JOURNAL OF QUATERNARY SCIENCE (2020) **35**(1-2) 81–92 ISSN 0267-8179. DOI: 10.1002/jqs.3170

JQS Journal of Quaternary Science QRA Quaternary Research Association

Machine learning classifiers for attributing tephra to source volcanoes: an evaluation of methods for Alaska tephras

MATTHEW S. M. BOLTON,[1]* BRITTA J. L. JENSEN,[1] KRISTI WALLACE,[2] NORE PRAET,[3] DAVID FORTIN,[4] DARRELL KAUFMAN[5] and MARC DE BATIST[3]

[1]Department of Earth and Atmospheric Sciences, University of Alberta, Alberta, Edmonton, Canada
[2]US Geological Survey/Volcano Science Center, Anchorage, Alaska, USA
[3]Renard Centre of Marine Geology, Ghent University, Ghent, Belgium
[4]Department of Geography and Planning, University of Saskatchewan, Saskatoon, SK, Canada
[5]School of Earth & Sustainability, Northern Arizona University, Flagstaff, Arizona, USA

Received 2 May 2019; Revised 13 November 2019; Accepted 14 November 2019

ABSTRACT: Glass composition-based correlations of volcanic ash (tephra) traditionally rely on extensive manual plotting. Many previous statistical methods for testing correlations are limited by using geochemical means, masking diagnostic variability. We suggest that machine learning classifiers can expedite correlation, quickly narrowing the list of likely candidates using well-trained models. Eruptives from Alaska's Aleutian Arc-Alaska Peninsula and Wrangell volcanic field were used as a test environment for 11 supervised classification algorithms, trained on nearly 2000 electron probe microanalysis measurements of glass major oxides, representing 10 volcanic sources. Artificial neural networks and random forests were consistently among the top-performing learners (accuracy and kappa > 0.96). Their combination as an average ensemble effectively improves their performance. Using this combined model on tephras from Eklutna Lake, south-central Alaska, showed that predictions match traditional methods and can speed correlation. Although classifiers are useful tools, they should aid expert analysis, not replace it. The Eklutna Lake tephras are mostly from Redoubt Volcano. Besides tephras from known Holocene-active sources, Holocene tephra geochemically consistent with Pleistocene Emmons Lake Volcanic Center (Dawson tephra), but from a yet unknown source, is evident. These tephras are mostly anchored by a highly resolved varved chronology and represent new important regional stratigraphic markers. Copyright © 2019 John Wiley & Sons, Ltd.

KEYWORDS: Alaska; classification; glass geochemistry; machine learning; tephra.

Introduction

Volcanic ashes (tephra) are useful chronostratigraphic markers for studies in geology, archaeology, and palaeoenvironmental sciences, forming the basis for the field of tephrochronology (Lowe, 2011; Lowe *et al.*, 2017). Some of the most challenging tephrochronologic work can be confidently cross-correlating tephras—developing 'tie-lines' between profiles, cores, sites and source volcanoes.

Meaningfully correlating disparate tephra layers requires multiple lines of evidence, including stratigraphic, physical, and geochemical characterisation. Glass geochemistry is a fundamental part of this process, where a sample's geochemical characteristics are compared with those of possible correlatives from a reference dataset, attempting to match the 'geochemical fingerprint' of the unknown to that of a known tephra. This process is complicated in regions that have experienced ashfalls from successive eruptions, thereby increasing the list of possible correlatives. Assessment complexity is increased further when tephras possess related magmatic origins and similar compositions. Arrays of bivariate plots are the primary way that tephrochronologists visually assess the relationships between glass geochemical parameters (Pearce *et al.*, 2008), regardless of additional analytical techniques that may be employed.

Pioneering efforts have examined statistical methods to overcome limitations in glass-based tephra correlations (Lowe *et al.*, 2017). These methods include 'machine learning' or 'computer intelligence' techniques. A type of machine learning called supervised classification employs algorithms that accept labelled data points (e.g. glass-shard geochemistry of a tephra with a known source) as inputs to train a model that in turn generates predictions relative to those labels. Supervised classifiers for tephra correlation are rare. However, several examples have shown that the concept is viable for tephra classification, including applications of linear discriminant analysis (LDA) (Beaudoin and King, 1986; Stokes *et al.*, 1992; Shane and Froggatt, 1994; Bourne *et al.*, 2010), and support vector machines (SVM) (Petrelli *et al.*, 2017).

Supervised models are divided into two categories based on their approach to classification (Ng and Jordan, 2002). The differentiation depends on whether a model purely calculates the probability of a class label (y) given certain predictor characteristics (x), i.e. $P(y|x)$, or whether the joint probability of both are considered first, i.e. $P(y, x)$. Based on the approach adopted, classification models are categorised as discriminative or generative, respectively (Ng and Jordan, 2002).

Not all classifiers approach the problem of multiple potential labels in the same way. Most simply, classification is the task of identifying a target class (i.e. the positive class) relative to examples of another label (i.e. the negative class). Some algorithms are intrinsically extensible from the binary situation (e.g. decision trees, nearest neighbour methods, and multi-output neural networks). Others must approach multiclass problems by combining binary classifications. For example, in a one-versus-all approach, multiclass problems

*Correspondence: Matthew S. M. Bolton, as above.
E-mail: bolton1@ualberta.ca

are interpreted as multiple binary problems by treating one class as the positive class and considering all other examples as the negative class (Rifkin and Klautau, 2004). This process is repeated until sub-models are trained with each label as the positive class. Alternatively, in one-versus-one classification, pairs of individual classes are evaluated against one another, with each having a turn as the positive class (Knerr *et al.*, 1990; Galar *et al.*, 2011). Under both approaches, multiple binary sub-models must be reconciled into a final classification. Some common methods include evaluating the confidence of the individual models (Rifkin and Klautau, 2004), or by voting (Hsu and Lin, 2002).

Although some discriminative algorithms only define the boundaries of classification in the feature space to produce discrete 'raw' label-only predictions without meaningful probabilistic interpretations (e.g. SVM), generative and many other discriminative algorithms produce probabilistic outputs. Such probabilistic outputs are important for applications in tephra identification, as they impart information about the confidence of the classification and the degree of similarity between samples.

The purpose of this article is to assess the applicability and performance of select supervised machine learning methods in determining the volcanic sources of compositionally complicated tephras based on their glass geochemistry. Late Quaternary tephra geochemical data from Alaska serve as the training and evaluation sets. We explore classification algorithms that have proven successful in tephra correlation (LDA and SVM), and some other methods shown to perform well in classification trials (Fernández-Delgado *et al.*, 2014). Several model ensembles are also evaluated. Finally, we test the most promising algorithms on the glass geochemical dataset of the tephras from Eklutna Lake, Alaska, that has been partially presented in Boes *et al.* (2018) and Fortin *et al.* (2019). These data are used to examine the models' capacity to make predictions in a dataset that had initially been evaluated through manual plotting.

We focus on the practical application and results of this study so that future analysts and tephrochronologists may adapt these methods to meet their own goals. A more detailed methodological rationale, procedure and discussion are provided in the electronic supplement (Appendix S1) along with the R code itself (Script S1).

Regional setting

More than 100 volcanoes from Alaska's Aleutian Arc-Alaska Peninsula and Wrangell volcanic field have erupted in the Quaternary Period, and more than 50 of those have been historically active (Cameron and Schaefer, 2016). Tephra from these volcanoes are widely used as chronostratigraphic markers, and a well-developed tephrostratigraphy exists for the Pleistocene from interior Alaska and Yukon (e.g. Preece *et al.*, 1999, 2011; Jensen *et al.*, 2008, 2013). Systematic work on Alaska Holocene tephra deposits has been more challenging, but has the additional goal of building eruption histories for hazard assessments, particularly for the active Cook Inlet volcanoes adjacent to Alaska's largest population centre, Anchorage (Fig. 1). Regional distal tephras are a key component in building hazard assessments because proximal records can be sparse and/or difficult to access. However, considering the temporal and spatial density of relatively recent eruptions, the list of possible correlatives for tephras in this region is large, particularly if distal cryptotephras are assessed as well.

The geochemical results from analyses of glass shards of tephras derived from the volcanoes highlighted in Fig. 1 are used as the basis for this case study. Although there are

