Application of discriminant function analysis and change-point problem in dating volcanic ashes

B. W. Lee and J. Bacon-Shone
Social Sciences Research Centre, The University of Hong Kong; *h0002750@hkusua.hku.hk*

Abstract

The application of Discriminant function analysis (DFA) is not a new idea in the study of tephrochrology. In this paper, DFA is applied to compositional datasets of two different types of tephras from Mountain Ruapehu in New Zealand and Mountain Rainier in USA. The canonical variables from the analysis are further investigated with a statistical methodology of change-point problems in order to gain a better understanding of the change in compositional pattern over time. Finally, a special case of segmented regression has been proposed to model both the time of change and the change in pattern. This model can be used to estimate the age for the unknown tephras using Bayesian statistical calibration.

Kew words: Tephrochrology; Discriminant function analysis (DFA); Change-point problem; Bayesian statistical calibration; Segmented regression.

1 Introduction

In the past, tephrochrology, the dating of volcanic eruptions by the study of tephras (volcanic ashes), relied largely on radiocarbon dating to suggest a likely candidate eruption followed by comparing the geochemical characteristics of unknown tephras with reference to tephras from a suspected source using mean and standard deviations of major oxides, plus binary and ternary plots of the selected major oxides (Charman & Grattan, 1999). Clearly, such a type of analysis does not allow the dating of volcano ashes directly, as it still requires a high input of radiocarbon analyses to provide an initial likely candidate eruption and it is very dependent on the accuracy of those age estimates. Besides, as more tephras are discovered, the pattern of deposition becomes more complex, so this approach is no longer working as effectively as before. Moreover, this analysis is too subjective and relies on the judgement of individual researchers. Different geologists may select different sub-compositions to compare in the analysis due to the absence of clear guideline for the selection of sub-compositions, and different conclusions may be drawn. It does not provide any quantitative assessment of the best discriminating oxides or the probability of correct identification of a given tephra. The robustness of such subjective comparisons is surely in doubt. Although this approach may be useful as an ad-hoc comparison, it does not properly utilize the full complement of geochemical information available. Therefore, geologists always proposed the use of statistical technique to make the process more scientific and objective.

The application of DFA is not a new idea in the field (Borchardt & others, 1972; Stokes & Lowe, 1988; Forggatt, 1992; Charman & Gartten, 1999). The analysis consists of two parts (Khattree & Naik, 2000; Timm, 2002). In the first part, a series of discriminant functions, which are linear combinations of the major oxides, is derived from the reference set of data from a series of known tephras. These functions form the classification model to be applied to the analysis. The linear combination of the major oxides allows a more effective use of the information inhered in the compositional data set. Usually, the first several functions are already enough to explain up to 90% of the variation (Charman & Gartten, 1999). Therefore, instead of working with a large number of oxides scores, one or two canonical variables containing most of the chemistry information are analyzed to understand the difference in compositions. Differences between each pair of known tephras can be measured by the Mahalanobis distance squared statistics (D^2). The separation of the tephras can be displayed graphically on discriminant function axes. In the second part of the analysis, these derived discriminant functions were applied to geological data from unknown tephras. With this type of analysis, unknown tephras could be identified to the reference set with an objective probability of misclassification.

Results of the first part of analysis in four different compositional datasets have been demonstrated in the third section, whereas the information about the four datasets and the manipulation of the data before analysis is found in the coming section. The canonical variables from the analysis have been further investigated. It shows a constantly repeating pattern for the first canonical variable over time among the four datasets. The variable decreases linearly over time, jumps abruptly after a period of time, and then decreases linearly with a similar slope again. This suggests that the composition inside the volcano changes over time in a constant manner with an event occurs that changes the base composition; probably these may be related to the injection of raw material inside the earth. Change-point analysis (Chen & Gupta, 2000) can be applied here in order to model the repeated changing pattern. It may be useful to estimate the time of abrupt jump, the level of jump and the pattern between each jump, as these may give information about the occurrence of volcano eruption. A special case of segmented regression (Mahmoud, 2004) has been proposed in the fourth section. This model can be used to estimate the age of unknown tephras by calibration (Alfassi & others, 2005). The second part of DFA and estimation of age of unknown tephras by calibration with the first canonical variable have been done for one of the four datasets. Comparison of results is found in the fifth section. Finally, conclusion for the whole research is given in the sixth section.

2 Data description and data pre-treatment

Four compositional datasets have been studied in the analysis. They represent the constituencies of major oxides of two different types of tephras, black ashes and pumice layers from Mountain Ruapehu in New Zealand and Mountain Rainier in USA.

2.1 Data description

Ten most commonly used major oxides (Shane & Froggatt, 1994; Cronin and others; 1996) have been chosen to be analyzed in the study. The average chemistry for the two types of tephras from the two mountains is shown in the following table.

Table 1. Average chemistry for the two types of tephras from the two mountains.

