section 4, we present an in-depth analysis of the obtained results and a discussion of findings. Finally, Section 5 concludes the paper with a discussion of results and interesting new directions for further research.

2. Materials and methods

2.1. Dataset of 285 plays

For this study, we have used a collection of 285 English plays from the sixteenth and seventeenth centuries assembled by our two experts in the field (Hugh Craig and Gabriel Egan) which are part of an ongoing project on stylistic aspects of Folio versions of Shakespeare plays. This collection is a selection from the surviving printed and manuscript play versions from the period, with a bias towards original plays which had been performed by a professional company, rather than translations, plays written for school, university and Inns of Court productions, or for readers as opposed to live audiences. There are 223 plays attributed to a sole author; 53 different playwrights are represented in this group. In addition, there are 26 multi-author plays and 36 plays of uncertain authorship. Craig and Egan took the earliest printed version as the basis for the machine-readable texts, except where this version is a manifestly corrupt one, as when it is obviously missing large sections or has extensive garbled content. In some cases, we included alternative versions of the plays, bearing in mind the scholarly interest in alternative Shakespeare versions in particular. Using early versions is preferable to using more recent ones since they have not been subject to modern editing, but this choice means that the spelling is variable. Spelling was not standardised in England until the late seventeenth century and before that multiple variant spellings were tolerated – perhaps hardly noticed – even within a single work. This creates difficulties for statistical methods based on word counting. The proliferation of variant spellings in these works is considerable, and confounds the expectations of anyone used to modern standardised spelling. de Grazia and Stallybrass (1993) found fourteen different spellings of the word ‘one’ in printed works from this period. The latitude in manuscript works is wider still. Jackson found sixteen spellings in a short manuscript which were not repeated anywhere in a large corpus of sixteenth and seventeenth century printed works (Jackson, 2007b). Many words that are distinct in modern English overlapped in spelling in early modern English. The spelling ‘weeke’, for instance, was used for the different senses ‘weak’, ‘week’ and ‘wick’; the forms ‘travel’ and ‘travail’, ‘hart’ and ‘heart’, and ‘metal’ and ‘mettle’ were interchangeable (Craig & Whipp, 2010, pp. 37–38). For these reasons, we modernised and standardised the spelling in the texts, using the Variant Detector (VARD) 2 software[1] (Baron & Rayson, 2008; Baron et al., 2009) which offers assistance by prompting the user with probable modern equivalents and allowing global changes where the user feels confident there is only one possible modern equivalent for all the instances of a variant spelling in a work. (Fig. 1, below, shows the compression in word types in a section of the corpus that this step in pre-processing caused.)

We also marked up the works in the Text Encoding Initiative (TEI) P4 format, which uses a customised set of XML tags chosen to suit textual matter, so that stage directions, speech prefixes, prefaces, dedications and other non-dialogue material is identified and can be programmatically excluded from word counts.

The standardisation of spelling and parsing of text into dialogue and other materials is laborious, and no comprehensive collection of texts prepared in this way is available, so 285 texts is a large collection compared to other studies apart from those using raw texts based on the Optical Character Recognition of digitised page images and machine-only standardisation and parsing, where a considerable volume of error is encountered (Hill & Hengchen, 2019).

The largest comparable open access manually curated collection of Shakespeare-era plays known to us is the first two components of the Enhanced Shakespeare Corpus (ESC). These include 36 Shakespeare plays and 46 plays by other authors. The ESC has a third, much more extensive component, including many more plays, as well as works in other text types, but the spelling standardisation in this part was carried out programmatically, and those responsible warn that this produces a lower level of reliability than manual standardisation by humans (Murphy, 2019).

Using XML tags, we also marked a subset of words for part of speech so as to separate different uses of some grammatical words. These tags enable us to count instances of “that” as either conjunctions (as in “she said that she would”), relatives (“the book that I left”), or demonstratives (“see that sword”), for instance. In all 19 grammatical words are marked in this way, yielding 48 separate word-forms for counting. The effect of this separation of some homographs, along with the impact of standardising spelling, is illustrated in Fig. 1.

2.1.1. Metadata

We use a standard reference work, the multi-volume *Catalogue of British Drama 1533–1642* (Wiggins & Richardson, 2012/2018), for dates of the first performance. This work offers a single best guess date for first performance for each play based on the latest theatre-historical investigations.

2.1.2. Data availability

After acceptance of this manuscript for publication, the complete dataset will be provided via the UCI Machine Learning library.

2.2. Dataset being used for training

Note that we refer to the plays in the dataset as “samples”, and the frequencies of the words appearing in those plays as “features”. The goal here is to use these frequencies to determine the year each play was first performed in public. A standard best guess for date of best performance is also included in the dataset and is used for training our algorithm and measuring our accuracy. It is worth noting that the year a play was first performed is usually earlier than or the same as its year of publication, but need not be: a play may be published before being performed.

We next examined the distribution of the plays in date ranges, to check for thinly populated ranges. We used the common formula of the square root of N to establish bin ranges, giving us seventeen bins after rounding up to the nearest integer. Four bins covering the years 1587–1611 each contained more than fifty plays, whereas none of the eight bins of earlier plays contained more than ten plays, and the best-populated of the five later bins contained just 21 plays (Fig. 2). We decided to concentrate on the period 1585–1610 and created a new dataset containing only the 181 plays from this date range. The new set includes 135 single-author plays by 36 individual authors, 17 multi-author plays and 29 plays of uncertain authorship. The set of word types appearing in these plays has size 51 256, or 51 183 if the word types that can serve as different parts of speech are each counted once rather than counted once for each of those functions. As the dataset contains a large number of features, we have to apply some feature selection methods to reduce the dimensionality of the data. We chose to use the full range of words available, avoiding the exclusion of ‘stop words’ that is common in text mining to economise on computer resources and focus on rarer words. We used dictionary-type headwords, including inflected forms, rather than lemmas, in order to retain the extra stylistic information they carry.

[1] http://ucrel.lancs.ac.uk/vard.

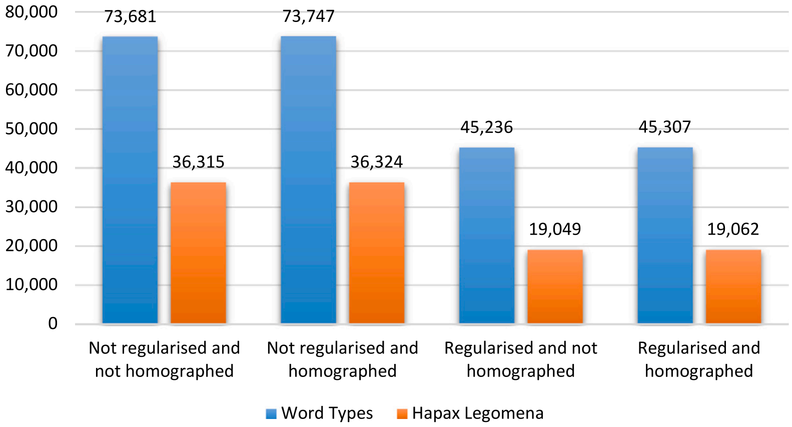


Fig. 1. Word types and hapax hegomena in 143 Plays. In these plays, a subset of the corpus, the mark-up of the text allows us to retrieve the state of the text before regularisation and the tagging of homographs. Marking a select list of homographs makes only a small difference in totals. Hapax legomena (word types represented in the corpus by only one instance) are half the total in regularised text, and less than half in regularised text. After regularisation, the remaining word types are three-fifths of the original total and the remaining hapax legomena are around half the original total.

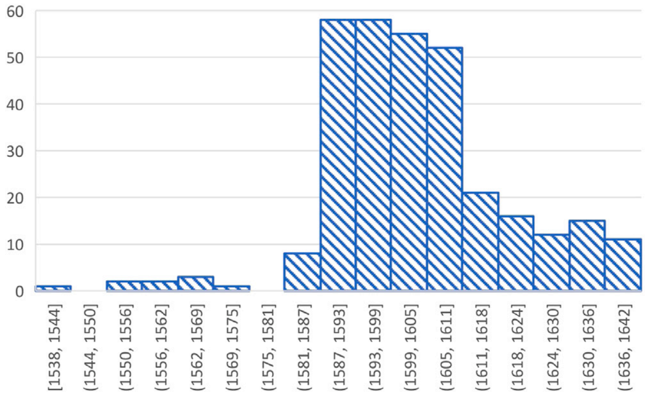


Fig. 2. Histogram showing the number of plays produced in each date range within the years 1538–1642. The majority of plays were first performed in the period of 1585–1610, so this is the range on which samples are extracted for training in our quest to find mathematical models.