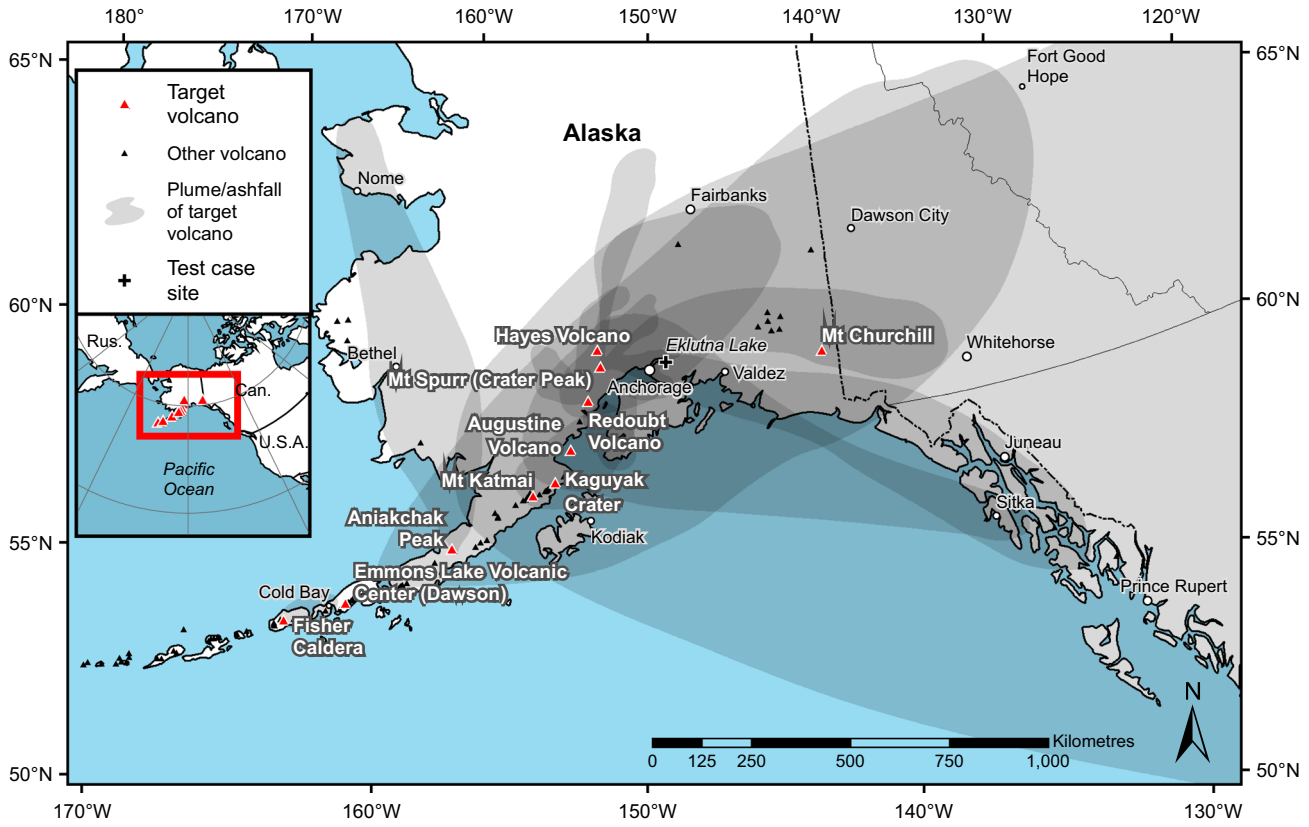

Figure 1. Map of the study area with the plume areas of substantial late Quaternary eruptions demonstrating high geographic overlap and density of substantial tephra deposits. (Plumes redrawn from Scott and McGimsey, 1994; Fierstein *et al.*, 1998; McGimsey *et al.*, 2001; Froese *et al.*, 2002; Stelling *et al.*, 2005; Fierstein and Hildreth, 2008; Bull *et al.*, 2012; Hildreth and Fierstein, 2012; Preece *et al.*, 2014; Davies *et al.*, 2016). [Color figure can be viewed at wileyonlinelibrary.com]

numerous tephras known from Alaska, this study utilises only those unambiguously tied to a source volcano, with a focus on the Holocene timeframe. This naturally skews the study towards more recent eruptions.

Materials and methods

Source data

Electron probe microanalyses of individual glass shards were used as training and validation data for modelling. They included 1953 geochemical data points from 55 samples, representing 28 tephras traced to 10 volcanic sources: Aniakchak, Augustine, Mount Churchill, Mount Spurr, Emmons Lake Volcanic Center (Dawson tephra), Fisher Caldera, Hayes, Kaguyak, Katmai-Novarupta and Redoubt. The majority of analysed samples and data are from the University of Alberta (UA) tephra collection, supplemented by data and/or samples from the Alaska Volcano Observatory. Training data drew heavily on geochemical data first reported in Davies *et al.* (2016) but include new analyses of early/mid-Holocene eruptions of Redoubt. Modelling data included weight percentage measures of major oxides: SiO_2, TiO_2, Al_2O_3, FeO, MnO, CaO, Na_2O and K_2O.

Data were parsed to remove poor or non-glass analyses. Glass geochemistries are summarised per source in Table 1. Further sample information is listed in the electronic supplement (Table S1) in addition to details of data pre-processing (Appendix S1). Glass analyses (n = 1793) were conducted at the UA's Electron Microprobe Laboratory, except Augustine data, which were collected by the US Geological Survey Electron Microprobe laboratory at Menlo Park, California (n = 160). Both laboratories were part of a formal interlaboratory comparison evaluation (Kuehn *et al.*, 2011) and were deemed to produce comparable data. For example, glass data from mid-Holocene Hayes eruptions (Wallace *et al.*, 2014) have been analysed at both laboratories, and the University of Alaska Fairbanks (Mulliken, 2016). The data from all three institutions compare favourably. This is an important consideration, as comparisons between tephras and use of models trained on those tephras are only useful so long as the analyses are accurate and reproducible.

Our training and testing dataset was selected for its size and geochemical characteristics. Tephras from Alaska can be difficult to correlate using traditional glass-shard major element analysis (e.g. plotting alone). There are numerous potential sources and eruptions, all with various degrees of geochemical similarity between them (Table 1). Plotting and comparing unknowns to the reference data becomes extremely time-consuming because there are simply so many options. These data, replete with overlapping geochemical fields, bimodal distributions, and diversity of chemistries (Table 1; Fig. 2), can meaningfully test these computational methods and their ability to reliability identify potential correlatives and make this process more efficient. In particular, geochemical similarities between Kaguyak and Augustine (Fig 2; Table 1), and Katmai and some older Redoubt material, should test the limits of the algorithms. Given the complexities and range of data in this training set, we can expect algorithms trained and tested on it to perform similarly well on comparable datasets.

Modelling

Supervised classification algorithms have two primary steps. First, a model is 'trained' on data so it can 'learn' the relationships that relate variable attributes to classification groups, $P(y|x)$. Once a model has been trained, it can be used to predict the labels of unknown data presented to it.

We trained and tested eight different learning algorithms to classify data points to source. These fitted models are referred to as 'base learners'. We also used the probabilistic outputs of the base learners as input features to three second-layer ensemble models (see Sebestyen, 1962). One ensemble was an unweighted average of two high-performing base learners (a classifier fusion ensemble; a form of non-trainable combiner

Table 1. Geochemical summary of glass data (weight percentage) used for model training, including the number of analyses and eruptive events or layers comprising each source's data pool. See Supplemental Data (Table S1) for information on individual tephras and samples in the dataset. Note: Aniakchak data in this table are divided into two geochemical populations based on silica content, although all data from this source were given the same label, 'Aniakchak', for training. SD = standard deviation.

Source		SiO_2	TiO_2	Al_2O_3	FeO_t	MnO	MgO	CaO	Na_2O	K_2O	Cl	n	events/layers
Aniakchak ($SiO_2 < 65\%$)	Average	59.17	1.38	16.50	7.56	0.21	2.79	6.39	4.34	1.55	0.13	117	2
	SD	1.40	0.08	0.24	0.75	0.04	0.32	0.59	0.40	0.14	0.03		
Aniakchak ($SiO_2 > 65\%$)	Average	71.09	0.49	15.16	2.54	0.14	0.51	1.80	5.10	2.98	0.20	175	2
	SD	0.61	0.07	0.21	0.30	0.03	0.08	0.20	0.28	0.13	0.03		
Augustine	Average	76.26	0.35	13.01	1.99	0.06	0.44	2.23	3.45	1.89	0.31	160	5
	SD	1.35	0.08	0.52	0.30	0.04	0.10	0.31	0.34	0.11	0.05		
Churchill	Average	74.30	0.19	14.26	1.44	0.05	0.31	1.74	4.07	3.32	0.33	491	2
	SD	1.03	0.06	0.58	0.23	0.03	0.09	0.26	0.22	0.23	0.04		
Emmons (Dawson)	Average	74.17	0.26	13.61	2.08	0.07	0.23	1.24	4.48	3.64	0.23	117	1
	SD	0.31	0.04	0.23	0.08	0.03	0.03	0.06	0.15	0.13	0.03		
Fisher	Average	68.97	0.54	15.69	4.04	0.19	0.48	2.26	5.25	2.43	0.17	209	2
	SD	1.36	0.09	1.23	0.60	0.05	0.15	0.62	0.56	0.31	0.03		
Hayes	Average	74.38	0.23	14.24	1.70	0.08	0.48	2.26	3.82	2.53	0.36	139	6
	SD	2.03	0.08	1.05	0.51	0.03	0.40	0.59	0.47	0.38	0.09		
Kaguyak	Average	77.56	0.29	12.48	1.38	0.05	0.27	1.87	4.00	1.95	0.18	32	1
	SD	0.29	0.04	0.16	0.18	0.03	0.04	0.07	0.11	0.05	0.02		
Katmai	Average	77.08	0.25	12.49	1.58	0.06	0.23	1.17	3.97	3.00	0.19	245	1
	SD	2.43	0.16	0.79	0.72	0.03	0.24	0.74	0.20	0.24	0.04		
Redoubt	Average	74.51	0.39	13.70	1.76	0.07	0.42	1.82	4.17	3.05	0.15	218	7
	SD	3.17	0.12	1.61	0.73	0.03	0.30	0.90	0.46	0.53	0.06		
Spurr	Average	63.18	0.85	15.94	6.44	0.18	2.00	4.89	4.69	1.64	0.26	50	1
	SD	0.52	0.06	0.34	0.23	0.04	0.17	0.22	0.22	0.09	0.04		

All analyses are normalised to 100%. n = number of shards analysed. FeO_t = all Fe as FeO