Oxide wt. %	New Zealand				U. S. A			
	Black ashes		Pumice layers		Black ashes		Pumice layers	
	Mean	s. d.	Mean	s. d.	Mean	s. d.	Mean	s. d.
SiO2	65.01	2.21	65.64	5.06	69.53	6.05	64.09	5.46
Al2O3	14.61	0.80	15.10	1.52	14.47	2.26	15.49	1.81
TiO2	1.12	0.15	0.80	0.28	1.05	0.48	1.16	0.33
FeO	6.14	0.70	4.59	2.03	3.15	1.66	5.15	1.57
MnO	0.21	0.13	0.16	0.09	0.14	0.09	0.17	0.10
MgO	1.66	0.68	1.54	1.13	1.01	1.27	2.04	1.37
CaO	4.14	0.95	3.91	1.79	2.86	1.89	4.25	1.97
Na2O	3.11	0.55	3.24	0.50	3.88	1.08	3.55	0.50
K2O	2.90	0.56	2.98	0.87	3.15	1.04	2.53	0.82
Cl	0.13	0.08	0.19	0.08	0.12	0.15	0.16	0.08
sample size	238		130		324		145	

To allow comparisons of the compositional patterns over time, there should be enough samples of tephra units from different time in each reference set and number of individual shards in each tephra (Table 2). Besides the four reference sets, the unknown set for New Zealand black ashes has also been included to demonstrate the second part of DFA and the estimation of the unknown age by calibration. A very approximate age has been given as a reference for the calibration.

Table 2. Information about each tephra unit for the four reference sets.

New Zealand						U. S. A.					
Black ashes			Pumice layers			Black ashes			Pumice layers		
Tephra unit	age (ka)	*n*	Tephra unit	age (ka)	*n*	Tephra unit	age (ka)	n	Tephra unit	age (ka)	*n*
TF19	0.01	17	OK(6)	10.1	3	R029	0.4	33	R310	4.7	23
TF14	0.4	11	OK(Mg)	10.1	12	R0930	0.4	9	R036	5	9
TF10	0.5	12	OK2	10.1	13	R021	0.5	20	R05	5	5
TF9	0.55	20	OK3	10.1	10	R0820	1.1	32	R301	5.5	26
TF8	0.6	8	BL17	10.8	11	R08	2.3	28	R013	6	14
TF7	0.63	19	BL15	12	17	R024	2.5	22	R035	6	25
TF6	0.65	16	BL13	13	21	R0763	2.5	10	R015	6.4	8
TF5	0.7	15	BL11	14	9	R064	2.53	12	R037	6.4	11
TF4	0.83	26	BL6	16	10	R065	2.55	26	R025	6.5	13
TF2	1.8	10	BL5	17	10	R031	2.7	22	R02000	8.75	11
			BL4	17.9	4	R03000	3	18			
			BL3	19	10	R09	4.8	27			
						R034	6.43	20			
						R032	6.48	30			
						R01020	6.55	15			
	Sum	144		Sum	130		Sum	324		Sum	145

n=no. of individual shards

Table 3. Approximate age and no. of individual shards in each tephra unit for the unknown set of New Zealand black ashes.

Tephra unit	appro. Age (ka)	*n*
NZ08	0.7	5
NZ359	0.75	17
NZ361	0.76	25
NZ374	0.8	23
NZ500	0.5	14
	Sum	84

2.2 Data pre-treatment

One of the conflicting issues in comparison of geochemical microprobe data is that of the data pre-treatment (Charman & Grattan, 1999). It is basically due to the very important characteristic of compositional vector; each variable represents a proportion of some whole. Therefore, all value sum up to a constant, which is the unit-sum constraint as mentioned by Aitchison (1986). Since then, lot of researchers have started to doubt about the correctness of applying standard unconstrained statistical analysis directly to the constrained compositional dataset (Aitchison, 2003). Obviously, manipulation of the dataset should be done prior to the analysis in order to switch the compositional dataset from its sample space to the real number space, which is assumed in our simple multivariate analysis.

2.2.1 Log-ratio transformation

According to Charman and Gratten (1999), data is usually expressed as either raw percentage data or normalized to 100% total. Hunt & Hill (1993) suggest the former while the INQUA guideline recommends the latter (Forgatt, 1992). Whichever approach is adopted, the famous problem of a constant sum constrict exists (Aitchison, 1983; Stokes & Lowe, 1988). Aitchison (1983) suggest the use of log-ratio transformations, the additive log-ratio transformation $alr(x) = [log(x_1/x_D)\ log(x_2/x_D)\ \dots\ log(x_{D-1}/x_D)]$ and the centered log-ratio transformation $clr(x) = [log(x_1/g(x))\ log(x_2/g(x))\ \dots\ log(x_D/g(x))]$, where $g(x)$ denotes the geometric mean of the D components of the major oxides x.

In the centered log-ratio transformation, it has the disadvantage that the covariance matrix formed based on such transformation is singular, while the operational problem of the additive log-ratio transformation is that a common divisor has to be chosen. As proved in the monograph of Aitchison (1986), the choice of common divisors would not affect the results of analysis due to scale invariance property, so the choice could be arbitrary, but the clear disadvantage of the additive log-ratio transformation is that the chosen common divisor could not be used in the analysis. Therefore, some geologists resist the log-ratio transformation and continue to analyze the data with the pathological approach (Aitchison, 2003). To allow consensus between statisticians and geologists, the selected common divisor for the log-ratio transformation should be of moderate abundance and relatively small variance. As a result, Cl was chosen to be the common divisor (table 1).

2.2.2 Treatment of missing data

Another common problem in dealing with compositional data analysis is the problem of missing data, rounded or trace zeros for those missing due to a very infinitesimal value below detection level and essential zeros for those which is truly zero (Aitchison, 1986). A common approach suggested in the monograph of Aitchison (1986) is simply replacing the zeros by half of the lowest possible value observed. More advanced and sophisticated zero replacement strategies have been introduced by other researchers recently (Fry and others, 2000) and even in last conference (Aitchison & Kay; 2003; Bacon-Shone; 2003; Martin-Fernández and others, 2003). However, as shown in the analysis by Stokes and Lowe (1988), the presence of multivariate outliers had only minor effects on the performance of discriminant function procedure. Therefore, the simple replacement method is still in use by most geologists and also in this study.