2.3. A note on feature selection

Our task is to find a vector – a summed weighting of the selected features we count – which gives the closest approximation to the date variable with the smallest possible set of features or word-variables. Since the binary Min k-Feature Selection problem is NP-complete and also W[2]-complete (Cotta & Moscato, 2003), it is unlikely that either a polynomial-time algorithm or a fixed-parameter tractable algorithm can be found for this problem. This means that the selection of optimal sets of features for multivariate regression analysis needs to be done with some other external heuristic technique that selects them, iteratively trying combinations that lead to regression models with a progressively closer fit. Two approaches used in this study are discussed next: Lasso regression and the memetic algorithm for continued fraction regression.

2.3.1. Lasso regression

The subset of words chosen to be included in the model was determined using the Lasso regression analysis. The Lasso is a well-known regularisation technique for linear regression that identifies a sparse set of features (Santosa & Symes, 1986; Tibshirani, 1996). Given a linear model $y = X\beta$, where y is the dependent variable, X is a matrix with each column being an independent variable, and β is the vector of p parameters, along with a regularisation parameter λ, Lasso regression minimises the sum of squares errors of n samples with Lasso penalty in following objective function:

Table 1
A table containing all 14 words that appeared in at least one of 100 Lasso regression trials using the dataset containing the chosen 181 plays and all 16 383 words. The number in parentheses is the percentage (pct.) of trials each word appeared in at least once.

Word	pct.	Word	pct.	Word	pct.
'and'	(100%)	'a'	(100%)	'you'	(100%)
'thou'	(100%)	'it'	(100%)	'is'	(99%)
'your'	(99%)	'thy'	(96%)	'sir'	(56%)
'my'	(15%)	'that[conjunction]'	(14%)	'to[infinitive]'	(8%)
'the'	(7%)	'of'	(1%)		

$$\sum_{i=1}^{n}\left(y_i - \sum_{j} X_{ij}\beta_j\right)^2 + \lambda \sum_{j}^{p} |\beta_j|. \tag{1}$$

To ensure stability, a Lasso regression model was fitted on a random subset of the data containing 80% of the samples, repeated over 100 independent trials. The words that appear in the Lasso models are shown in Table 1. A regularisation parameter $\lambda = 1$ was used.

Each word that appears in at least 90% of Lasso trials is in the top 50 words whose frequency of occurrence has the highest Pearson correlation with performance year. Thus, these 50 words are a useful subset of features to deploy in further analysis.

2.3.2. Continued fraction regression

In 2019, a regression approach based on 'Continued Fraction' (CFR) was proposed; it views multivariate regression as a non-linear optimisation problem and the authors used a memetic algorithm to find approximations to the unknown target functions from experimental data (Sun & Moscato, 2019). Memetic algorithms are a population-based approach to solve computational problems that are posed as optimisation tasks and have been heavily used for other data analytics in combinatorial optimisation problems (Gabardo et al., 2020; Haque & Moscato, 2019; Zaher et al., 2019) and that are also showing impressive results for non-linear regression problems (Moscato, Haque et al., 2020; Moscato, Sun et al., 2020; Moscato et al., 2021; Sun & Moscato, 2019) and other machine learning problems (Moscato & Mathieson, 2019).

Continued fractions are a type of mathematical expression consisting of the sum of an integer and a quotient, whose denominator is again the sum of an integer and a quotient. These expressions may be finite or infinite (Sun et al., 2019). Euler's continued fraction formula allows us to write the sum of products as a continued fraction, as follows:

$$x = a_0 + a_0a_1 + a_0a_1a_2 + \cdots + a_0a_1a_2 \dots a_n$$
$$= \cfrac{a_0}{1 - \cfrac{a_1}{1 + a_1 - \cfrac{a_2}{1 + a_2 - \cfrac{\ddots}{\ddots \, \cfrac{a_{n-1}}{1 + a_{n-1} - \frac{a_n}{1+a_n}}}}}}.$$

This simple yet powerful equation displays a general continued fraction approximation for the ratio of two higher-order polynomials. We can use the same idea to approximate a function $f(\boldsymbol{x})$ by replacing each a_i and b_i with other functions of $\boldsymbol{x}$. Sun and Moscato (2019) proposed that we can approximate the "target function" of a multivariate regression problem, given a set of examples, and that it can be expressed as a multivariate function $f : \mathbb{R}^n \to \mathbb{R}$ of the form:

$$f(\mathbf{x}) = g_0(\mathbf{x}) + \cfrac{h_0(\mathbf{x})}{g_1(\mathbf{x}) + \cfrac{h_1(\mathbf{x})}{g_2(\mathbf{x}) + \cfrac{h_2(\mathbf{x})}{g_3(\mathbf{x}) + \ddots}}}. \tag{2}$$

Then we have $g_i(\mathbf{x}) \in \mathbb{R}$ for all integers $i \geq 0$, and each function $f_i : \mathbb{R}^n \to \mathbb{R}$ is associated with a different array $\mathbf{a_i} \in \mathbb{R}^n$ and with a different constant $\alpha_i \in \mathbb{R}$:

$$g_i(\mathbf{x}) = \mathbf{a_i}^{\mathrm{T}}\mathbf{x} + \alpha_i, \tag{3}$$

For each function $h_i : \mathbb{R}^n \to \mathbb{R}$ we also have a different array $\mathbf{b_i} \in \mathbb{R}^n$ as well as a different constant $\beta_i \in \mathbb{R}$:

$$h_i(\mathbf{x}) = \mathbf{b_i}^{\mathrm{T}}\mathbf{x} + \beta_i. \tag{4}$$

The "depth" of a continued fraction refers to the number of "sub-fractions" in the overall fraction. For example, the depth 0 form of the fraction in Eq. (2) would be $x = g_0(\mathbf{x})$, the depth 1 form would be $x = g_0(\mathbf{x}) + \frac{h_0(\mathbf{x})}{g_1(\mathbf{x})}$, and so on.

It is often useful to represent continued fractions in a way that explicitly states each numerator and denominator, particularly when a continued fraction is difficult to visualise in the standard representation. To do this, we simply state the expression for each $g_i(\mathbf{x})$ and $h_i(\mathbf{x})$ term. To illustrate this, we will use the concrete example of the Mills ratio. This value is used in probability and its definition is shown in Eq. (5), where $D(\mathbf{x})$ and $P(\mathbf{x})$ are the distribution and probability density functions, respectively (Weisstein, 2021).

$$m(\mathbf{x}) = \frac{1 - D(\mathbf{x})}{P(\mathbf{x})} \tag{5}$$

This quantity can be approximated by the following continued fraction, which appears in the two equivalent forms in Eqs. (6) and (7) (Gasull & Utzet, 2014). We will use both representations throughout this paper.

$$f(\mathbf{x}) = 0 + \cfrac{1}{\mathbf{x} + \cfrac{1}{\mathbf{x} + \cfrac{2}{\mathbf{x} + \cfrac{3}{\mathbf{x} + \ddots}}}} \tag{6}$$

$$\begin{array}{lll} g_0(\mathbf{x}) = 0 & h_0(\mathbf{x}) = 1 & g_1(\mathbf{x}) = \mathbf{x} \\ h_1(\mathbf{x}) = 1 & g_2(\mathbf{x}) = \mathbf{x} & h_2(\mathbf{x}) = 2 \\ g_3(\mathbf{x}) = \mathbf{x} & h_3(\mathbf{x}) = 3 & g_4(\mathbf{x}) = \mathbf{x} + \ddots \end{array} \tag{7}$$

In situations like the one we are addressing in this study, finding a multivariate regression of a single target variable, we need to approximate the unknown target function given a dataset $S = \{(\mathbf{x^{(i)}}, y^{(i)})\}$, i.e. a set of pairs of an unknown multivariate target function $f : \mathbb{R}^n \to \mathbb{R}$ on which the image values are known (ideally, with no uncertainties). In general, better generalisation outcomes are expected if we identify the subset of the variables of $\mathbf{x}$, which are more relevant for prediction. Minimisation of the MSE on the values of the training set S is used to identify the sets of coefficients $\{\mathbf{a_i}\}$, $\{\mathbf{b_i}\}$, $\{\alpha_i\}$, and $\{\beta_i\}$. One of the advantages of our method is that, since it selects subsets of variables as well as adapting the coefficients in the formula, it may lead to insights about the classes of variables that are more relevant for prediction. We are going to utilise that advantage of CFR in this contribution.