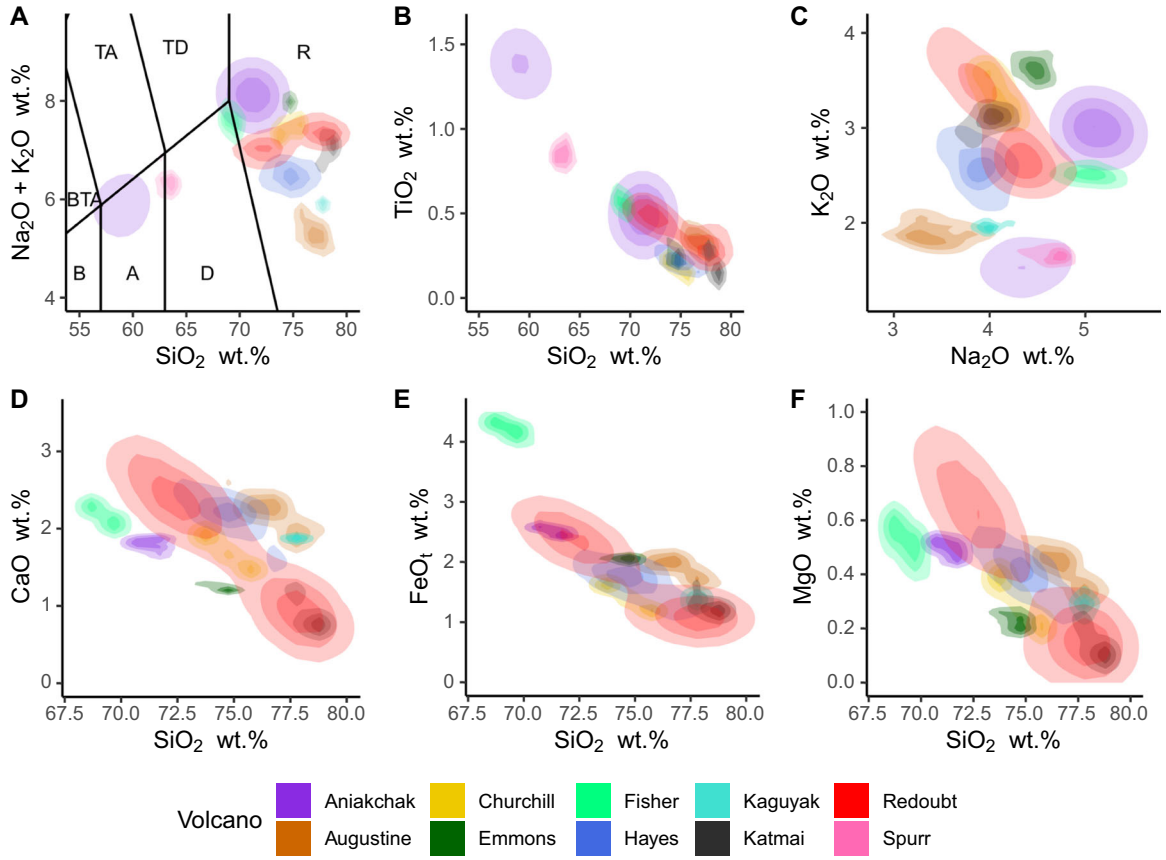

Figure 2. Glass geochemical plots for the full training set, with data summarised by density fields (darker colours mean higher density). Note: Spurr and the low silica population of Aniakchak are easily differentiable based on SiO_2 (A, B); they are excluded from the lower row plots (D, E, F) to show more detail for the remaining distributions. Total alkali-silica (TAS) plot (A) following le Bas *et al.* (1986); B = basalt, BTA = basaltic trachyandesite, TA = trachyandesite, TD = trachydacite, R = rhyolite, A = andesite, D = dacite. [Color figure can be viewed at wileyonlinelibrary.com]

(Kuncheva, 2004)), while the second and third ensembles were trainable combiner 'model stacks', or 'meta-models' that fit higher-level models (artificial neural network (ANN) and random forest (RF)) using all probabilistic base learner predictions. This ensemble technique is synonymous with 'stacked generalisation' and is employed with the goal of minimising the error rate (Wolpert, 1992). The base learners and ensembles trialled are detailed in Table 2. The nuances of each method are well documented in other sources, including their respective R package documentation (Table 2) and in Kuhn and Johnson (2013). However, brief summaries of the fundamental concepts behind each learner's approach and key references for the methods are included in supplementary Table S2.

Two SVM variants were trained. By default, SVMs provide non-probabilistic class predictions; a second-layer model must be used to calibrate the outputs into a probabilistic format (Platt, 1999). We tested the discrete SVM classifier (SVM raw), and one adapted to produce probabilities using a sigmoid function (SVM prob.) (Platt, 1999; Wu *et al.*, 2004). Only the SVM prob. predictions were used in the meta-models. Of the algorithms tested, SVM was the only method that could return different predictions in 'raw' and 'probabilistic' modes.

Each of the base learners and trainable ensembles was 'optimised' such that the tuning parameters for each allowed the maximisation of a performance measure, Cohen's kappa statistic (κ), within a subset of possible permutations. Tuning was conducted using cross-validation among data specifically partitioned for training and tuning. The data were split into three partitions, stratified by volcanic source, following Arlot and Celisse (2010). Of all data, 40% were allotted to train base learners, another 40% were used in training the subsequent ensembles, and 20% were reserved for evaluation (the 'test set'). Performance was measured through cross-validation during training and directly on the test set. Cohen's kappa and overall accuracy were used to evaluate model efficacy. Cross-validation performance was also assessed for final models using the complete dataset. Also, performance was evaluated using an exact one-sided binomial test (Clopper and Pearson, 1935), resolving the p value that accuracy is greater than the null hypothesis (the 'no information rate' NIR; i.e. the accuracy if all records are labelled as the most common class) by chance alone.

Modelling was conducted in the R programming language and software environment (R Core Team, 2019). An easily adaptable R script is supplied in the supplementary material (Script 1) such that the code can be used to train new models and adapted to different use-cases as users see fit. Further, 'final' models, trained on the full dataset, are presented as stand-alone R objects (Models S1), so that users can make source predictions using our fitted models on new data.

Evaluation on new samples

As a trial case, a suite of 11 tephras from Eklutna Lake was assessed using this study's machine learning methods and traditional techniques. Initial correlations used traditional plotting methods, comparing the glass analyses of the tephras to internal glass geochemical data and analyses from selected literature (Riehle, 1985; Begét and Nye, 1994; Payne and Blackford, 2008). Some of these results have been reported by Boes *et al.* (2018) and Fortin *et al.* (2019). Here, we present the entire tephrostratigraphy and reassess all geochemical results using machine learning. Instead of seeking an explicit measure

Table 2. Table summarising classifiers used in this study and the abbreviations by which they are referred to throughout.

Algorithm name	Abbreviation	Caret method	Parent package	Approach	Explanatory variables
Classification Tree	CART	rpart	rpart (Therneau and Atkinson, 2018)	Discriminative, multiclass	Geochemistry
Random Forest	RF	rf	randomForest (Liaw and Wiener, 2002)	Discriminative, multiclass	Geochemistry
Support Vector Machine with Radial Kernel	SVM	svmRadialSigma	kernlab (Karatzoglou *et al.*, 2004)	Discriminative, 'one against one' voting (e.g. Hsu and Lin, 2002)	Geochemistry
K Nearest Neighbours	KNN	knn	caret (Kuhn, 2008)	Discriminative, multiclass	Geochemistry
Naive Bayes	NB	naive_bayes	naivebayes (Majka, 2019)	Generative, multiclass	Geochemistry
Linear Discriminant Analysis	LDA	lda	MASS (Venables and Ripley, 2002)	Generative, multiclass	Geochemistry
Artificial Neural Network	ANN	nnet	nnet (Venables and Ripley, 2002)	Discriminative, multiclass	Geochemistry
C5.0	C5.0	C5.0	C50 (Kuhn and Quinlan, 2018)	Discriminative, multiclass	Geochemistry
Ensemble Average	Average Ensemble	none	base (R Core Team, 2019)	Discriminative, multiclass	Base learner predictions (probability), RF and ANN
Random Forest Ensemble	Meta-RF	rf	randomForest (Liaw and Wiener, 2002)	Discriminative, multiclass	Base learner predictions (probability), all
Artificial Neural Network Ensemble	Meta-ANN	nnet	nnet (Venables and Ripley, 2002)	Discriminative, multiclass	Base learner predictions (probability), all

of model performance, we strove to assess the feasibility of machine learning for identifying the sources of unknown tephras (not necessarily the specific eruption), while exploring the method's practicality.

Results

Learner performance

All base-learning and ensemble algorithms resulted in mean kappa values of greater than 0.56 on their respective training sets. Minimum mean training accuracy was 0.63. Of the fitted models, the SVM prob. performed the worst, with all others resolving mean and median kappa and accuracy statistics >0.95 (Fig. 3). Given the good cross-validation results of ANN and RF, and their reputation for producing realistic probabilistic outputs (Niculescu-Mizil and Caruana, 2005), these two best-performing probabilistic base learners were combined to form the average ensemble model. Further analysis of learner performance is included in the supplementary material (Appendix 1).

When models were trained on only their respective training sets and evaluated on test data, performance trends followed much the same patterns as in the training cross-validation (Fig. 4). For all models tested, accuracy was significantly higher than the no information rate (NIR = 0.2513; p values < 2×10^{-55}; as low as 1.1×10^{-212} for the Meta-RF ensemble).

The use of kappa as a class-size-weighted metric is valuable in this dataset, and we suggest it is appropriate to use for many other tephra datasets given unequal label frequencies found in most studies. Following the arbitrary ranking of kappa values presented by Altman (1990), predictions of all the learners exhibited 'very good agreement' with the test data (kappa 0.8–1), with the exception of CART, which demonstrated 'good agreement' (0.6–0.8), and SVM prob., which showed only 'moderate agreement' (0.4–0.6) (Fig. 4).

Generally, the most computationally complex models, including the various ensemble methods and the base ANN, performed the best (Fig. 4). Given the good final performance of all the models, except SVM prob. and CART, the question remains if increased computational complexity (i.e. time) is worth a modicum of heightened classification performance. This decision will require a value judgement by the end-user. For this study, we believe the most reasonable compromise between complexity and performance is the average ensemble of ANN and RF base learners. This simple ensemble delivers consistently high performance, reduces the variance inherent in a single base learner, while limiting the intensity of training and cross-validation required.