3 Results of the first part of DFA

The DFA performs quite well in all the four datasets. It could discriminate between those tephra units in all the four data sets even with just the first two components. The first two canonical variables explain up to nearly 80% (table 4) of the variation in the compositional pattern. Moreover, it is obvious that there is a clear pattern of moderate changing patterns of the first canonical variable for all the four cases, while there is not a common pattern observed from the case for the second canonical variable. It appears that the first canonical variable decreases linearly with time, jumps abruptly at some time point, and then decreases linearly again. This fact is very useful for setting a mathematical model to analyze the change of compositional pattern over time.

Table 4. Cumulative proportion of variation explained by the first two canonical variables.

cum. prop. of explained variation	New Zealand		U. S. A.	
	Black ashes	Pumice layers	Black ashes	Pumice layers
1	0.5265	0.7268	0.6724	0.6394
2	0.7592	0.83	0.8357	0.8616

3.1 Black ashes from New Zealand

From the figure 1A, at the first sight, it seems as if that the first canonical variable decreases from time = 0 to time = 0.83, and then increases again to the same level as the beginning at time = 1.8, but the problem is that no data is actually available between time = 0.83 and time = 1.8. Therefore, we just focused on the pattern up to time = 0.83 (fig. 1B). It is obvious that there are two jump-points, one is between time = 0.55 to time = 0.6, the other one is after time = 0.83, but there may also be one jump-point between time = 0.01 to time = 0.4, so there may be two or three change points existed in this cases. For the second canonical variable, the scatter plot is just like the bottom part of an upward-opened curve of a quadratic function, the value decreases moderately to the minimum point, and then increases moderately again. There is not a clear change point over time.

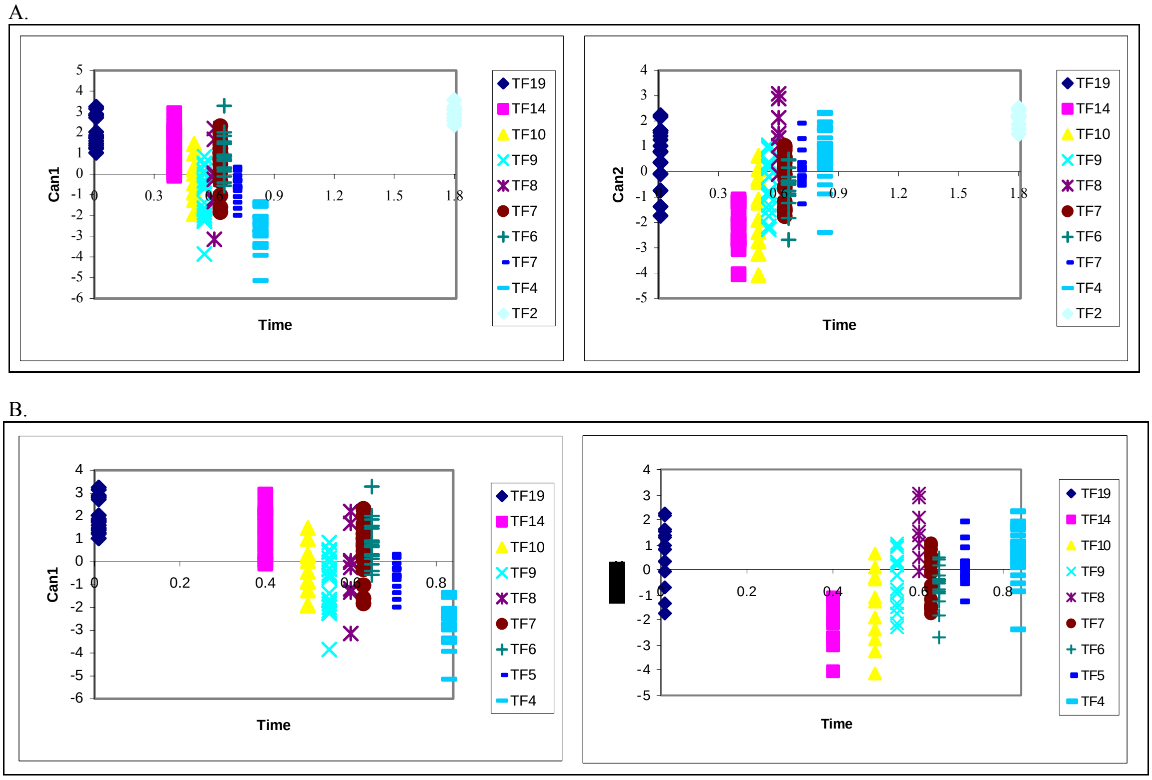

Figure 1: Scatter plots of the first two canonical variables against time for New Zealand black ashes: A. with time = 1.8 B. without time = 1.8

3.2 Pumice layers from New Zealand

For the first canonical variable, three obvious change-points could be noted, the first one is between time = 10.8 and time = 12, the second one is between time = 13 and time = 14, and the last one is between time = 17.9 and time = 19. Besides, there may be a change point between time = 14 and time = 16. Altogether, there may be three or four change points in the whole process.

For the second canonical variable, the pattern is similar to the one of New Zealand black ashes. It is a bit like the patterns of sine-cosine function, the value decreases moderately from a "peak" at time = 10 to the local minimum at time between 12 and 13, and then increases moderately again to the "peak" at time = 14. The pattern roughly repeats between time = 14 and time = 19.