2.4. Memetic algorithm for iterative continued fraction regression

Memetic Algorithms (MAs) is a type of population-based approach used for solving complex problems which are generally posed as an optimisation task with one or multiple objectives and constraints. In these methods, we start by initialising the search using a "population" of potential solutions (generally feasible solutions of the problem at hand), which are evaluated based on some heuristic (such as Mean Squared Error, or MSE). The fittest ones, according to this heuristic, are modified and combined to generate a new population of solutions, for which this process is repeated. MAs then follow similar process to other evolutionary type of algorithms and heuristics but they are characterised by the inclusion of an additional step of individual optimisation. Each solution is then independently improved using the given heuristic before the "recombination" operation processes them. This increases, on average, the accuracy of solutions as well as the diversity of the new generation (Moscato et al., 2011; Neri et al., 2012).

The CFR algorithm has a number of default parameters, which we will describe here. Unless stated otherwise, these were the parameters used for each experiment. No normalisation is done on the data. The objective function, used to measure the accuracy of each potential solution, is the MSE by default. The penalty in the fitness function (the "delta"), is 0.10 with this dataset. A larger value of this parameter prevents overfitting to the data in order to accommodate outliers. The program runs for 200 generations, where each generation is a new population of the potential solutions. The mutation rate is 0.10, which affects how much the potential solutions are altered at each stage. The root of the population tree, which determines which potential solutions will be generated, gets reset if the MSE does not improve after five generations. The local search algorithm is performed at each generation to improve current potential solutions. At the local search step, the Nelder–Mead algorithm is run four times, with each run producing at most 250 generations, and the algorithm resets after ten consecutive generations without improvement. Local search optimisation is run serially. All data samples are used in the local search.

The depth of the continued fraction solution generated begins at 0. Once we have the depth 0 solution (using a random function as its initial solution), we use that as the initial solution to find a new solution of depth 1. This process is repeated until we reach a solution with MSE worse than that at the previous depth. At this point, we take the solution of the previous depth to be our final solution. This approach of iteratively increasing the depth of the CFR algorithm as long as the fitness improves is referred to as iterative continued fraction regression (abbreviated to `iter-CFR`).

2.4.1. An univariate example of the performance of the memetic algorithm for regression using continued fractions

As an example of the power of the memetic algorithm to do a regression of non-linear functions, we show results on approximating an unknown highly non-linear target function, namely $1 + Sin(x)/x$ on the interval $[-10, 10]$ and with an added normally distributed random noise with mean 0 and standard deviation of 0.01. Fig. 3 shows the approximation found with the memetic algorithm and a continued fraction of depth equal to three. We have instructed the algorithm to make use of the original variable x and the metafeature x^2.

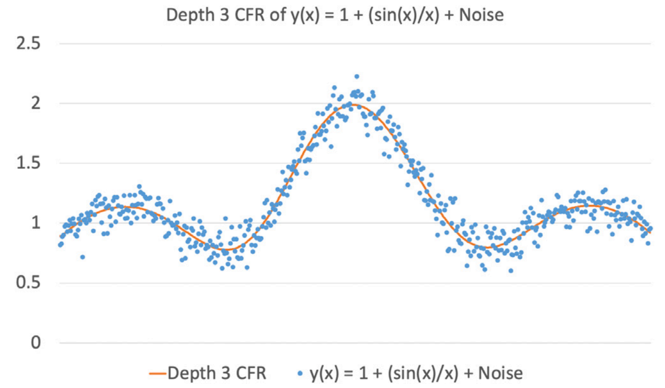


Fig. 3. Model produced by CFR algorithm at depth 3 on a benchmark dataset produced by adding noise to $y(x) = 1 + \frac{\sin x}{x}$ on 500 points equally separated in the interval $x \in [-10, 10]$. The noise was normally distributed with mean 0 and standard deviation 0.1. The memetic algorithm found a truncated continued fraction approximation of $y(x)$ (i.e. $f(x)$ as given by Eq. (2)) having a Mean Squared Error of 0.00968963.

Using the notation for $f(x)$ given by Eq. (2), we can then write:

$$\begin{aligned}
g_0(x) &= 1.29492 - 0.0162327\,x^2,\\
h_0(x) &= 33.8386 - 4.84268\,x^2,\\
g_1(x) &= 16.4545 + 0.580148\,x - 3.54912\,x^2,\\
h_1(x) &= -98.6612 - 3.87476\,x - 17.5014\,x^2,\\
g_2(x) &= -6.07812 - 0.0996804\,x^2,\\
h_2(x) &= 51.4633 - 0.00939706\,x + 2.38741\,x^2,\\
g_3(x) &= 16.9629 - 0.134414\,x^2.
\end{aligned} \tag{8}$$

3. Experiments with 10 regression techniques well-known in machine learning

3.1. Computational environment

All experiments were conducted on a computer equipped with a 6 Core Intel(R) Core(TM) i7-9750H CPU of 2.60 GHz clock-speed and 16.0 GB of RAM running on 64-bit Windows 10 Operating System. We used Python 3.7.6 as the execution environment of the regression methods.

3.2. Dataset and parameter settings of the regression methods

To test the performance of many machine learning algorithms, we employed a dataset consisting of 181 plays and, as variables, the percentages of occurrences of the 50 words having the highest Pearson correlation with the performance year of the play (ranging from 1585 to 1610). To ensure reproducibility, we employed the implementations of 11 machine learning regression methods – comprising a set of 9 regressors from the popular *Scikit-learn* machine learning library (Pedregosa et al., 2011), one from *XGBoost* (Chen & Guestrin, 2016) and the iterative Continued Fraction presented in this paper – to predict the year using 100 randomised runs with 80–20 training/test splits. The names of the regression methods studied are shown in Table 2.

In our initial testing we found that `krnl-r`, `l-svr`, `mlp` and `sgd-r` performed poorly in terms of the MSE score. We used the `‘squared_epsilon_insensitive’` loss function for SGD Regressor (with `learning_rate = ‘adaptive’`) and Linear SVR. This loss function applies the squared penalty by ignoring any residuals $(y - y') > \epsilon$ and linear penalty in the other case. It is computed as $Loss = max\{0, |y - y'| - \epsilon\}^2$, where $\epsilon = 0.1$, y and y' are the actual/target and predicted value. For the Kernel Ridge, we used the `‘polynomial’` kernel with degree 3. As the solver in `mlp` and `sgd-r` were not converged with the default parameter value for maximum iteration, `‘max_iter’`, we set the value as 25 000 and 100 000, respectively. We kept the default parameters of other machine learning regression algorithms unchanged. For the convenience we have tabulated these parameter values in Table 2.

3.3. Experimental results

We used Mean Squared Error (MSE) as the evaluation metric to compare the performance of the regression methods. This metric quantifies the goodness of the method by averaging the squared error of prediction (y') with the actual/target (y) value of all n samples in the dataset as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2. \tag{9}$$

Table 3 shows the descriptive statistics of the MSE scores obtained for 100 runs by the regressors. Here, we can see that `grad-b` obtained the best average MSE score of 0.08 for the training data. The best average testing MSE value of 15.65 is obtained by `ada-b`. However, `grad-b`, `l-regr`, `rf` and `xg-b` also obtained nearly the same value of average MSE score in testing (ranging from 16.03 to 16.58). The `iter-CFR` is the next closest method to that group of regressors in terms of the average test MSE of 21.25. Some other regressors performed significantly worse.

In addition to the summary table, we show in Fig. 4 the box plot for the testing MSE scores of the regressors obtained for 100 runs. From this plot, it can be seen that the range of MSE scores is wide. To better understand the MSE scores obtained by a good subset of regressors, we show the zoomed plot as an inset for Test MSE scores up to 100. From the inset we can see that `ada-b`, `rf`, `xg-b`, `grad-b` and `l-regr` exhibited similar results to `iter-CFR` as the closest performing regressor to the group.

3.4. Statistical comparison of the rankings of regression methods

We conducted the Friedman test for repeated measures (Friedman, 1937) to validate the significance in results obtained by different regression methods for 100 independent runs. We used the ranking of the methods based on their MSE scores obtained for the test set to help us determine if the experiment's techniques are consistent in their generalisation performance. The statistical test found $p = 2.74845 \times 10^{-176}$ which rejected the null hypothesis of *all the models perform the same* and we proceeded with the post-hoc test.