Shard versus sample performance

Despite variations in classification performance, the majority of single-point measurements (e.g. individual shards) within each sample were correctly identified more often than not by all learners, even the probabilistic SVM. This result indicates that applying classification schemes to glass data on a per-shard basis can be accurate when predictions are pooled for each sample and the mode is accepted as the final prediction. This method, called voting, is common for aggregating discrete predictions, as in the SVM raw model. However, a more valuable option is to use a model's average probabilistic outputs per sample as a pooled prediction.

Because the class probabilities for each shard analysed from a sample can also be composited in the mean predictions per source, the result is a 'probability' for each sample's source as a whole. This shard-wise aggregation diverges from the

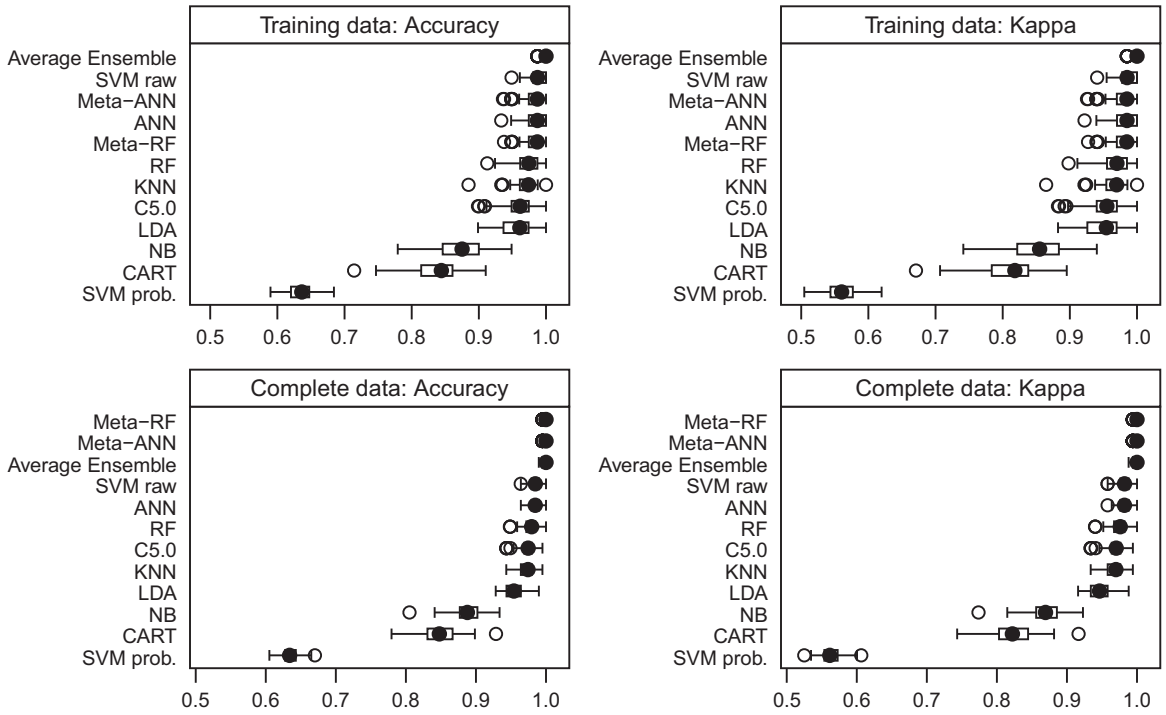

Figure 3. Box and whisker plots of model performance on cross-validation resamples (100 each) for training data only and final models (complete data). Boxes represent interquartile range (IQR), with filled circles indicating median; whiskers are 1.5 times IQR above and below the box or minimum/maximum data limits if minimum/maximum points are within 1.5 times IQR range; hollow circles are fold measurements that fell outside the whisker range. See Table 2 and Appendix S1 for explanation of algorithms used.

common method of using the mean geochemical values for each sample as central point estimates (e.g. Shane and Froggatt, 1994). Note, despite being the accepted term for the type of prediction in question (i.e. probabilistic classification), the word 'probability' may be misleading. Predictions are derived from a probability distribution restricted to the set of classes considered by the model. As such, the sum of a record's probabilities from all labels must equal 100%. In other words, the values calculated per class represent conditional probabilities, given that the unknown evaluated was in the training set. We made no attempt to calibrate probabilities outside of modelling (cf. Niculescu-Mizil and Caruana, 2005).

Based on predictions from the test data, all evaluated probabilistic learners were more accurate when the averages of probabilities per sample were used for classification instead of their single-point (raw) classifications alone. The sample-wise accuracy for probabilistic models was, on average, 3% higher than the evaluation of independent shards. In fact, the average ensemble, meta-ANN, ANN, and LDA classified the 54 samples included in the test set perfectly when sample averages were used. All probabilistic learners with the exception of CART and SVM prob. correctly classified over 96% of samples when evaluated in this way. Meta-RF, RF, and NB each misclassified one sample, C5.0 and KNN misclassified two, while CART misclassified six, and SVM prob. misclassified 20. When mean probabilities were used for predictions, NB accuracy increased by 7.6% and LDA increased by 5.9%, but SVM prob. increased by only 0.1%. In the case of raw-only predictions, when the mode was used per sample, the SVM raw model predicted classes perfectly in the test set, up from 97% accuracy on individual shards.

By evaluating a sample's individual point data, a more complete assessment of the composition can be made than if only geochemical means are used. For example, nuances such as polymodal distributions can be used in classification and

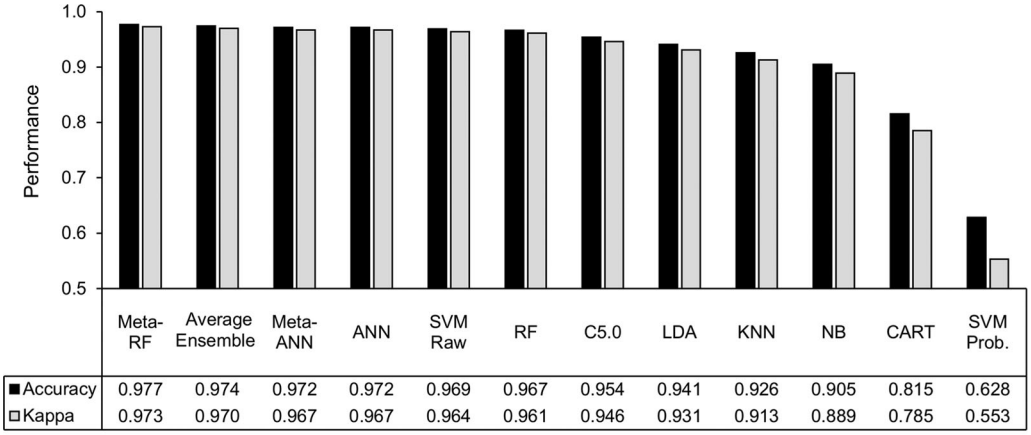

Figure 4. Algorithm performance; models trained on the training set and evaluated on the test set. See Table 2 and Appendix S1 for explanation of algorithms used.

detected in unknowns. Shard-wise predictions can also be useful for discerning the sources of shards that have been redeposited and intermixed within tephra layers, though are compositionally distinct on a statistical level (e.g. Pouget *et al.*, 2014).

Comparison with prior work

The relatively poor performance of SVM prob. on this dataset was unexpected given the high accuracy of a raw (non-probabilistic) model of similar design (>0.90) that discerned volcanic sources in Italy (Petrelli *et al.*, 2017). Even averaging the predicted probabilities per sample, the predictive accuracy of SVM prob. was hardly better than considering each shard independently (0.630 vs 0.628). However, Petrelli *et al.* (2017) used P_2O_5 and trace elements as predictors in addition to major elements, which undoubtedly contributed to their SVM's accuracy. Interestingly, the comparatively poor SVM prob. accuracy here does compare to the raw accuracy of Petrelli and Perugini (2016) (0.69) when SVMs were trained only on major oxides to discriminate rock tectonic environments. SVM's poorer performance on our dataset suggests that the present multiclass implementation of the algorithm and associated probability model can be problematic for some datasets. The difference between probabilistic and raw predictions is noted by Wu *et al.* (2004) as a result of how sigmoid functions split classes. They suggest no solution for when probability calibration fails. However, given the good performance of the SVM raw model, and following evaluation of the decision values produced by the uncalibrated SVM model that underlies the SVM prob. model, we are confident that reasonable decision boundaries can be found by SVM for this dataset. Unless an alternative multiclass probability calibration method is employed for SVMs, care should be taken when evaluating learner performance or using predictions of SVM probabilistic models.

Most of the algorithms selected for use in this study are known to work particularly well with non-linear data relationships, many features, and non-normal distributions. One significant departure is LDA, which is known to have problems coping with non-normal data, and particularly, multicollinearity (Naes and Mevik, 2001). Both characteristics are common in geochemical datasets, including ours, where pairwise correlations—quantified by R^2—are frequently > 0.9. Nevertheless, LDA demonstrated that discrimination rules can be well defined, even where the eigenvalues that determine them are relatively small, while still allowing for reliable classification of tephra. In relation to other LDA efforts for geochemical classification of tephra-derived glass, our final model performed comparably in terms of accuracy on reference data: our study = 95.1%; Tryon *et al.* (2009) = 95–97%; Stokes and Lowe (1988) = 97.5%; Charman and Grattan (1999) = 90.3%.