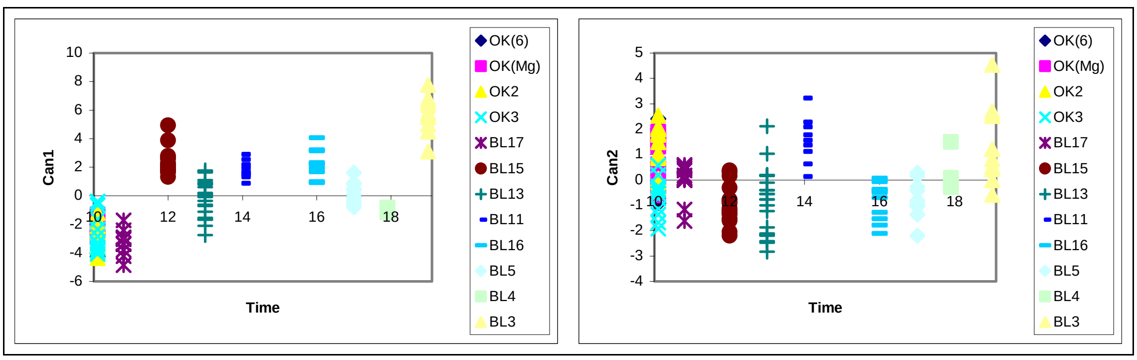

Figure 2: Scatter plots of the first two canonical variables against time for New Zealand pumice layers

3.3 Black ashes from U. S. A.

The pattern for both canonical variables is not very clear for the U. S. A. black ashes (fig. 3). It may be because there is not an evenly distributed time for the available tephra units. No much chemistry information is available after time = 3.5. For the first canonical variable, the changing pattern could be seen from time = 0.4 to time = 3.4, there should be a change-point between time = 1.1 to time = 2.3. After time = 3.4, the value for the first canonical variable seems to remain at the same level, while for the second variable, the value stays at similar level for all time points.

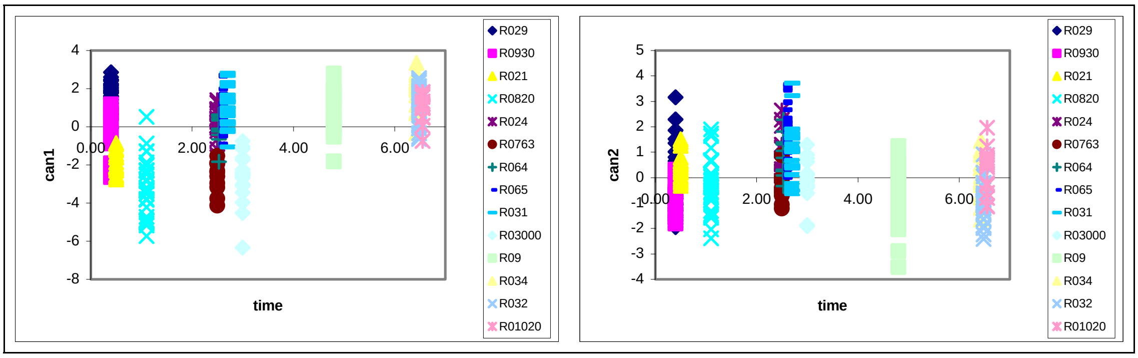

Figure 3: Scatter plots of the first two canonical variables against time for U. S. A. black ashes

3.4 Pumice layers from U. S. A.

For the first canonical variable, it is obviously that there is at least one change point, which may be between time = 5 to and time = 5.5 or between time = 5.5 to time = 6, but it seems as if that the value decrease continuously from time = 6 to the end point at time = 8.75, but it is inconclusive whether there is any change-points between time = 6.5 to time = 8.75 as no information is available. For the second variable, the pattern is different from those in pervious examples. It seems as if that the value decreases moderately and becomes stabilized at the end.

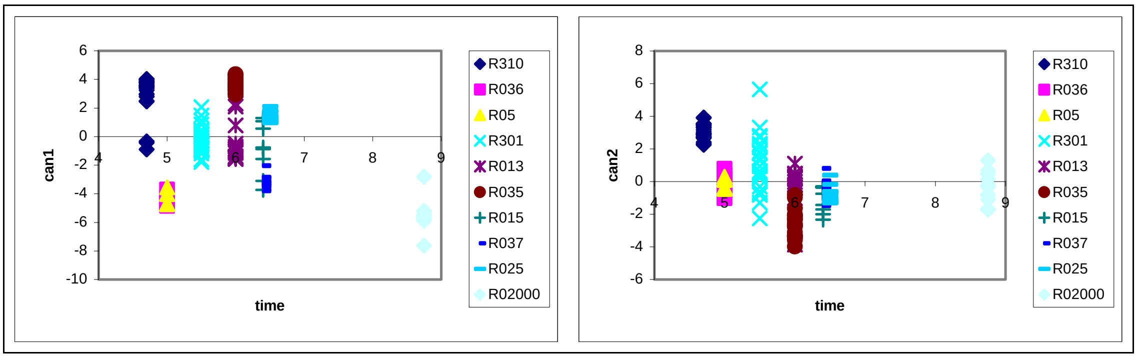

Figure 4: Scatter plots of the first two canonical variables verse time for U. S. A. pumice layer

4 Change point problems and calibration

In order to understand the changing pattern including the time of abrupt jump, the level of jump and the pattern between jump points, a specific mathematical model based on the methodology of change problems has been studied. This model can be used to estimate the age of unknown tephra by calibration.