The Friedman's post-hoc test on the ranking of 11 regressors computed for the test *MSE* scores was obtained for 100 runs on 80–20 split. In Fig. 5 the p-values obtained for the test are plotted as a heatmap. It is noticeable that there exist *no significant differences* (symbolised as NS in Fig. 5) in performances of `iter-CFR` with `l-regr` and `grad-b`.

Additionally, we generated the Critical Difference (CD) diagram proposed by Demšar (2006) to visualise the differences among the regressors regarding their median ranking. The CD plot uses the Nyemeni post-hoc test and hence the results may differ from the results obtained from Friedman's post-hoc test, and it places the regressors on the x-axis of their median ranking. It then computes the critical difference of rankings between them. It connects those methods which are closer than the critical difference with a horizontal line denoting that them as statistically non-significant.

In Fig. 6 we plot the CD graph using the implementation from the Orange data mining toolbox (Demšar et al., 2013) in Python. The Critical Difference is found to be 1.397. We can see that there are no significant differences in the median rankings of `rf`, `xg-b`, `l-regr` and `grad-b` with the top-ranked `ada-b`. The ranking of `iter-CFR` is statistically not similar to any other regressors; however, it is next to the group of regressors that ranked first.

Table 2
The regression methods and their parameter values used in this study. For most of the regressors, we have kept the default parameters unchanged from their source (Scikit and XGBoost libraries for Python language); however, we have tabulated any changes to the parameter values in the table.

Regression method	Parameter settings of the method
AdaBoost (`ada-b`)	Scikit default
Gradient Boosting (`grad-b`)	Scikit default
Kernel Ridge (`krnl-r`)	`kernel='polynomial'`
Lasso Lars (`lasso-l`)	Scikit default
Linear Regression (`l-regr`)	Scikit default
Linear SVR (`l-svr`)	`loss='squared_epsilon_insensitive'`
MLP Regressor (`mlp`)	`max_iter=25 000`
Random Forest (`rf`)	Scikit default
Stochastic Gradient Descent (`sgd-r`)	`loss='squared_epsilon_insensitive'`, `learning_rate='adaptive'`, `max_iter=100 000`
XGBoost (`xg-b`)	XGBoost default
Iterative Continued Fraction Regression (`iter-CFR`)	`max_depth = 20`, Δ=0.10, `generations=200`, `mu_rate=0.10`, `reset_pop=5`, `NM_run=4`, `NM_gen=250`, `NM_reset=10`

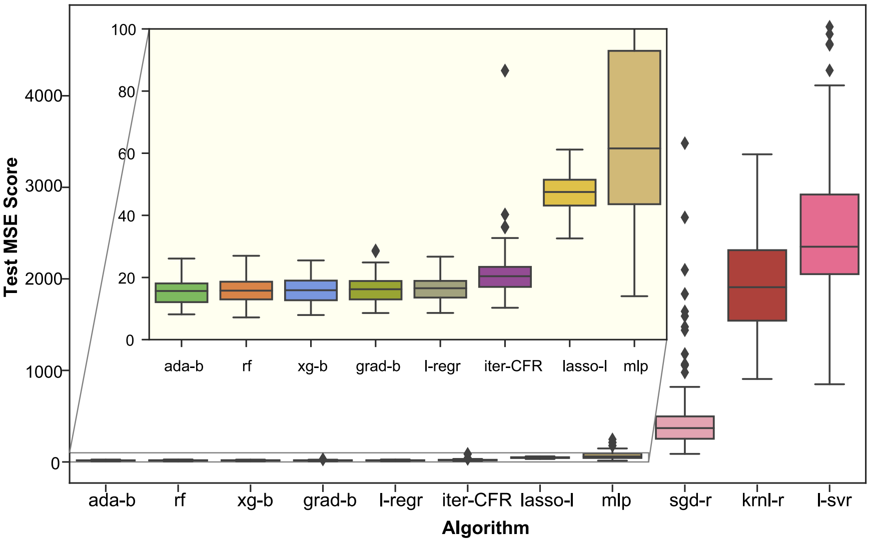


Fig. 4. Bar and whisker plot showing the MSE scores of regressors obtained for 100 runs in the testing sets. As the MSE scores of the regressors vary in a wide range, we show the subset of regressors having the upper bound of MSE score of 100 in the inset.

Table 3
Descriptive statistics for the 100 runs of regressors on the 50 most correlated words of 181 plays with 80-20 training/test splits.

Regression method	Training MSE score			Testing MSE score		
	Avg.	Med.	std.	Avg.	Med.	std.
ada-b	2.88	2.85	0.25	**15.49**	**15.65**	4.07
grad-b	**0.09**	**0.08**	**0.02**	16.28	16.22	4.24
iter-CFR	12.14	11.93	1.91	21.25	20.41	8.56
krnl-r	1722.95	1726.55	28.72	1974.71	1908.89	559.22
l-regr	6.13	6.17	0.56	16.56	16.54	3.97
l-svr	1830.26	1835.50	31.53	2545.38	2351.11	781.39
lasso-l	47.99	48.01	1.54	47.67	47.52	6.04
mlp	5.69	5.46	1.91	73.91	61.55	42.60
rf	2.23	2.24	0.18	15.93	15.77	4.10
sgd-r	266.58	197.66	280.18	520.90	369.56	529.74
xg-b	0.54	0.55	0.08	16.11	15.89	**3.95**

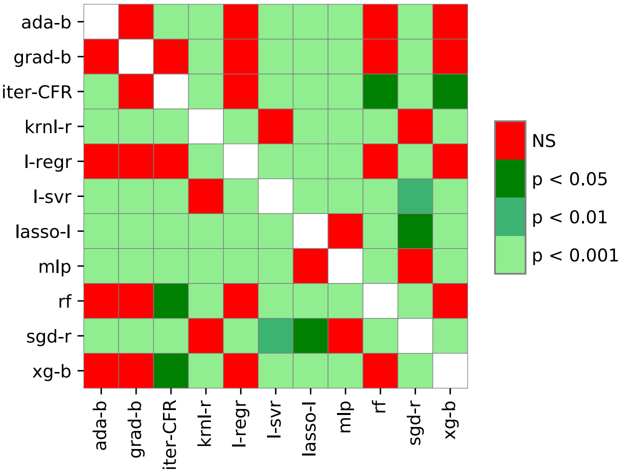


Fig. 5. Heatmap showing the statistical significance levels of p-values obtained by the Friedman post-hoc test.

4. Discussion

Table 4 shows the number of times a regressor was ranked first and its minimum and maximum ranking for 100 runs on the dataset. We can see that the same set of regressors, the ones that are ranked first in the CD plot of Fig. 6, have a maximum ranking of 6. In terms of the number of times a regressor was ranked 1th, `l-regr` has the highest value (32 times). The proposed `iter-CFR` was ranked first 11 times in 100 runs, which is the same as the value for the regressor `rf`. The set of regressors consisting of `krnl-r`, `l-svr`, `lasso-l`, `mlp` and `sgd-r` were never ranked first in the 100 runs. Moreover, `l-svr` exhibited the worst ranking with the minimum ranking of 10 out of 11 regressors.

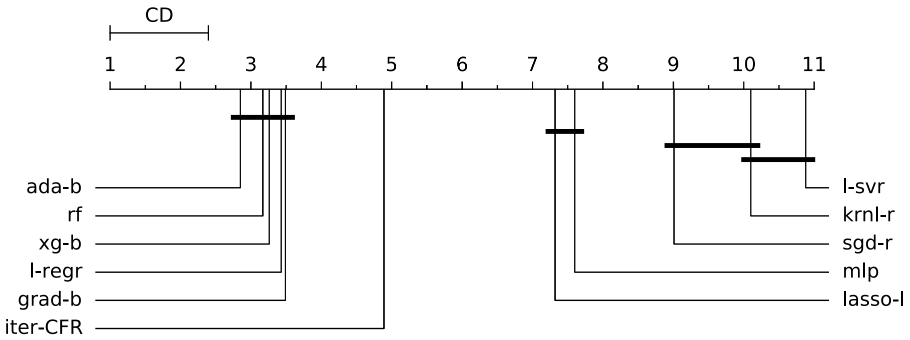


Fig. 6. Critical Difference (CD) plot showing the statistical significance of rankings achieved by the regression methods.

Table 4
Number of times each regression method came first and the value of maximum and minimum ranking achieved for 100 runs on the dataset.