CART methods have been successful in tracing obsidian to source with 96% accuracy using trace metal data (Sheppard *et al.*, 2011). This result is substantially more accurate than our CART models, although this separation can probably be ascribed to differences in data characteristics, not implementation. Further, an example of a classification tree is presented in Lowe *et al.* (2017) for glass data, although this was provided as a data exploration tool, not strictly for deriving class predictions. As such, no performance measure was given. One other early study of statistical cheomstratigraphy, Malmgren and Nordlund (1996), compared ANN, KNN, and LDA for classifying volcanic ash zones utilizing major oxide geochemistry. Their basic findings agree with ours, indicating high accuracy of ANN on held-out data (our study = 97.7%, Malmgren and Nordlund (1996) = 90.8%). Their assessment of KNN and LDA were less favorable (69.2% and 61.6% accuracy respectively) (Malmgran and Nordlund, 1996). For the other algorithms, the authors know of no comparable performance baseline in literature for the classification of glass composition.

Test application: Eklutna Lake tephras

Eklutna Lake is a glacially fed lake approximately 45 km northeast of Anchorage (61°22′36"N 149°02′07"W, Fig. 1). The lake is an important paleoenvironmental archive not in small part because the sediments are varved (e.g. Loso *et al.*, 2017; Praet *et al.*, 2017; Boes *et al.*, 2018). Tephras in the lake cores have been important in developing the chronology for these studies, but not all tephras present have been reported (e.g. Boes *et al.*, 2018; Fortin *et al.*, 2019). These tephras are of interest because they are regionally distributed, most are exceptionally well dated through the varve chronology, and they represent past eruptions where ashfall impacted what is now the most densely populated region of Alaska.

Eleven distinct tephra layers were characterised, with an additional four samples containing four different, mixed populations. Sampling of background sediments shows that glass is a component of the lake sediment. The tephras' glass geochemistry and average median ages (Fortin *et al.*, 2019) are summarised in Table 3 and included as supplementary data (Table S3). A diagram indicating the relative core depths/stratigraphy of the tephras examined is included in the supplement (Fig. S1). Overall, of the 15 populations tested, representing 13 samples, an agreement between machine learning and plotting was found in all 12 tephras initially assigned to a source volcano (Table 4).

The source classification strongly supports the initial geochemical correlations of Tephras 1 and 2 presented in Boes *et al.* (2018), and Tephras 5, 7, 10 and 12 in Fortin *et al.* (2019). Tephra 1 was correlated with the AD 1992 Crater Peak eruption of Spurr, and Tephra 2 to the AD 1989/1990 eruption of Redoubt Volcano. Tephras 5, 10 and 12 are all attributed to Redoubt, and Tephra 7 to Augustine (Fortin *et al.*, 2019). These assertions are strengthened by known activity of Redoubt around 450 cal a BP (Begét and Nye, 1994; Begét *et al.*, 1994; Schiff *et al.*, 2010) and Augustine around 700 cal a BP (Waitt and Begét, 2009).

New data presented here are for Tephras 3, 4, 6, 8, 9, 11, 17 and 19. Tephras 3, 6, 8 and 9 are largely comprised of mixed geochemical populations. Some have identifiable geochemical groupings, but inconsistencies between cores and a detrital component preclude their identification as primary deposits. For example, Tephra 3, younger than ~AD 1930, contains geochemical populations sourced to Katmai (AD 1912), and the Dawson tephra. Dawson tephra, from a late Pleistocene caldera-forming eruption, also appears in several of the other detrital-rich samples. All other tephras were considered to be primary based on their presence across multiple cores, consistent stratigraphy, and purity of the samples.

Tephra 4 is attributed to Redoubt and is geochemically similar to late 20th and 21st century eruptions from that volcano. The most likely event would be the well-documented eruption of AD 1902; however, the age estimate for this tephra is closer to ~AD 1880. There are no documented eruptions from Redoubt at this time, except for a short note in Cordeiro (1910) about a potential event in AD 1881. Tephra 11 correlated to Emmons Caldera, which was initially problematic due to the reworking of Dawson tephra in lake sediments. However, the purity of the samples across multiple cores suggests that Tephra 11 represents a primary Holocene event. In fact, most of the tephras from this lake were effectively cryptotephras (i.e. non-visible tephras, detected in

Table 3. Weight percentage averages and standard deviations (SD) of glass-shard analyses of primary tephras identified in Eklutna Lake, their correlatives, and their modelled ages where present in varved cores (following Fortin *et al.*, 2019). Note: Tephra 11 is interpreted as a primary tephra and analyses of its constituent glass shards are geochemically consistent with those of the Dawson tephra, but at present we have no evidence of silicic glass being produced from Emmons in the Holocene.

Sample	SiO_2	TiO_2	Al_2O_3	FeO_t	MnO	MgO	CaO	Na_2O	K_2O	Cl	H_2O_d	n	Average modelled median age (a BP)	Average SD (a)
Tephra 1	63.14	0.85	16.12	6.31	0.17	1.98	4.91	4.59	1.67	0.26	2.39	39		
Crater Peak	0.69	0.06	0.52	0.19	0.04	0.12	0.31	0.30	0.13	0.03	0.84			
Tephra 2	77.43	0.29	12.43	1.18	0.05	0.16	1.04	3.80	3.49	0.13	1.81	93		
Redoubt	1.38	0.08	0.82	0.22	0.03	0.07	0.38	0.34	0.29	0.05	0.91			
Tephra 4	77.47	0.25	12.50	1.12	0.05	0.20	1.11	3.84	3.32	0.13	2.22	63	70	3
Redoubt	0.51	0.05	0.25	0.18	0.03	0.03	0.12	0.22	0.16	0.04	0.59			
Tephra 5	75.67	0.26	13.44	1.37	0.07	0.28	1.54	4.10	3.13	0.14	2.40	126	454	5
Redoubt	1.97	0.08	0.88	0.40	0.02	0.13	0.49	0.30	0.23	0.04	1.87			
Tephra 7	74.04	0.44	13.49	2.33	0.06	0.58	2.53	4.29	1.92	0.31	0.54	101	729	6
Augustine	1.33	0.08	0.55	0.35	0.02	0.14	0.42	0.26	0.26	0.06	1.25			
Tephra 10	77.11	0.26	12.49	1.19	0.05	0.21	1.18	3.97	3.37	0.17	2.87	88	1312	13
Redoubt	0.90	0.05	0.45	0.18	0.02	0.07	0.26	0.15	0.18	0.03	1.26			
Tephra 11	74.17	0.28	13.59	2.04	0.07	0.24	1.27	4.53	3.65	0.22	3.35	52	1579	14
Emmons	0.37	0.05	0.11	0.17	0.02	0.06	0.14	0.15	0.21	0.02	1.79			
Tephra 12	74.66	0.30	13.70	1.68	0.08	0.32	1.65	4.38	3.06	0.17	2.27	72	1749	15
Redoubt	0.54	0.04	0.28	0.18	0.03	0.05	0.11	0.12	0.10	0.03	1.02			
Tephra 17	70.72	0.50	15.04	2.85	0.10	0.75	2.79	4.44	2.65	0.18	2.61	82		
Redoubt	1.48	0.07	0.56	0.43	0.03	0.17	0.47	0.17	0.26	0.03	1.40			
Tephra 18	75.03	0.21	13.63	1.67	0.07	0.39	2.12	3.90	2.53	0.46	3.34	22		
Hayes	0.52	0.04	0.20	0.15	0.01	0.05	0.17	0.10	0.10	0.04	2.20			
Tephra 19	73.10	0.40	14.21	2.23	0.09	0.51	2.28	4.33	2.65	0.20	2.84	35		
Redoubt	1.77	0.08	0.71	0.43	0.03	0.15	0.51	0.18	0.20	0.02	1.59			

All analyses are normalised to 100%. n = number of shards analysed. FeO_t = all Fe as FeO

this study by magnetic susceptibility). The presence of Tephra 11 as a single-population, highly glass-dense unit accentuates its importance and helps identify it as a primary ashfall unit, and not just reworked Dawson. This interpretation is also supported by the slightly lower hydration of glass from Tephra 11 (~<3 wt%) compared with that of reworked Dawson (~>3.5 wt%). However, we must emphasise that there is no published evidence of silicic glass products from Emmons in the Holocene and that modern eruptions from the area (e.g. Pavlof Volcano) are more mafic than Dawson (e.g. andesitic composition; not rhyolitic) (Waythomas *et al.*, 2017).

Two of the three oldest units, Tephra 17 and 19, were tentatively identified as originating from Redoubt when traditional methods were employed. However, these layers also shared geochemical characteristics with material from the Katmai Volcanic Cluster, especially Tephra 19. An early version of the RF/ANN average model trained only with Redoubt data from the AD 1989–90 and AD 2009 eruptions indicated that these samples were statistically most similar to eruptives from Katmai. However, predictions from the average ensemble trained on the full dataset that included new early- and mid-Holocene Redoubt data were more or less definitive, favouring Redoubt over Katmai as the source for Tephra 17 with 11 times the probability. Tephra 19 was less clearly separable, but Redoubt was still 4.27 times more likely than Katmai. This notable shift in predictive outcomes following a

Table 4. Shard-averaged membership probabilities for unknown populations of tephras from Eklutna Lake resulting from ANN/RF average ensemble, coupled with the perceived correlations resulting from initial traditional plotting and machine learning (in this case, the maximum probability per population). Populations within tephras are denoted with decimal suffixes.