4.1 Segmented regression technique found in literature

This model is commonly used in the field of industrial management or quality control in order to ensure a smooth process of production and stable quality of the products. The simple *s*-segmented piecewise regression model given in Mahmoud (2004) 's dissertation is that:

$$
\begin{aligned}
Y_i &= A_{01} + A_{11}X_i + \varepsilon_i,\ \theta_0 < i \le \theta_1 \\
Y_i &= A_{02} + A_{12}X_i + \varepsilon_i,\ \theta_1 < i \le \theta_2 \\
&\vdots \\
Y_i &= A_{0s} + A_{s1}X_i + \varepsilon_i,\ \theta_{s-1} < i \le \theta_s
\end{aligned}
\tag{1}
$$

where $i = 1, 2, \ldots, N$ and θ_j 's are the change points between segments (usually $\theta_0 = 0$ and $\theta_s = N$) and the ε_i 's are assumed to be i.i.d $N(0, \sigma_j^2)$, where σ^2_j is the segment error term variance, where $j = 1, 2, \ldots, s$. In simple terms, in segment regression modeling, the parameters of intercepts and slopes may change suddenly at some points. This technique could be used to estimate the number of segments s and the locations of the change point θ_j 's.

4.2 Proposed segment regression for the analysis

As mentioned in section three, a clear changing pattern is observed for the first canonical variable. In order to set a mathematical model for such a changing patterns, several assumptions have been given based on the observation.

- The slope parameter is constant over time.
- The jump time may be related to the occurrence of volcanic eruption, so it follows gamma distribution in the whole process.
- The change of intercept parameter is due to abrupt jump, which may be related to the level of volcanic eruption, so it follows normal distribution in the whole process.
- The initial time point is at time = 0

Based on the above assumptions, the following mathematical model has been set up:

$$
\begin{aligned}
Y_i &= A_{01} + A_1t_i + \varepsilon_i,\ R_0 < t_i \le R_1 \\
Y_i &= A_{02} + A_1t_i + \varepsilon_i,\ R_1 < t_i \le R_2 \\
&\vdots \\
Y_i &= A_{0s} + A_st_i + \varepsilon_i,\ R_{s-1} < i \le R_s
\end{aligned}
\tag{2}
$$

where $A_{0i} = A_{01} + c_1 + c_2 + \ldots + c_{i-1}$ for i from 2 to s and c_i is the change of intercept from $i - 1^{\text{th}}$ segment to i^{th} segment, that is the jump level which is assumed to follow iid $N(\mu_c, \sigma_c^2)$, where μ_c and σ_c^2 are the mean and variance of the jump level; ε_i 's are assumed to be iid $N(0, \sigma_j^2)$, where σ_j^2 is the segment error term variance as the same as the simple model quoted in previous section; R_0 is initial point of the process, which is assumed to be zero, R_i is the i^{th} change-point number of time points and $r_i = R_i - R_{i-1}$ is the length of time for the i^{th} segment, which is assumed to follow iid $\Gamma(\alpha, \beta)$, that is, $R_i \sim \Gamma(i\alpha, \beta)$.

4.3 Estimation of model parameters and calibration of age of unknown tephras

To fit the dataset, the parameters for all the distributions in the model have to be estimated. Bayesian approach (Lee, 2004) has been applied for the convenience of combining the procedure of estimation of the model parameters and calibration of age of the unknown tephra. Due to special characteristics of the segmented regression model, multiple solutions may be arrived in the calibration procedure. With Bayesian approach, those impossible solutions could be eliminated or pulled down by giving a suitable prior distribution to the calibrated age. The analysis could be easily done by WinBUGS, a piece of computer software for the Bayesian analysis of complex statistical models using MCMC method and Gibbs sampler, which could be downloaded in the website of the Biostatistics Unit of the Medical Research Council in the University of Cambridge (http://www.mrc-bsu.cam.ac/bugs). With such software, estimation for model parameters model and calibration of the age of unknown tephras could be done simultaneously by simulation based on the prior distribution and the model given. Non-informative prior have been set for all the parameters, while a uniform prior has been given to the calibration with adjusted lower and upper bound for each case of unknown tephra. The syntax for the analysis could be found in the appendix.

5 Aging the unknown tephras

As mention in the introduction section, to age the unknown tephra, the second part of the DFA is helpful in correlating the unknown tephra to the reference set. On the other hand, it could be done by calibration of the first canonical variable to the age of the tephra with the proposed segmented regression. The box-plot of the first canonical variable is shown in the following figure. It shows that NZ08 and NZ500 are far different from NZ359, NZ361 and NZ374. It could be further confirmed by finding the value of D^2 (table 5), NZ359, NZ361, and NZ374 are similar, but NZ08 and NZ500 are far away from the other tephras. The separation of the three different groups of unknown tephra is further shown by the scatter plot of the first two canonical variables (fig. 6).