Regressor	#1st	Rank (min, max)	Regressor	#1st	Rank (min, max)
l-regr	32	(1, 6)	mlp	0	(2, 9)
ada-b	19	(1, 6)	lasso-l	0	(6, 8)
grad-b	14	(1, 6)	sgd-r	0	(8, 11)
xg-b	13	(1, 6)	krnl-r	0	(9, 11)
rf	11	(1, 6)	l-svr	0	(10, 11)
iter-CFR	11	(1, 7)			

Table 5
20 most frequent words showing the number of times (x) each has appeared in 100 models of iterative Continued Fraction Regression (`iter-CFR`) using the 50 words whose frequencies are most correlated with date.

Word	x	Word	x	Word	x	Word	x
‘ah’	66	‘that[conjunction]’	54	‘own’	49	‘business’	35
‘goodness’	60	‘your’	53	‘women’	48	‘does’	34
‘known’	59	‘beseems’	52	‘aside’	38	‘wherein’	32
‘therefore’	56	‘for[conjunction]’	52	‘content’	38	‘thy’	25
‘like[preposition]’	55	‘has’	50	‘thou’	36	‘threats’	25

4.1. Looking in depth at the best model of iterative continued fraction

We look at the best `iter-CFR` model found in the 100 repetitions of the experiment. The model which fitted the training data best had a training MSE of 7.63661 and produced a test MSE of 14.4752. The continued fraction model is at depth = 0 and it is as follows:

$$\begin{aligned} f(x) =& 1608.8 + 13.0358 \times (has) - 16.5958 \times (ah) - 7.10359 \\ & \times (for[conjunction]) - 7.00464 \times (that[conjunction]) - 4.7321 \\ & \times (thy) + 54.2334 \times (known) - 400.11 \times (beseems) + 31.6509 \\ & \times (women) - 104.339 \times (wherein) - 104.749 \times (aside) + 1.39045 \\ & \times (that[demonstrative]) + 1.27499 \times (mighty) + 110.119 \\ & \times (goodness) - 2.8458 \times (a) - 68.8078 \times (saith) - 44.4536 \\ & \times (triumph) + 15.6812 \times (like[preposition]) - 8.98319 \times (words). \end{aligned}$$

12 out 18 of these words, ‘ah’, ‘goodness’, ‘known’, ‘like[preposition]’, ‘that[conjunction]’, ‘beseems’, ‘for[conjunction]’, ‘has’, ‘women’, ‘aside’, ‘wherein’ and ‘thy’ are in the 20 most frequent words (Table 5). In Fig. 7 we show how well the continued fraction model predicts the year for both the training and the testing portions of the data.

4.2. Top 20 words by interpretable models

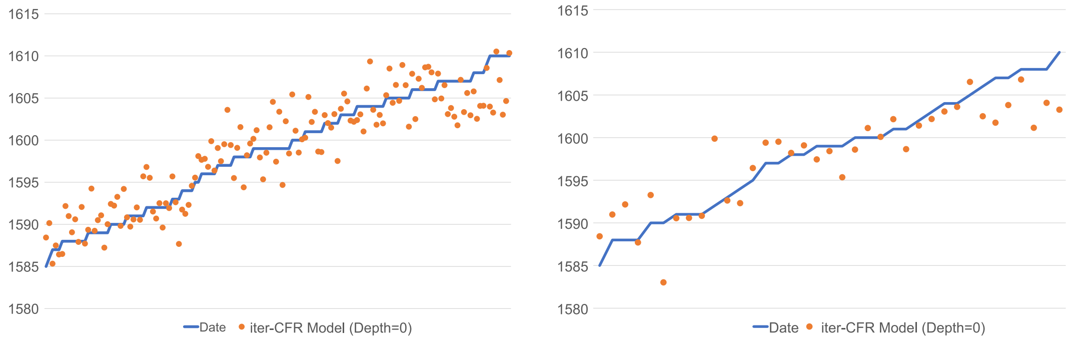
We selected three interpretable regressors (`grad-b`, `xg-b` and `l-regr`) which ranked highly in the CD plot shown in Fig. 6. Then we collected the *feature importance* score of the words given by each of their 100 models. From those scores we selected the top 20 words for each regressor and compared them with the 20 most frequently appearing words from `iter-CFR` in the Venn diagram in Fig. 8 created with an online Venn Diagrams tool developed by Van de Peer Lab.[2] We can observe strong agreement in selecting words by `iter-CFR` and other regressors. Our `iter-CFR` has 13 common words with each of `grad-b` and `xg-b`, and 10 common words with `l-regr`. Among the 20 words of `iter-CFR`, only three words – ‘own’, ‘like[preposition]’, ‘thy’ – did not appear in any of the top 20 words given by other methods. Due to these strong correspondences with the 20 most frequently appearing words found by `iter-CFR` and other regressors, we analyse these words’ roles in `iter-CFR` models in the following section.

4.2.1. The 20 most frequently appeared words in iterative continued fraction models

Fig. 9 shows the percentage of plays in five genre groupings by the period of 1585–1610. “Comedy” is well-represented here; however, “History” plays decline sharply after the third half-decade. We will analyse the association of words with the genre. Table 5 lists the twenty words most often included in the functions which emerge from the CFR process. They are evidently markers of change over time in the style of the plays. The exclamation ‘ah’ is the word-variable most commonly found in the functions, appearing 66 times. Its incidence declines over the period. It has been noticed before in discussions of words used in early modern English drama. The editors of the *Encyclopedia of Shakespeare’s Language* offer it as an example of the way “certain words, meanings, structures, etc. are peculiar to tragedies, comedies or histories, to certain social groups — and to specific periods” (Culpeper et al., 2018; Plescia, 2016, pp. 1). They note that in Shakespeare’s works ‘ah’, which “signal[s] emotional distress or pity”, “is characteristic of the histories, and occurs more than twice as densely in the speech of female characters compared with male”. This word is “used relatively frequently by Shakespeare, compared with his contemporaries, and, despite being characteristic of the histories, is strongly colloquial in flavour, occurring densely in speech-related genres (e.g. trial proceedings)”.

From our study, we can add to this that usage declines in play dialogue in general over the period 1585–1610. As we have seen, the Encyclopedia editors comment that ‘ah’ is unusually common in Shakespeare’s history plays, and we might infer that the change in usage over time can be explained by the fact that in the drama more generally, as well as in Shakespeare’s canon, history plays cluster in the years before 1600, but if we account for the genre effect associated with history plays by looking exclusively at comedies (Fig. 10), there is still a significant negative correlation between date and the probability of ‘ah’ ($r = -0.257$, $p = 0.0005$). For this purpose, the genre of comedy includes plays described in Wiggins and Richardson as “Classical Legend (Comedy)”, “Domestic Comedy”, and “Romantic Comedy”, as well as simply “Comedy”. If we include in a broader History Play

[2] The Venn Diagrams tool can be accessed from the website of the Bioinformatics and Evolutionary Genomics group of Ghent University, Belgium at http://bioinformatics.psb.ugent.be/webtools/Venn/.



(a) Training Performance

(b) Testing Performance

Fig. 7. The result of the best model found for iterative CFR algorithm on the dataset containing only the top 50 words whose frequencies of occurrence have the highest Pearson correlation with performance year. The blue function is the target, and the orange dots are the approximation. (a) Result of the CFR algorithm at depth 0 on the training portion of the data with MSE score of 7.637. (b) Result of the CFR algorithm at depth 0 on the testing portion of the data with MSE score of 14.475.

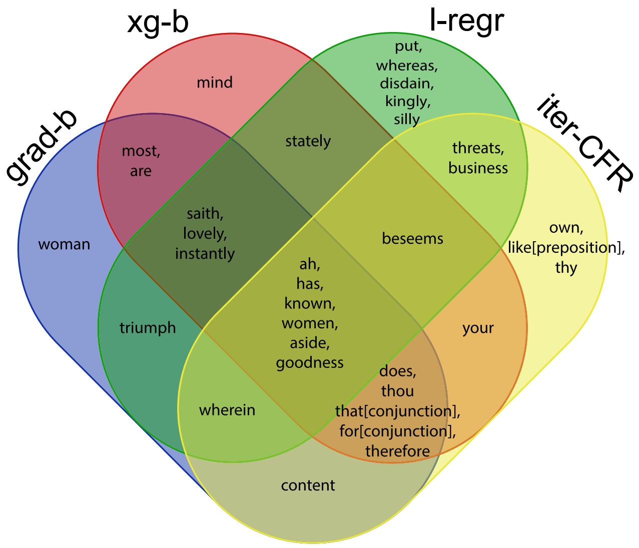

xg-b
l-regr
grad-b
iter-CFR
mind
put, whereas, disdain, kingly, silly
most, are
stately
threats, business
woman
saith, lovely, instantly
beseems
own, like[preposition], thy
ah, has, known, women, aside, goodness
triumph
your
does, thou
wherein
that[conjunction], for[conjunction], therefore
content

Fig. 8. Venn diagram showing the agreement in top 20 words of the models by `grad-b`, `xg-b`, `l-regr` and `iter-CFR`. A subset of six words is common to all four methods: ‘ah’, ‘has’, ‘known’, ‘women’, ‘aside’ and ‘goodness’.

category plays described in Wiggins and Richardson as “Biblical History”, “Classical History”, “Legendary History” and “Pseudo-history”, as well as simply “History”, we get the following percentages of History Plays compared to all plays by half-decade: 1585–89, 30.4%; 1590–94, 31.4%; 1595–99, 37.1%; 1600–04, 20%; 1605–10, 4.2%.