	Aniakchak	Augustine	Churchill	Emmons	Fisher	Hayes	Kaguyak	Katmai	Redoubt	Spurr	Traditional Plotting Correlation	Average RF/ ANN Most Probable Correlation
Tephra 1	4.5%	0.4%	0.0%	0.0%	0.4%	0.9%	0.0%	1.8%	0.2%	91.8%	Spurr	Spurr
Tephra 2	0.1%	0.1%	3.2%	3.4%	0.0%	3.9%	0.2%	9.1%	79.8%	0.0%	Redoubt	Redoubt
Tephra 3.1	0.2%	0.0%	3.2%	94.9%	0.0%	0.5%	0.0%	0.4%	0.7%	0.0%	--	Dawson
Tephra 3.2	0.0%	0.4%	0.1%	0.1%	0.0%	0.9%	0.3%	86.8%	11.3%	0.0%	Katmai	Katmai
Tephra 4	0.0%	0.1%	2.6%	0.2%	0.0%	2.7%	0.2%	19.4%	74.9%	0.0%	Redoubt	Redoubt
Tephra 5	0.6%	0.2%	15.8%	0.5%	0.1%	8.5%	0.2%	8.4%	65.7%	0.0%	Redoubt	Redoubt
Tephra 7	0.3%	77.1%	0.5%	1.2%	0.6%	5.7%	1.0%	4.4%	8.7%	0.5%	Augustine	Augustine
Tephra 8.1	0.2%	0.0%	1.6%	96.4%	0.0%	0.2%	0.0%	0.3%	1.3%	0.0%	--	Dawson
Tephra 8.2	0.0%	89.1%	0.1%	0.0%	0.2%	4.2%	0.2%	4.0%	2.2%	0.0%	Augustine	Augustine
Tephra 10	0.0%	0.1%	3.3%	0.2%	0.0%	2.8%	0.4%	9.3%	83.7%	0.0%	Redoubt	Redoubt
Tephra 11	0.1%	0.0%	2.6%	88.2%	0.0%	0.1%	0.0%	0.3%	8.7%	0.0%	--	Dawson
Tephra 12	1.3%	0.4%	27.7%	4.9%	0.2%	3.5%	0.3%	5.1%	56.5%	0.0%	Redoubt	Redoubt
Tephra 17	2.3%	0.2%	0.5%	0.7%	3.6%	1.5%	0.0%	7.6%	83.4%	0.2%	Redoubt	Redoubt
Tephra 18	0.0%	0.5%	1.1%	0.0%	0.0%	97.2%	0.1%	0.5%	0.5%	0.0%	Hayes	Hayes
Tephra 19	1.5%	1.0%	1.9%	0.3%	1.2%	7.5%	0.3%	16.4%	69.9%	0.1%	Redoubt	Redoubt

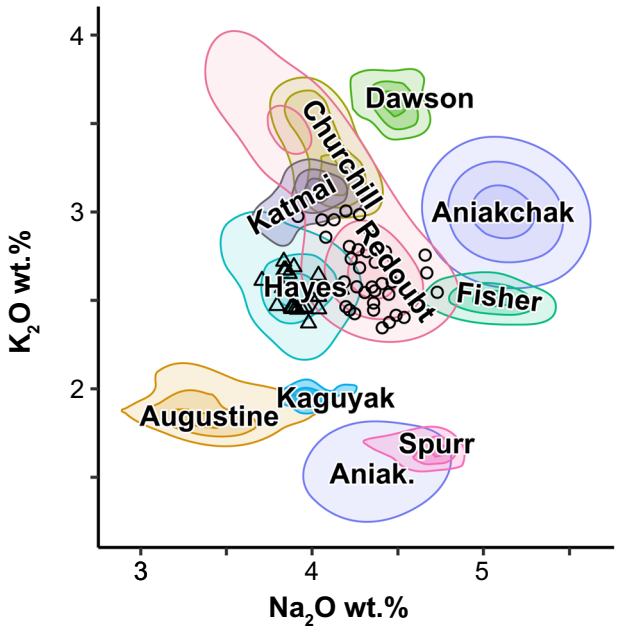

Figure 5. Glass geochemical plot highlighting separability between volcanic sources. Shaded areas represent high-density regions based on the training dataset. Hollow triangles = Tephra 18; Hollow circles = Tephra 19. [Color figure can be viewed at wileyonlinelibrary.com]

change to the training data highlights the importance of exhaustive classifier training data relative to the potential tephras being analysed. While we were able to improve the training dataset for Redoubt by including newly collected data from older eruptions, our Katmai volcanic cluster data are limited to the AD 1912 eruption. The third older unit, Tephra 18, was clearly identified as a Hayes-derived tephra.

All correlations were tested with bivariate plotting as well as statistical analysis. Visual discrimination between sources in this study is clearest in a Na_2O vs K_2O plot (Figs. 2,5). We also highlight that the two oldest tephras (Tephra 18, from Hayes, and Tephra 19, from Redoubt) from Eklutna Lake cluster in the high-density regions of the Na_2O – K_2O chemical-space for their respective assignments, further supporting our statistical identification of these tephras (Fig. 5).

The Fortin *et al.* (2019) age model does not extend to the most distal core (site 1), which is the only core that extended down to tephras older than Tephra 12. Fortunately, Tephra 18 was further correlated to Hayes tephra unit H2, a component of tephra H (Fig. 6; Wallace *et al.*, 2014), through comparisons with reference Hayes glass chemistry. Tephra unit H of Wallace *et al.* (2014) also correlates to Tephra T1 of Combellick and Pinney (1995), which has upper and lower bounding dates. Utilising their ^{14}C dates from peat above and below, and wood above the tephra, we have calculated a median age for the tephra of 3713 ± 72 cal a BP by modelling the age of a boundary, 'tephra unit H', between terminus ante/post quem phases in OxCal v4.3.2 (Ramsey, 2009), calibrated with the IntCal13 curve (Reimer *et al.*, 2013). Tephra 17 is younger than Tephra 18, but below the base of the age model, thus between ~2200 and 3700 cal a BP. Without radiocarbon dates constraining the age of the lowest tephra, Tephra 19, we can only say that this tephra is older than Tephra 18/Hayes H. However, the interval between the tephras is probably substantial (perhaps on the order of one thousand years), given the ~2 m of sediment between them.

It needs to be kept in mind that prediction probabilities can be biased when the training data are missing specific eruption geochemistry, or when there are little or no data relating to a source. For example, we are limited to Katmai 1912, and have no data from Iliamna. Eruptives from both volcanoes do or may exhibit geochemical characteristics similar to products from Redoubt. We have examples through time of Redoubt's many Holocene eruptions, and careful plotting reveals the glass analyses of the Redoubt tephras (4, 5, 10, 12, 17 and 19) are on trend with reference data. However, without a clear match to a specific eruption, we must recognise that attribution cannot be unequivocally determined.

We believe that the machine learning approach to tephra sourcing has proven reliable when provided with appropriately diverse training data, and produces informative source predictions. The average ensemble, built from RF and ANN learners, was both relatively quick to train and fast in making predictions (< 1 s for >800 point analyses, each with nine geochemical parameters).

Further considerations

Petrelli and Perugini (2016) discussed a number of caveats about machine learning in geochemical discrimination that are applicable to this study. They asserted that machine learning models should be evaluated carefully, and that models do not magically classify data with unquestionable labels. Results from machine learning must be integrated with and tested by other analyses that include traditional geochemical assessment by plotting. We agree with Lowe *et al.* (2017) that robust correlations must consider some combination of physical, stratigraphic, compositional and chronologic parameters. In addition, we strongly suggest the use of plotting to assess input data before models are fit to ensure that only the best quality data are used. Thus, the resulting models are preconditioned by human experts.

We also must emphasise that any supervised learning method will fail if the examples of the true class of unknowns were not included in the data used to fit the model. This step is obvious

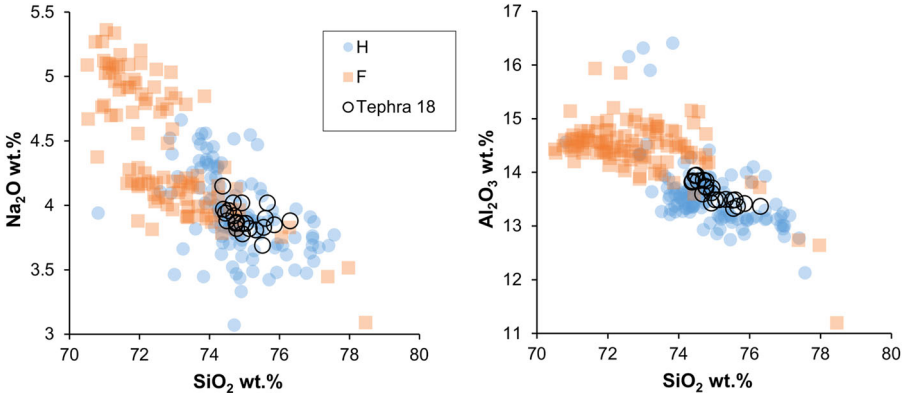

Figure 6. Geochemical plots of glass compositions demonstrating the good correlation of Tephra 18 with Tephra H, and not with the other chronologically proximate Tephra F/Jarvis Ash. [Color figure can be viewed at wileyonlinelibrary.com]

but underscores the importance of using exhaustive and confidently identified tephra data as inputs for model training. Although not tested in this work, 'outlier detection' methods may be helpful for pre-evaluating unknown data before presenting them to a classifier. This could help solve the problem of predictions only representing classes within the training set. These methods could filter out samples unlikely to belong to any of the reference classes and exclude them from predictions. Examples include one-class SVM (Schölkopf *et al.*, 2001), soft independent modelling of class analogy (SIMCA) (Wold and Sjöström, 1977), or training some other algorithm to differentiate training set data from a background of random data (Hastie *et al.*, 2009; Kuhn and Johnson, 2013). No matter what filtering protocol may be adopted, it is always advisable to evaluate unknowns against reference data from their supposed labels using traditional plotting and comparison, including the use of stratigraphic, physical, and age data where possible.

Conclusions

Conclusions drawn from this work are divided into three categories: 1) the applicability of machine learning classifiers for tephra source attribution using glass composition overall; 2) the more specific practicality of using classifiers to help determine volcanic sources based on major oxides in tephra glass from south-central Alaska; and 3) how machine learning can aid in the assessment of unknown tephras, including findings from Eklutna Lake.