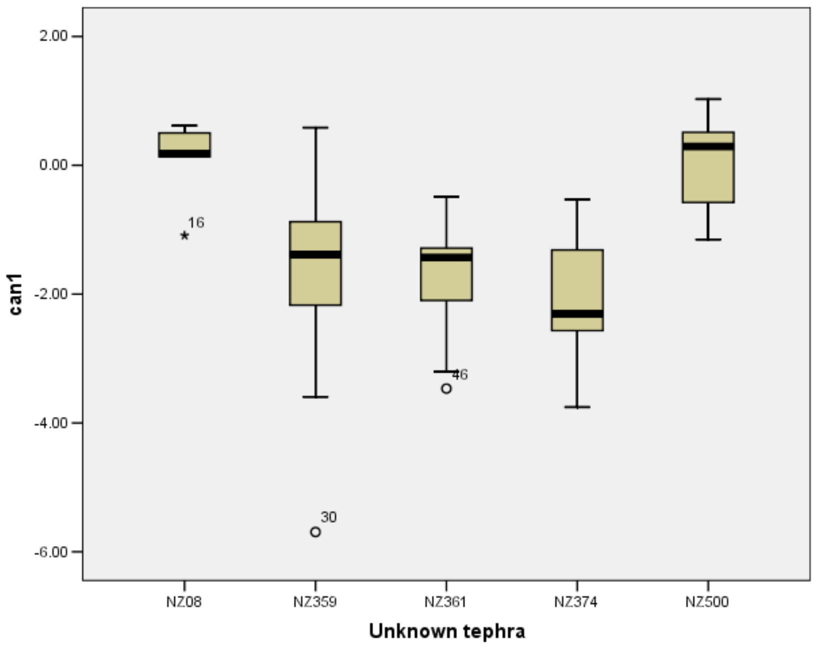

Figure 5: Box-plot of the first canonical variables for U. S. A. pumice layer

Table 5. Mahalanobis distance squared statistics (D^2) among unknown tephras. The cells with the largest four values are shaded in pink, while those with the least three values are shaded in orange

	NZ08	NZ359	NZ361	NZ374	NZ500
NZ08	0	8.99	9.17	20.45	78.16
NZ359		0	2.23	6.18	70.44
NZ361			0	1.89	115.26
NZ374				0	145.37
NZ500					0

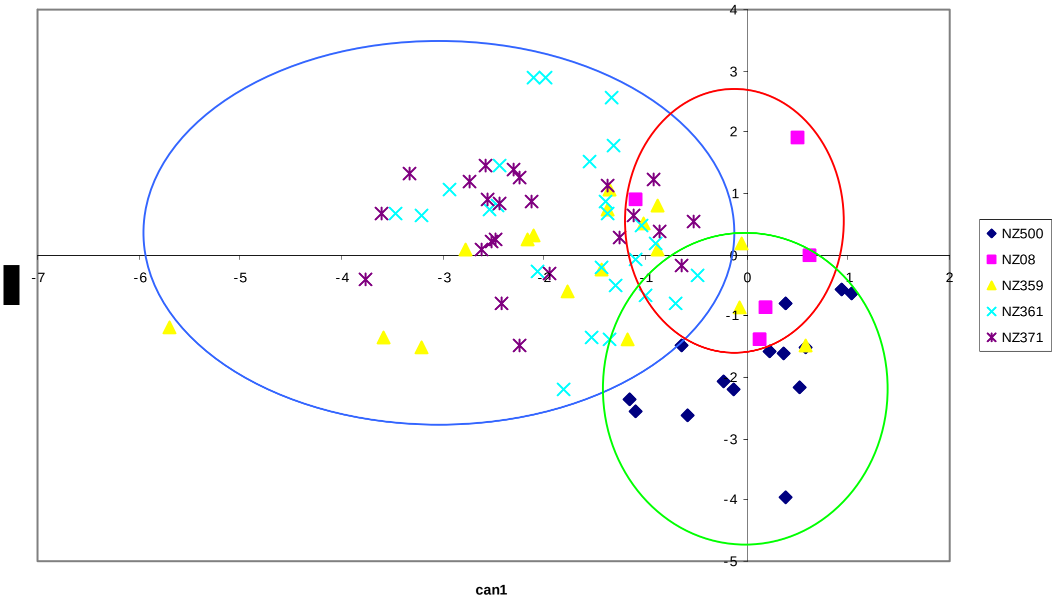

Figure 6: Scatter plot of the first two canonical variables for U. S. A. pumice layer. The three groups of unknown tephras are separated by the three ellipses.

5.1 Results of the second part of DFA

The assignation of the unknown tephra units to the reference set could be found in the following table, the classification of the unknown tephras to the reference set is not very clear in the analysis, more than one groups could be assigned to each tephra, and even for NZ08, which only possesses five tephra units. NZ500 could be assigned to three groups including TF10, TF9 or TF4; NZ361 could be assigned to TF9 or TF4, while NZ08 is assigned to TF9, NZ359 and NZ374 are assigned to TF4. It may be because TF4, TF9 and TF10 are close to each others (table 7). D^2 could be used as a criterion to further decide if the only one reference set is allowed to identify each unknown tephra.

Table 6. Assignation of the unknown tephra units to the reference set. The possible identified reference sets for each unknown tephra are highlighted in pink.

Sample	TF19	TF14	TF10	TF9	TF8	TF7	TF6	TF5	TF4
NZ500		1	4	5					4
NZ08				3				1	1
NZ359				6					11
NZ361				12					13
NZ374				6					17

Table 7. Mahalanobis distance squared statistics (D^2) between the unknown tephras and reference set and among the reference set. The cells with D^2 less than 5 are highlighted in pink.

D^2		Reference set							
		TF10	TF14	TF4	TF5	TF6	TF7	TF8	TF9
Unknown tephra	NZ08	21.60012	115.97753	15.31083	6.05909	65.63358	23.24747	533599	1.74661
	NZ359	19.64835	78.99815	1.11433	208.7856	194.40084	71.50327	122248	3.94546
	NZ361	27.40658	134.01035	1.44096	107.93689	182.5308	65.13286	63578	6.00691
	NZ374	20.1007	193.21254	3.8624	221.74358	178.51582	63.31685	6046	8.93996
	NZ500	8.79423	15.387	14.0904	115.09422	219.48576	92.22103	263250	3.73382
Reference set	TF10	0	28.62199	26.39923	38.71744	126.9224	60.21051	624562	4.736
	TF14		0	72.20409	57.25712	72.06714	32.59449	3422759	13.39212
	TF4			0	112.64177	162.24598	64.81576	87384	4.42305
	TF5				0	13.61556	13.25078	387530	6.64534
	TF6					0	12.2045	1245109	9.39774
	TF7						0	117299	7.08063
	TF8							0	6.4021
	TF9								0

5.2 Result of calibration with segmented regression model

The analysis could be divided into two parts, the estimation of the parameters for the model and the calibration of the age to the compositional pattern of unknown tephra.