A number of the words in Table 5 are already well known as forms whose incidences were increasing or decreasing in the English language in general at this time as part of overall changes in Early Modern English. The auxiliary verbs ‘does’ and ‘has’ are incoming forms, replacing the outgoing forms ‘doth’ and ‘hath’, respectively. The older forms remained current but became progressively less common. The pronouns ‘thou’ and ‘thy’ are outgoing forms, and the pronoun ‘your’ is an incoming form, part of the larger change whereby ‘thou’ forms in general lost their function as singular second person forms, and ‘you’ forms were increasingly used in for singular as well as plural referents.

Some other words in Table 5 which decline in use in the plays – ‘beseems’, ‘for’ as a conjunction, ‘that’ as a conjunction, and ‘wherein’ – sound archaic to a modern ear and it is plausible that playwrights might use fewer of them over time in the period of our study as a reflection of contemporary usage outside the theatre.

The remaining words have not, to the best of our knowledge, been discussed in the context of language change before. Four of them decrease in incidence over the period: ‘aside’, ‘content’, ‘therefore’ and ‘threats’. The use of ‘threats’ declines significantly in the full corpus ($r = -0.0412$, $p < 0.0001$), but does not also decline in a separate sub-corpus composed exclusively of comedies ($r = -0.0674$, $p = 0.367$), and in this case we might suspect that a genre factor might best explain its power to mark change over time and hence its presence in Table 5. The other three show a highly significant correlation between probability and date in comedies as well as in the full set. Five words not yet mentioned increase in incidence over the period: ‘business’, ‘goodness’, ‘known’, ‘like’ as a preposition, ‘own’, and ‘women’. All of them have a highly significant correlation between probability and date in the comedies sub-corpus.

4.3. Out of domain performances of the model

To test the generalisation capability of the regressors, we conducted an out-of-domain test. For this purpose, we drew 80% data uniformly at random from the set of data with date of plays within 1585–1610 range as a training set. We trained the model on these random samples of training data and tested its generalisation capability on the out-of-domain test data, containing the samples outside the range of 1585–1610. This process is repeated 100 times to get a statistically sound understanding of their performances. The descriptive summary of the regressors sorted by Testing MSE score in ascending order is shown in Table 6. Here we can see that `mlp` has shown the best generalisation performances among 11 methods. Our `iter-CFR` is ranked 3th for the average MSE score obtained in Test set for 100 runs among 11 regressors.

To understand the importance of the words whose probabilities are used as features, we look at the best model of `iter-CFR` on the out-of-domain test. The continued fraction model is given by:

$$f(\mathbf{x}) = g_0(\mathbf{x}) + \frac{h_0(\mathbf{x})}{g_1(\mathbf{x})}$$

where

$$\begin{aligned} g_0(x) =& 1604.17 + 15.4971 \times (has) - 40.1605 \times (therefore) - 5.3539 \\ & \times (thou) + \ 22.5657 \times (own) - 81.8433 \times (stately) - 31.7315 \\ & \times (mighty), \\ h_0(x) =& -75.3675 + 812.856 \times (has) - 1730.8 \times (therefore) + 962.143 \\ & \times (own) - \ 614.681 \times (stately) - 81.0375 \times (mighty), \\ g_1(x) =& 5.99535 + 4348.8 \times (has) + 499.345 \times (therefore) - 1.50877 \\ & \times (thou) - \ 130.705 \times (own) - 99.6367 \times (stately) + 157.17 \\ & \times (mighty). \end{aligned}$$

Interestingly, this `iter-CFR` model is able to obtain a 423.982 MSE score on the out-of-domain test set and uses only six words (‘has’,

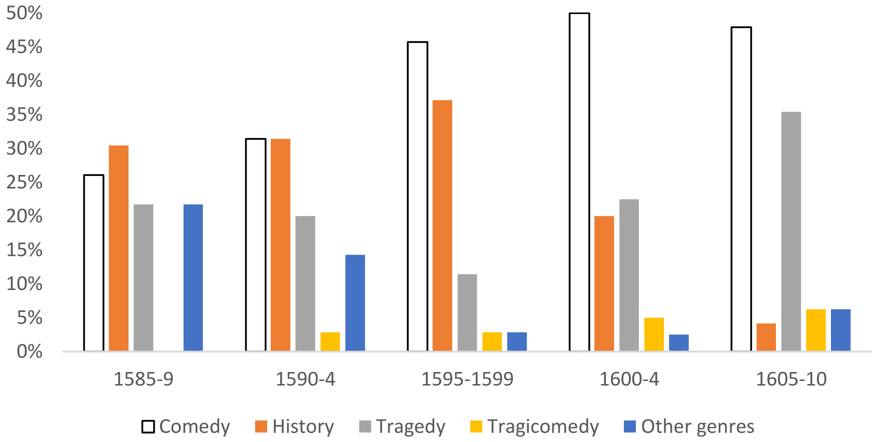


Fig. 9. Percentage of plays in five genre groupings, by half-decade. Comedy is well represented throughout. History plays decline sharply after the third half-decade.

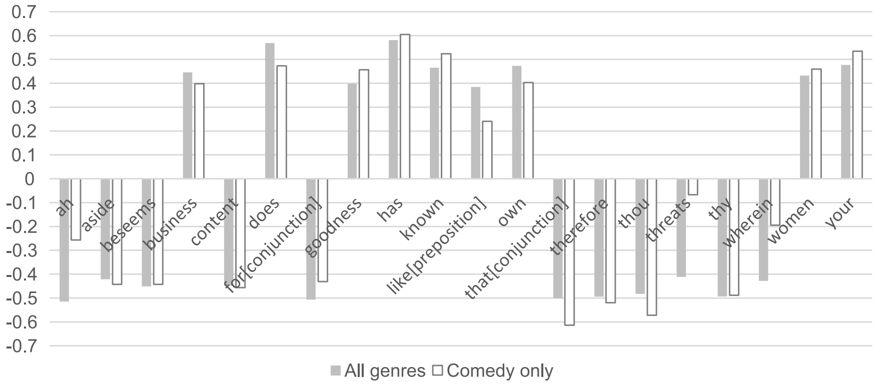


Fig. 10. The Pearson product-moment correlation between date and probability for 20 word-variables found by iterative Continued Fraction Regression models in all genres and in comedies. The correlations are all significant at the $p < 0.01$ level for the all-genres and comedies sets except for the correlation for the word 'threats' in comedies ($r = -0.067$, $p = 0.370$).

'therefore', 'thou', 'own', 'stately' and 'mighty'). Among these six words, only 'stately' and 'mighty' did not appear in top 20 words used by `iter-CFR` tested on the data with date 1585–1610 (presented in Table 5). 'Stately' and 'mighty' both have a strong negative correlation with date in the full set of 285 plays ($r = -0.2403$, $p < 0.0001$ and $r = -0.2759$, $p < 0.001$) but the correlation with date is not significant in the subset of comedies ($r = -0.290$, $p = 0.7564$ and $r = -0.949$, $p = 0.3087$). It is likely that some of the change over time in the probabilities of these words is linked to the replacement of high-scoring genres with lower-scoring genres in the later plays.

5. Conclusions

We analysed the frequency of words most correlated with the date of publication of 181 English plays from the sixteenth and seventeenth centuries (ranging between 1585 and 1610). We employed a set of 11 machine learning methods on the dataset of words with their frequencies to predict the date of the first performance of the plays. In our effort to learn the significance of the words during the Shakespearean era as markers for publication date, we trained each of the machine learning regression methods with an 80% of the data samples taken uniformly at random. We tested the methods' predictive performance on the remaining 20% of the data and repeated this process 100 times, each with a separate set of train and test samples but with the same ratio. AdaBoost, Random Forest, XGBoost, Gradient Boosting, and Linear Regression have shown the best performance in terms of predictive capability. However, most of these models are non-interpretable, in terms of the usage information of the words. The next best performing method, supported by statistical tests, is the iterative Continued Fraction Regression (`iter-CFR`), which has the advantage of offering interpretable models.