1. In terms of the broad-scale adoption of machine learning in tephra analysis:

- Classifiers can be useful for quickly parsing glass geochemistry datasets.
- Classifiers can generate probabilistic predictions of volcanic source.
- Aggregated point analysis predictions are more useful than classifications from geochemical means or raw (label-only) predictions.
- By using point analyses, mixed geochemical populations can be detected and discriminated.
- Algorithm performance is not always consistent given differing problem questions or data presented. As such, algorithms from differing methodological families should be evaluated when new research questions are addressed.
- Ensemble models can effectively improve classification performance and reduce variance, but their use may be limited by their heightened computational requirements.
- Any learning algorithm is only as good as the data on which it is based. In order to best utilise classification methods, wisely curated and appropriately expansive and reliable glass geochemistry datasets must be available as reference data.

2. Specific conclusions from our exploration of the classification of select Alaska tephras include:

- RF and ANNs appear to be among the most robust base learners explored, and their averaged prediction (forming a simple ensemble) is a computationally cost-effective method that yields high performance through cross-validation and on 'true unknowns'.
- LDA, despite being among the least computationally complex methods trialled, still proved highly accurate. Although learner complexity is often correlated with performance, this outcome is not always the case.
- The methods trialled view data synoptically and can define multidimensional decision boundaries, even when geochemical overlap exists.
- SVM prob. performed consistently poorly, indicating that, despite its potential as a useful and accurate classifier in other studies, certain data contexts may produce suboptimal results.

3. Applying what we believe was the best compromise between performance and complexity, the RF/ANN average ensemble model, to a geochemical dataset from Eklutna Lake produced results in agreement with manual plotting and correlation. Other conclusions from this case study include:

- Where tephra layers are correlated for the first time, the chronologies of known eruptions from these intervals lend credibility and context to the model predictions.
- Even when tephras not present in the training set are encountered, predictions can still be reliable, though this is most effective a) when tephras are geochemically consistent between eruptions, and b) the training set is appropriately representative to include the variability and sources present.
- Special care should be taken when maximum probabilities are particularly low (e.g. 1/C; C = number of classes). But even at values much higher than that, misclassifications may be more frequent if the training data lack the unknown's eruption. Plotting and statistical tools can help assess this scenario.

Continuing comments on the Eklutna Lake case study beyond the technical application of classifiers, we demonstrated that the tephra layers are predominantly products from Redoubt Volcano, showing that the Anchorage area has repeatedly been impacted by ashfall from this Cook Inlet volcano. Such impact is predictable given the volcano's proximity but using magnetic susceptibility as the primary indicator of tephras may have skewed the tephra record somewhat because tephras from Redoubt tend to be richer in Fe-bearing minerals than those from some other regional volcanoes. Nonetheless, ongoing field studies at Redoubt do suggest that it is the most active of the Cook Inlet volcanoes in terms of the number of tephra falls preserved in Holocene-age sediments (e.g. Schiff *et al.*, 2010). Although the late Pleistocene Dawson tephra does appear as a detrital population in several samples, Tephra 11 appears to be a true primary tephra, representing a previously undocumented eruptive geochemically identical to Emmons' Dawson tephra, but dating to around ~1580 cal a BP. There exist many poorly documented eruptions and poorly characterised tephra deposits from Alaska, and even among those characterised, most have not been correlated to source (Mangan *et al.*, 2003). Still, it seems unlikely that Tephra 11 is a product from the Emmons Lake Volcanic Center unless rapid changes in melt characteristics occurred between ~1580 cal a BP and modern times. Overall, the varved chronology from Eklutna Lake bottom sediments presents a unique opportunity to develop a more complete understanding of ashfall hazards for the Anchorage area and is worth further examination.

Finally, as increasingly complete glass compositional databases are developed, the potential for employing efficient and accurate predictive models for tephra classification increases concurrently. We have shown that machine learning algorithms have great capacity to discern the sources of tephras from the chemically diverse and complex late Quaternary record of Alaska. Ours is the first

large-scale comparative study of machine learning for classifying glass geochemistry. However, work on discriminating tephras at a finer resolution (i.e. per eruption) is ongoing. Such studies are important for evaluating discriminatory power on increasingly similar geochemical populations, and where decision boundaries may be even less clear. But this work is just one component of the expanding computational toolset available to tephrochronologists. The potential for machine learning in this application will depend on its adoption by researchers, and especially, adaptation to new problems.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web-site.

Models S1. Archived final R models.

Appendix S1. Expanded machine learning methodological rationale, procedure, and discussion.

Script S1. Example R script for data pre-processing, classifier training, and prediction.

Figure S1. Sample stratigraphy from Eklutna Lake core localities 1 and 2.

Table S1. Training data composition (volcanic sources, eruptions, samples, and number of analyses).

Table S2. Summary of learners with key references.

Table S3. Eklutna Lake glass geochemical data.

Acknowledgements. This research was funded in part by a Research Foundation – Flanders (FWO-Vlaanderen) grant (G042812N) to M De Batist, an NSF award (1602106) to D. Kaufman and D. Fortin, a Natural Sciences and Engineering Research Council (NSERC) Discovery Grant to B. J. L. Jensen, and an NSERC Canada Graduate Scholarship-Master's (CGS M) to M. S. M. Bolton. We thank three anonymous reviewers and Roger Denlinger (U.S. Geological Survey) for their insightful and constructive comments that greatly benefitted this paper. We also thank the numerous researchers who contributed to this work by putting in many hours in the field, lab, operating the microprobe, and entering data into databases. Without their hard work, glass compositional studies and comparisons such as this would be impossible. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Altman DG. 1990. *Practical Statistics for Medical Research*. Chapman and Hall/CRC: London.

Arlot S, Celisse A. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys* **4**: 40–79.

le Bas MJ, le Maitre RW, Streckeisen A, *et al*. 1986. A chemical classification of volcanic rocks based on the total alkali-silica diagram. *Journal of Petrology* **27**: 745–750.

Beaudoin AB, King RH. 1986. Using discriminant function analysis to identify Holocene tephras based on magnetite composition: a case study from the Sunwapta Pass area, Jasper National Park. *Canadian Journal of Earth Sciences* **23**: 804–812.

Begét JE, Nye CJ. 1994. Postglacial eruption history of Redoubt Volcano, Alaska. *Journal of Volcanology and Geothermal Research* **62**: 31–54.

Begét JE, Stihler SD, Stone DB. 1994. A 500-year-long record of tephra falls from Redoubt Volcano and other volcanoes in upper Cook Inlet, Alaska. *Journal of Volcanology and Geothermal Research* **62**: 55–67.

Boes E, Van Daele M, Moernaut J, *et al*. 2018. Varve formation during the past three centuries in three large proglacial lakes in south-central Alaska. *Geological Society of America Bulletin* **130**: 757–774.

Bourne AJ, Lowe JJ, Trincardi F, *et al*. 2010. Distal tephra record for the last ca 105,000 years from core PRAD 1-2 in the central Adriatic Sea: implications for marine tephrostratigraphy. *Quaternary Science Reviews* **29**: 3079–3094.

Bull KF, Cameron C, Coombs ML, *et al*. 2012. The 2009 Eruption of Redoubt Volcano, Alaska. In *Report of Investigations of the Alaska Department of Natural Resources, Division of Geological & Geophysical Surveys 2011-5*, Schaefer JR (ed). State of Alaska, Department of Natural Resources: Fairbanks, AK.

Cameron CE, Schaefer JR. 2016. Historically Active Volcanoes of Alaska: Alaska Division of Geological & Geophysical Surveys Miscellaneous Publication 133 v. 2. Alaska Division of Geological & Geophysical Surveys: Fairbanks, AK.

Charman DJ, Grattan J. 1999. An assessment of discriminant function analysis in the identification and correlation of distal Icelandic tephras in the British Isles. *Geological Society, London, Special Publications* **161**: 147–160.

Clopper CJ, Pearson ES. 1935. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**: 404–413.

Combellick RA, Pinney DS. 1995. Radiocarbon age of probable Hayes tephra, Kenai Peninsula, Alaska. *Short Notes on Alaska Geology, Alaska Division of Geological & Geophysical Surveys Professional Report PR0117*, 1–9.

Cordeiro FJB. 1910. The volcanoes of Alaska. *Appalachian Journal* **12**: 130–135.

Davies LJ, Jensen BJL, Froese DG, *et al*. 2016. Late Pleistocene and Holocene tephrostratigraphy of interior Alaska and Yukon: Key beds and chronologies over the past 30,000 years. *Quaternary Science Reviews* **146**: 28–53.

Fernández-Delgado M, Cernadas E, Barro S, *et al*. 2014. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research* **15**: 3133–3181.

Fierstein J, Hildreth W. 2008. Kaguyak dome field and its Holocene caldera, Alaska Peninsula. *Journal of Volcanology and Geothermal Research* **177**: 340–366.

Fierstein J, Hildreth W, Hendley JW II, *et al*. 1998. Can another great volcanic eruption happen in Alaska? US Geological Survey Fact Sheet 075-98.

Fortin D, Praet N, McKay NP, *et al*. 2019. New approach to assessing age uncertainties – The 2300-year varve chronology from Eklutna Lake, Alaska (USA). *Quaternary Science Reviews* **203**: 90–101.

Froese D, Westgate J, Preece S, *et al*. 2002. Age and significance of the Late Pleistocene Dawson tephra in eastern Beringia. *Quaternary Science Reviews* **21**: 2137–2142.

Galar M, Fernández A, Barrenechea E, *et al*. 2011. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition* **44**: 1761–1776.

Hastie T, Tibshirani R, Friedman J. 2009. *The Elements of Statistical Learning*. Springer: New York.

Hildreth W, Fierstein J. 2012. The Novarupta-Katmai eruption of 1912—largest eruption of the twentieth century; centennial perspectives. *U.S. Geological Survey Professional Paper* **1791**.

Hsu C-W, Lin C-J. 2002. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks* **13**: 415–425.