5.2.1 Fitting the dataset

Two change points have been estimated by the WinBUGS based on the segmented regression model. They are 0.3396 and 06016, which are reasonable based on the observation as mention in section 3.1. The fitting of the data set is quite good as observed in the figure 7. The MSEs calculated based on the segmented regression line have also been quoted as reference.

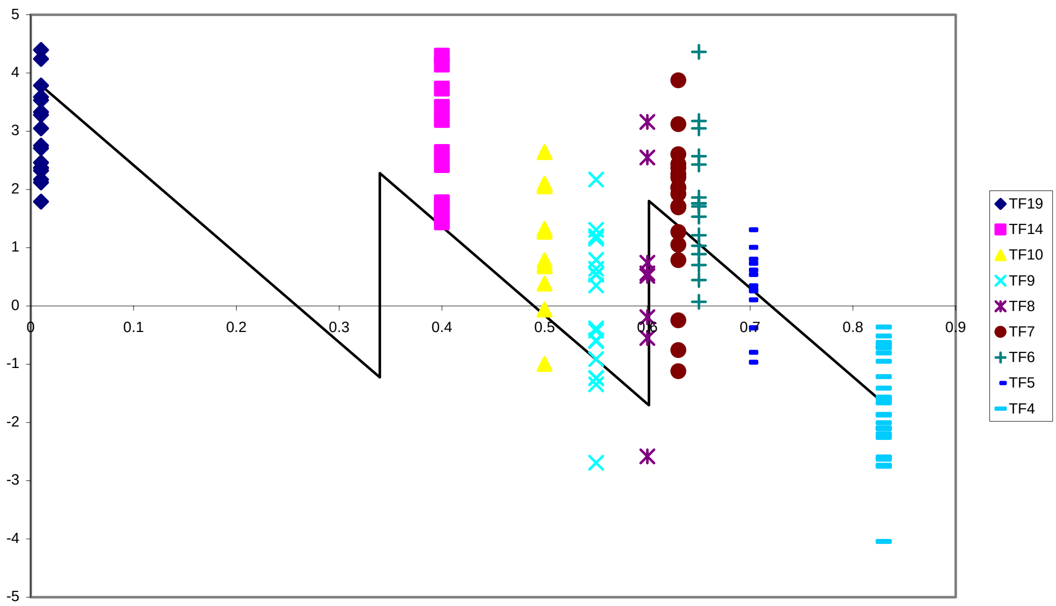

Figure 7: Scatter plot of the first canonical variables for U. S. A. black ashes (with the segmented regression line)

Table 8: MSE calculated based on the segmented regression model

MSE within	TF19	TF14	TF10	TF9	TF8	TF7	TF6	TF5	TF4
Group	1.2492	2.9608	2.2947	2.2674	7.6512	1.5672	1.698	0.3927	0.753
Whole reference set					1.8772				

5.2.2 Calibration of the unknown age

The kernel density for the calibrated age of the unknown tephras could be found in the following. As one the first canonical variable is used in the analysis, the result is quite similar for NZ500 and NZ08, as well as NZ359, NZ361 and NZ374. The possible calibrated age is picked up by the time at which the kernel density reached to the peak (table 8).

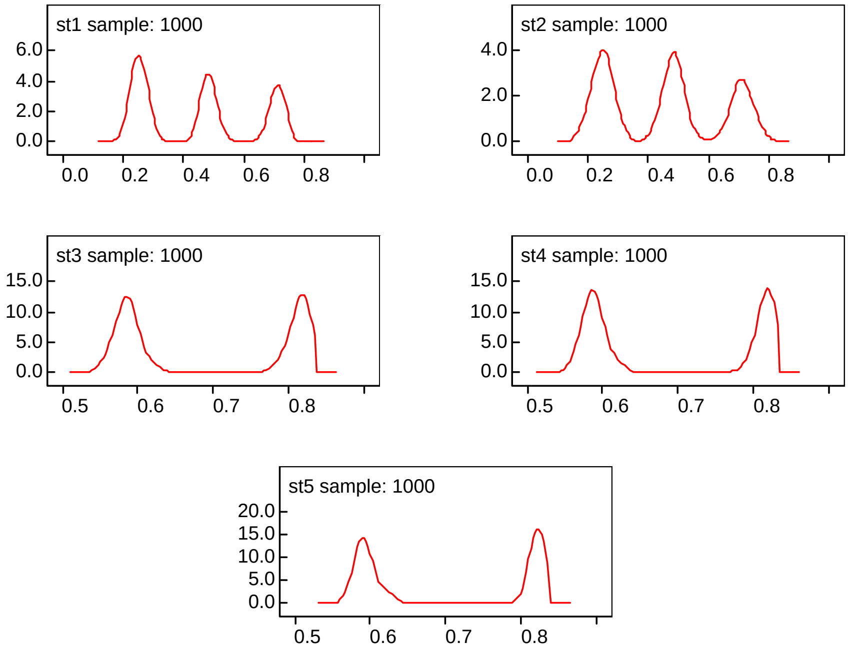

Figure 8: Kernel density of the calibrated age for the five unknown tephras (st1: NZ500, st2: NZ08, st3: NZ359, st4: NZ361, st5: NZ374)