Table 6
Descriptive statistics for the 100 runs of 11 regressors trained on the 50 most correlated words with each train set consisting of randomly drawn 80% samples from 181 plays (dated in 1585–1610) and tests of generalisation capability using the out-of-domain (plays with a date outside of the 1585–1610 range) test data. The regressors are sorted in ascending order of their average testing MSE score.

Regression method	Training MSE score			Out-of-domain testing MSE score		
	Avg.	Med.	std.	Avg.	Med.	std.
mlp	5.270	5.118	1.623	377.685	373.270	43.403
l-regr	6.130	6.167	0.542	443.024	441.430	19.799
iter-CFR	14.479	14.489	3.163	451.108	446.530	41.543
grad-b	0.089	0.086	0.018	456.207	455.175	12.087
xg-b	0.548	0.551	0.077	476.658	476.912	14.593
ada-b	2.869	2.851	0.247	478.238	477.300	11.566
rf	2.246	2.237	0.174	494.535	494.290	9.875
sgd-r	244.362	206.296	305.002	567.663	524.490	316.154
lasso-l	48.006	47.991	1.606	723.946	723.272	7.039
krnl-r	1722.366	1726.545	29.747	1957.720	1958.517	145.128
l-svr	1829.291	1835.390	32.747	2325.274	2258.324	232.467

We further analysed the 20 words from `iter-CFR`, and found that those are already well known in English plays during the Shakespearean era. As an obvious finding, the word 'ah' is the most frequently appearing in the `iter-CFR` model, which is indeed a signature word of Shakespeare for plays from the history play genre and has a negative correlation with date for comedies, as well as for plays in general.

The relatively good performance of Linear Regression in the out-of-domain test indicates that for the relatively short interval analysed a linear approximation already provides a good generalisation capability.

A more in-depth analysis revealed in the context of language change that a set of words ('aside', 'content', 'therefore' and 'threat') showed a significant decrease in usage over the period. However, the usage of 'threat' has declined over the years but opposite trends are exhibited in the comedy genre. This application of machine learning methods on the frequency of words from the plays not only uncovered some interesting insights about the relationship of word frequencies with genre but also provides an important new context for words which have been previously highlighted as signature words of William Shakespeare.

Given this demonstration of the competitiveness of the `iter-CFR` method in prediction of dates over a limited time-span, further experiments with this approach applied to a dataset extending over centuries rather than decades will be a valuable direction for future research. These would be likely to show some non-linear changes and here the capacity of `iter-CFR` will bring particular advantages. We hope that the approach presented here will encourage other researchers to apply quantitative methods to other important problems of literary chronology. In particular, they might test the possibility that improved performance could be achieved by using disaggregated datasets such as single-author and single-genre ones. We believe the combination of good predictive performance and ready interpretability shown here will be attractive to literary researchers.

CRediT authorship contribution statement

Pablo Moscato: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Hugh Craig:** Conceptualization, Methodology, Data curation, Writing – original draft, Writing – review & editing, Funding acquisition. **Gabriel Egan:** Data curation, Writing – review & editing. **Mohammad Nazmul Haque:** Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Kevin Huang:** Methodology, Software, Validation, Formal analysis, Investigation, Writing – review & editing. **Julia Sloan:** Methodology, Software, Validation, Formal analysis, Investigation, Writing – review & editing. **Jonathon Corrales de Oliveira:** Methodology, Software, Validation, Formal analysis, Investigation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abualigah, L., Diabat, A., Mirjalili, S., Abd Elaziz, M., & Gandomi, A. H. (2021). The arithmetic optimization algorithm. *Computer Methods in Applied Mechanics and Engineering*, *376*, Article 113609. http://dx.doi.org/10.1016/j.cma.2020.113609, URL: https://www.sciencedirect.com/science/article/pii/S0045782520307945.

Arefin, A. S., Riveros, C., Berretta, R., & Moscato, P. (2015). The MST-kNN with paracliques. In S. K. Chalup, A. D. Blair, & M. Randall (Eds.), *Artificial life and computational intelligence* (pp. 373–386). Cham: Springer International Publishing.

Arefin, A. S., Vimieiro, R., Riveros, C., Craig, H., & Moscato, P. (2014). An information theoretic clustering approach for unveiling authorship affinities in Shakespearean era plays and poems. *PLoS One*, *9*(10), 1–12. http://dx.doi.org/10.1371/journal.pone.0111445.

Baron, A., & Rayson, P. (2008). VARD2: A Tool for dealing with spelling variation in historical corpora. In *Postgraduate conference in corpus linguistics*. Birmingham, UK: Aston University, URL: https://eprints.lancs.ac.uk/id/eprint/41666.

Baron, A., Rayson, P., & Archer, D. (2009). Word frequency and key word statistics in corpus linguistics. *Anglistik*, *20*(1), 41–67.

Brainerd, B. (1980). The chronology of Shakespeare's plays: a statistical study. *Computers and the Humanities*, *14*, 221–230.

Bruster, D., & Smith, G. (2016). A new chronology for Shakespeare's plays. *Digital Scholarship in the Humanities*, *31*(2), 301–320. http://dx.doi.org/10.1093/llc/fqu068.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable tree boosting system. In *KDD '16, Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). New York, NY, USA: ACM, URL: http://doi.acm.org/10.1145/2939672.2939785.

Cotta, C., & Moscato, P. (2003). The k-Feature Set problem is W[2]-complete. *Journal of Computer and System Sciences*, *67*(4), 686–690, URL: http://www.sciencedirect.com/science/article/pii/S0022000003000813.

Craig, H. (2013). The date of Sir Thomas More. In P. Holland (Ed.), *Shakespeare survey: working with Shakespeare, Vol. 66* (pp. 38–54). Cambridge University Press, http://dx.doi.org/10.1017/SSO9781107300699.003.

Craig, H., & Whipp, R. (2010). Old spellings, new methods: automated procedures for indeterminate linguistic data. *Literary and Linguistic Computing*, *25*(1), 37–52. http://dx.doi.org/10.1093/llc/fqp033.

Culpeper, J., Archer, D., Rayson, P. E., Findlay, A. G., Hardie, a., Wattam, S. M., & Demmen, J. E. (2018). *Encyclopaedia of Shakespeare's language*. UK Research and Innovation, Research Grant: AH/N002415/1, URL: https://gtr.ukri.org/projects?ref=AH%2FN002415%2F1#/tabOverview (Accessed on Feb 19, 2021).

de Grazia, M., & Stallybrass, P. (1993). The materiality of the Shakespearean text. *Shakespeare Quarterly*, *44*(3), 255–283.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*(Jan), 1–30.

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., štajdohar, M., Umek, L., žagar, L., Žbontar, J., Žitnik, M., & Zupan, B. (2013). Orange: Data mining toolbox in python. *Journal of Machine Learning Research*, *14*(1), 2349–2353.

Fleay, F. G. (1876). *Shakespeare manual*. Macmillan.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*(200), 675–701.

Gabardo, A. C., Berretta, R., & Moscato, P. (2020). M-Link: A link clustering memetic algorithm for overlapping community detection. *Memetic Computing*, *12*(2), 87–99. http://dx.doi.org/10.1007/s12293-020-00300-x.

Gasull, A., & Utzet, F. (2014). Approximating Mills ratio. *Journal of Mathematical Analysis and Applications*, *420*(2), 1832–1853.

Gray, H. D. (1931). Chronology of Shakespeare's plays. *Modern Language Notes*, *46*(3), 147–150, URL: http://www.jstor.org/stable/2913639.

Haque, M. N., & Moscato, P. (2019). The cohesion-based communities of symptoms of the largest component of the DSM-IV network. *Journal of Interconnection Networks*, *19*(01), Article 1940002. http://dx.doi.org/10.1142/S0219265919400024.

Hill, M. J., & Hengchen, S. (2019). Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. *Digital Scholarship in the Humanities*, *34*(4), 825–843. http://dx.doi.org/10.1093/llc/fqz024.

Jackson, M. P. (1978). Linguistic evidence for the date of Shakespeare's addition to "Sir Thomas More". *Notes and Queries*, *CCXXIII*(apr), 154–156. http://dx.doi.org/10.1093/notesj/CCXXIII.197804.154.