Jensen BJL, Froese DG, Preece SJ, *et al*. 2008. An extensive middle to late Pleistocene tephrochronologic record from east-central Alaska. *Quaternary Science Reviews* **27**: 411–427.

Jensen BJL, Reyes AV, Froese DG, *et al*. 2013. The Palisades is a key reference site for the middle Pleistocene of eastern Beringia: new evidence from paleomagnetics and regional tephrostratigraphy. *Quaternary Science Reviews* **63**: 91–108.

Karatzoglou A, Smola A, Hornik K, *et al*. 2004. kernlab-an S4 package for kernel methods in R. *Journal of Statistical Software* **11**: 1–20.

Knerr S, Personnaz L, Dreyfus G. 1990. Single-layer learning revisited: a stepwise procedure for building and training a neural network, *Neurocomputing*, NATO ASI Series (Series F: Computer and Systems Sciences), Springer: Berlin Heidelberg; 41–50.

Kuehn SC, Froese DG, Shane PAR. 2011. The INTAV intercomparison of electron-beam microanalysis of glass by tephrochronology laboratories: Results and recommendations. *Quaternary International* **246**: 19–47.

Kuhn M. 2008. Building predictive models in R using the caret package. *Journal of Statistical Software* **28**: 1–26.

Kuhn M, Johnson K. 2013. *Applied Predictive Modeling*. Springer: New York, NY.

Kuhn M, Quinlan R. 2018. *C50: C5.0 Decision Trees and Rule-Based Models*. R package version 0.1.2. Retrieved from https://CRAN.R-project.org/package=C50

Kuncheva LI. 2004. *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley & Sons: Hoboken, NJ.

Liaw A, Wiener M. 2002. Classification and regression by randomForest. *R news* **2**: 18–22.

Loso M, Finney B, Johnson R, *et al*. 2017. Evaluating evidence for historical anadromous salmon runs in Eklutna Lake, Alaska. *Arctic* **70**: 259–272.

Lowe DJ. 2011. Tephrochronology and its application: A review. *Quaternary Geochronology* **6**: 107–153.

Lowe DJ, Pearce NJG, Jorgensen MA, *et al*. 2017. Correlating tephras and cryptotephras using glass compositional analyses and numerical and statistical methods: Review and evaluation. *Quaternary Science Reviews* **175**: 1–44.

Majka M. 2019. *naivebayes: High Performance Implementation of the Naive Bayes Algorithm in R*. R package version 0.9.6. Retrieved from https://CRAN.R-project.org/package=naivebayes

Malmgren B., Nordlund U. 1996. Application of artificial neural networks to chemostratigraphy. *Paleoceanography* **11**: 505–512.

Mangan MT, Waythomas CF, Miller TP, Trusdell FA. 2003. Emmons Lake Volcanic Center, Alaska Peninsula: source of the Late Wisconsin Dawson tephra, Yukon Territory, Canada. *Canadian Journal of Earth Sciences* **40**: 925–936.

McGimsey RG, Neal CA, Riley CM. 2001. Areal distribution, thickness, mass, volume, and grain size of tephra-fall deposits from the 1992 eruptions of Crater Peak vent, Mt. Spurr Volcano, Alaska. *U.S. Geological Survey Open-File Report 01-370*.

Mulliken KM. 2016. *Holocene volcanism and human occupation in the middle Susitna River Valley, Alaska*. M.A. thesis. University of Alaska Fairbanks: Fairbanks, AK.

Naes T, Mevik B-H. 2001. Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics* **15**: 413–426.

Ng AY, Jordan MI. 2002. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. In *Advances in Neural Information Processing Systems*, Dietterich TG, Becker S, Ghahramani ZMIT Press, 841–848.

Niculescu-Mizil A, Caruana R. 2005. Predicting good probabilities with supervised learning, *Proceedings of the 22nd International Conference on Machine Learning*. ACM: New York, NY; 625–632.

Payne RJ, Blackford JJ. 2008. Extending the late Holocene tephrochronology of the central Kenai Peninsula, Alaska. *Arctic* **61**: 243–254.

Pearce NJG, Bendall CA, Westgate JA. 2008. Comment on "Some numerical considerations in the geochemical analysis of distal microtephra" by A.M. Pollard, S.P.E. Blockley and C.S. Lane. *Applied Geochemistry* **23**: 1353–1364.

Petrelli M, Bizzarri R, Morgavi D, *et al*. 2017. Combining machine learning techniques, microanalyses and large geochemical datasets for tephrochronological studies in complex volcanic areas: New age constraints for the Pleistocene magmatism of central Italy. *Quaternary Geochronology* **40**: 33–44.

Petrelli M, Perugini D. 2016. Solving petrological problems through machine learning: the study case of tectonic discrimination using geochemical and isotopic data. *Contributions to Mineralogy and Petrology* **171**: 81.

Platt J. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, Smola A, Bartlett P, Schölkopf D, Schuurmans D (eds). MIT Press: Cambridge, MA.

Pouget S, Bursik M, Rogova G. 2014. Tephra redeposition and mixing in a late-glacial hillside basin determined by fusion of clustering analyses of glass-shard geochemistry. *Journal of Quaternary Science* **29**: 789–802.

Praet N, Moernaut J, Van Daele M, *et al*. 2017. Paleoseismic potential of sublacustrine landslide records in a high-seismicity setting (south-central Alaska). *Marine Geology* **384**: 103–119.

Preece SJ, McGimsey RG, Westgate JA, *et al*. 2014. Chemical complexity and source of the White River Ash, Alaska and Yukon. *Geosphere* **10**: 1020–1042.

Preece SJ, Westgate JA, Froese DG, *et al*. 2011. A catalogue of late Cenozoic tephra beds in the Klondike goldfields and adjacent areas, Yukon Territory. *Canadian Journal of Earth Sciences* **48**: 1386–1418.

Preece SJ, Westgate JA, Stemper BA, *et al*. 1999. Tephrochronology of late Cenozoic loess at Fairbanks, central Alaska. *Geological Society of America Bulletin* **111**: 71–90.

Ramsey CB. 2009. Bayesian analysis of radiocarbon dates. *Radiocarbon* **51**: 337–360.

R Core Team. 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Reimer PJ, Bard E, Bayliss A, *et al*. 2013. IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal BP. *Radiocarbon* **55**: 1869–1887.

Riehle JR. 1985. A reconnaissance of the major Holocene tephra deposits in the upper Cook Inlet region, Alaska. *Journal of Volcanology and Geothermal Research* **26**: 37–74.

Rifkin R, Klautau A. 2004. In defense of one-vs-all classification. *Journal of Machine Learning Research* **5**: 101–141.

Schiff CJ, Kaufman DS, Wallace KL, *et al*. 2010. An improved proximal tephrochronology for Redoubt Volcano, Alaska. *Journal of Volcanology and Geothermal Research* **193**: 203–214.

Schölkopf B, Platt JC, Shawe-Taylor J, *et al*. 2001. Estimating the support of a high-dimensional distribution. *Neural Computation* **13**: 1443–1447.

Scott WE, McGimsey RG. 1994. Character, mass, distribution, and origin of tephra-fall deposits of the 1989–1990 eruption of redoubt volcano, south-central Alaska. *Journal of Volcanology and Geothermal Research* **62**: 251–272.

Sebestyen GS. 1962. *Decision-making Processes in Pattern Recognition*, ACM monograph series, Macmillan: Indianapolis, IN.

Shane PAR, Froggatt PC. 1994. Discriminant function analysis of glass chemistry of New Zealand and North American tephra deposits. *Quaternary Research* **41**: 70–81.

Sheppard PJ, Irwin GJ, Lin SC, *et al*. 2011. Characterization of New Zealand obsidian using PXRF. *Journal of Archaeological Science* **38**: 45–56.

Stelling P, Gardner JE, Begét JE. 2005. Eruptive history of Fisher Caldera, Alaska, USA. *Journal of Volcanology and Geothermal Research* **139**: 163–183.

Stokes S, Lowe DJ. 1988. Discriminant function analysis of late Quaternary tephras from five volcanoes in New Zealand using glass shard major element chemistry. *Quaternary Research* **30**: 270–283.

Stokes S, Lowe DJ, Froggatt PC. 1992. Discriminant function analysis and correlation of Late Quaternary rhyolitic tephra deposits from Taupo and Okataina volcanoes, New Zealand, using glass shard major element composition. *Quaternary International* **13-14**: 103–117.

Therneau T, Atkinson B. 2018. *rpart: Recursive Partitioning and Regression Trees*. Retrieved from https://CRAN.R-project.org/package=rpart

Tryon CA, Logan MAV, Mouralis D, *et al*. 2009. Building a tephrostratigraphic framework for the Paleolithic of Central Anatolia, Turkey. *Journal of Archaeological Science* **36**: 637–652.

Venables WN, Ripley BD. 2002. *Modern Applied Statistics with S*. Springer: New York.

Waitt RB, Begét JE. 2009. Volcanic processes and geology of Augustine Volcano, Alaska. *U.S. Geological Survey Professional Paper* **1762**.

Wallace K, Coombs ML, Hayden LA, Waythomas CF. 2014. Significance of a near-source tephra-stratigraphic sequence to the eruptive history of Hayes Volcano, south-central Alaska. *U.S. Geological Survey Scientific Investigations Report 2014-5133*.

Waythomas CF, Haney MM, Wallace K, Cameron CE, Schneider DJ. 2017. The 2014 eruptions of Pavlof Volcano, Alaska. *U.S. Geological Survey Scientific Investigations Report 2017-5129*.

Wold S, Sjöström M. 1977. Chemometrics: Theory and Application, *SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy*, ACS Symposium Series, Vol. 52.

Wolpert DH. 1992. Stacked generalization. *Neural Networks* **5**: 241–259.

Wu T-F, Lin C-J, Weng RC. 2004. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research* **5**: 975–1005.