Jackson, M. P. (2006). The date and authorship of Hand D's contribution to Sir Thomas More: Evidence from 'Literature Online'. In P. Holland (Ed.), *Shakespeare survey, Vol. 59* (pp. 69–78). Cambridge University Press, http://dx.doi.org/10.1017/CCOL0521868386.006.

Jackson, M. P. (2007a). A new chronological indicator for Shakespeare's plays and for Hand D of Sir Thomas More. *Notes and Queries*, *54*(3), 304–307. http://dx.doi.org/10.1093/notesj/gjm126.

Jackson, M. P. (2007b). Is 'Hand D' of Sir Thomas More Shakespeare's? Thomas Bayes and the Elliott-Valenza authorship tests. *Early Modern Literary Studies*, *12*(3), 3.

Jackson, M. P. (2011). Deciphering a date and determining a date: Anthony Munday's *John a Kent and John a Cumber* and the original version of Sir Thomas More. *Early Modern Literary Studies*, *15*(3).

Jackson, M. P. (2014). Vocabulary links between Shakespeare's plays as a guide to chronology: A reworking of Eliot Slater's tables. *Shakespeare*, *11*(4), 446–458. http://dx.doi.org/10.1080/17450918.2014.985604.

Jackson, M. P. (2018). Vocabulary, chronology, and the First Quarto (1603) of *Hamlet*. *Medieval & Renaissance Drama in England*, *31*, 14.

Langworthy, C. A. (1931). A verse-sentence analysis of Shakespeare's plays. *Publications of the Modern Language Association of America*, 738–751.

Malone, E. (1778). An attempt to ascertain the order in which the plays attributed to Shakespeare were written. In S. Johnson, & G. Steevens (Eds.), *The plays of Shakespeare in ten volumes, Vol. 1* (pp. 269–346).

Marsden, J., Budden, D., Craig, H., & Moscato, P. (2013). Language individuation and marker words: Shakespeare and his Maxwell's Demon. *PLoS One*, *8*(6), 1–12. http://dx.doi.org/10.1371/journal.pone.0066813.

Moscato, P., Berretta, R., & Cotta, C. (2011). Memetic algorithms. In J. J. Cochran, L. A. Cox Jr., P. Keskinocak, J. P. Kharoufeh, & J. C. Smith (Eds.), *Wiley encyclopedia of operations research and management science*. American Cancer Society, http://dx.doi.org/10.1002/9780470400531.eorms0515.

Moscato, P., Haque, M. N., Huang, K., Sloan, J., & de Oliveira, J. C. (2020). Learning to extrapolate using continued fractions: Predicting the critical temperature of superconductor materials. ArXiv E-Prints, arXiv:2012.03774.

Moscato, P., & Mathieson, L. (2019). Memetic algorithms for business analytics and data science: A brief survey. In P. Moscato, & N. J. de Vries (Eds.), *Business and consumer analytics: new ideas* (pp. 545–608). Springer, http://dx.doi.org/10.1007/978-3-030-06222-4_13.

Moscato, P., Sun, H., & Haque, M. N. (2020). Analytic continued fractions for regression: Results on 352 datasets from the physical sciences. In *IEEE congress on evolutionary computation, CEC 2020, Glasgow, United Kingdom, July 19-24, 2020* (pp. 1–8). IEEE, http://dx.doi.org/10.1109/CEC48606.2020.9185564.

Moscato, P., Sun, H., & Haque, M. N. (2021). Analytic continued fractions for regression: A memetic algorithm approach. *Expert Systems with Applications*, *179*, 115018. http://dx.doi.org/10.1016/j.eswa.2021.115018, URL: https://www.sciencedirect.com/science/article/pii/S0957417421004590.

Murphy, S. (2019). Shakespeare and his contemporaries: Designing a genre classification scheme for Early English Books Online 1560–1640. *ICAME Journal*, *43*(1), 59–82. http://dx.doi.org/10.2478/icame-2019-0003.

Naeni, L. M., Craig, H., Berretta, R., & Moscato, P. (2016). A novel clustering methodology based on modularity optimisation for detecting authorship affinities in Shakespearean era plays. *PLoS One*, *11*(8), 1–27. http://dx.doi.org/10.1371/journal.pone.0157988.

Neri, F., Cotta, C., & Moscato, P. (Eds.), (2012). *Studies in computational intelligence*: *Vol. 379*, *Handbook of memetic algorithms*. Springer, http://dx.doi.org/10.1007/978-3-642-23247-3.

Oras, A. (1960). *Pause patterns in Elizabethan and Jacobean drama: An experiment in Prosody*. University of Florida Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Plescia, I. (2016). The shape of Early Modern English: An interview with Jonathan Culpeper on the Encyclopedia of Shakespeare's Language project. *Memoria Di Shakespeare. A Journal of Shakespearean Studies*, (3), 1–19, URL: https://ojs.uniroma1.it/index.php/MemShakespeare/article/view/14177.

Rosso, O. A., Craig, H., & Moscato, P. (2009). Shakespeare and other English Renaissance authors as characterized by Information Theory complexity quantifiers. *Physica A: Statistical Mechanics and its Applications*, *388*(6), 916–926. http://dx.doi.org/10.1016/j.physa.2008.11.018, URL: http://www.sciencedirect.com/science/article/pii/S0378437108009461.

Santosa, F., & Symes, W. W. (1986). Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, *7*(4), 1307–1330. http://dx.doi.org/10.1137/0907087.

Slater, E. (1975). Shakespeare: Word links between poems and plays. *Notes and Queries*, *22*(4), 157.

Slater, E. (1988). *The problem of The Reign of King Edward III: A statistical approach*. Cambridge University Press.

Sun, H., & Moscato, P. (2019). A memetic algorithm for symbolic regression. In *IEEE congress on evolutionary computation, CEC 2019, Wellington, New Zealand, June 10-13, 2019* (pp. 2167–2174). IEEE, http://dx.doi.org/10.1109/CEC.2019.8789889.

Sun, S., Ouyang, R., Zhang, B., & Zhang, T.-Y. (2019). Data-driven discovery of formulas by symbolic regression. *Materials Research Society Bulletin*, *44*(7), 559–564. http://dx.doi.org/10.1557/mrs.2019.156.

Taylor, G., & Loughnane, R. (2017). The Canon and Chronology of Shakespeare's works. In G. Taylor, & G. Egan (Eds.), *New Oxford Shakespeare*, *The New Oxford Shakespeare: ?Athorship Companion* (pp. 417–603). Oxford, UK: Oxford University Press, URL: https://kar.kent.ac.uk/60210/.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *58*(1), 267–288.

Waller, F. O. (1966). The use of linguistic criteria in determining the copy and dates for Shakespeare's plays. In *Pacific Coast Studies in Shakespeare* (pp. 1–19). University of Oregon.

Weisstein, E. W. (2021). Mills ratio. Available at https://mathworld.wolfram.com/MillsRatio.html (Accessed on Feb 19, 2021) MathWorld - A Wolfram Web Resource, Last Updated: Feb 16, 2021.

Wells, S., Taylor, G., Jowett, J., & Montgomery, W. (1987). *Oxford Shakespeare*, *William Shakespeare: A Textual Companion* (pp. 69–144). Clarendon Press, URL: https://books.google.com.au/books?id=Bcq1QgAACAAJ.

Wentersdorf, K. P. (1951). Shakespearean chronology and the metrical tests. In W. P. Fischer, & K. P. Wentersdorf (Eds.), *Shakespeare-studien: festschrift fur heinrich mutschmann* (pp. 161–193). Marburg: N. G. Elwert.

Wiggins, M., & Richardson, C. (2012/2018). *British drama 1533-1642: a catalogue, Vols. 1-9*. Oxford University Press.

Zaher, A. A., Berretta, R., Arefin, A. S., & Moscato, P. (2015). FSMEC: A feature selection method based on the minimum spanning tree and evolutionary computation. In K. Ong, Y. Zhao, M. G. Stone, & M. Z. Islam (Eds.), *CRPIT*: *Vol. 168*, *Thirteenth australasian data mining conference, AusDM 2015, Sydney, Australia, August 2015* (pp. 129–139). Australian Computer Society, URL: http://crpit.scem.westernsydney.edu.au/abstracts/CRPITV168Zaher.html.

Zaher, A. A., Berretta, R., Noman, N., & Moscato, P. (2019). An adaptive memetic algorithm for feature selection using proximity graphs. *Computers Intelligence*, *35*(1), 156–183. http://dx.doi.org/10.1111/coin.12196.