Table 9: Possible calibrated age for the five unknown tephras

	Possible calibrated age		
NZ500	0.2558	0.4872	0.7153
NZ08	0.2509	0.4859	0.7173
NZ359	0.5868	0.8161	
NZ361	0.5894	0.8184	
NZ374	0.5947	0.8233	

5.3 Comparison of the results from the two methodologies

In fact, the two results agree with each other, as the possible calibrated age is not too far from the age of the identified reference tephra (table 9). However, the second approach has an advantage that it allows a better use of information available from the reference set. It could pick up value of age other than those from the reference set. For instance, although we have not get any reference tephra between time = 0.01 and time = 0.4, as the value of the variable could be estimated with the segmented regression, so it allows a possible calibrated age to be 0.2558, which is between 0.01 and 0.4.

Table 10: Comparison of the results for estimating the unknown age by the two methodologies. The agreed classification and estimation are shaded in the same colours.

	TF14	TF10	TF9	TF5	TF4	Possible Calibrated Age		
	0.4	0.5	0.6	0.7	0.83			
NZ500	1	4	5		4	0.2558	0.4872	0.7153
NZ08			3	1	1	0.2509	0.4859	0.7171
NZ359			6		11		0.5868	0.8161
NZ361			12		13		0.5894	0.8184
NZ374			6		17		0.5947	0.8233

6 Conclusion

In this paper, the main aim is to solve the geological problem of aging the unknown tephra. Two methodologies of identification to the reference set by DFA and calibration of the first canonical variable have been demonstrated. The statistical methodology of change-point problem is useful in understanding the changing pattern of the canonical variable over time and the proposed segmented regression developed based on the change-point analysis enhances a better use of the chemistry information available from the DFA. It is a more flexible estimation of the calibrated age other than just a mapping of the unknown tephra to the reference set. As shown in section 5.1, it is not advisable to put tephras with similar age and compositional pattern in the reference set, as it will result in an unclear classification to several groups in the reference set. A secondary DFA need to be done until most of the tephra units of the unknown tephras could be assigned to the same reference group. Such problem would not exist in the case of calibration with segmented regression. The more chemistry information is allowed, the more precise segmented regression is formed, and results a more accurate calibration. The only problem is a prior distribution of the calibrated age has to be given in order to eliminate those impossible solutions. Nevertheless, the prior distribution could be very broad, and even a very approximate age or just a possible range for the calibrated age could be used to decide the prior distribution. Although the result showed the feasibility to apply calibration to estimate the unknown age, the accuracy of the calibrated age is still in doubt. It is highly dependent on the quality of the data and the accuracy or correctness of modeling the changing pattern with segmented regression model based on those assumptions. Therefore, more research needs to be done on such aspects to check the accuracy of the proposed segmented regression model. We hope that such model and methodology would be useful to the geologists in further research, to allow them to age the unknown tephra, to understand the changing pattern, and even to predict the occurrence of volcanic eruption better.

Acknowledgement

We thank to Dr. Sue Donoghue for supplying us with the four datasets in the analysis and also thank to Dr. John Aitchison for providing valuable advice in the beginning of the whole research.

Appendix

Syntax for the analysis by WinBUGS

```
model	{
	beta <- mul * taul
	alpha <- pow(mul, 2) * taul
	for (i in 1:Ns)		{
		r[i] ~ dgamma(alpha, beta)	}
	cumr[1] <- r[1]
	for (j in 2:Ns)		{
		cumr[j] <- cumr[j-1] + r[j]		}
	for (k in 1:Nt)		{
		for (l in 1:Ns)		{
			diff[k, l] <- cumr[l] - t[k]
			one[k, l] <- step(diff[k, l])		}
		none[k] <- sum(one[k, ])
		s1[k] <- Ns -none[k]
```

```
                    mu[k] <- a0 + a1 * t[k] + s1[k] * muc
                    tau[k] <- taue * tauc / (tauc + s1[k] * taue)
                              for (m in 1:N[k])       {
                                        can[m, k] ~ dnorm(mu[k], tau[k])             }              }
          mul ~ dgamma(1.0E-3, 1.0E-3)
          sigl <- 1 / sqrt(taul)
          taul ~ dgamma(1.0E-3, 1.0E-3)
          a0 ~ dnorm(0, 1.0E-6)
          a1 ~ dnorm(0, 1.0E-6)
          muc ~ dnorm(0, 1.0E-6)
          sigc <- 1/sqrt(tauc)
          varc <- sigc * sigc
          tauc ~ dgamma(1.0E-3, 1.0E-3)
          sige <- 1/sqrt(taue)
          vare <- sige * sige
          taue ~ dgamma(1.0E-3, 1.0E-3)
                    S[1] <- 0.01
                    S[2] <- cumr[1]
                    S[3] <- cumr[2]
                    S[4] <- 0.83
                              for (u in 1:3)          {
                                        var1[u] <- vare + (u-1) * varc
                                        tau1[u] <- 1 / var1[u]
                                        p1[u] <- (S[u+1] - S[u]) / ( S[4] - S[1])
                                        mu1[u] <- a0 + a1 * t1[u] + (u-1) * muc
                                        t1[u] ~ dunif(S[u], S[u+1])
                                        mu2[u] <- a0 + a1 * t2[u] + (u-1) * muc
                                        t2[u] ~ dunif(S[u], S[u+1])        }
                              for (o in 1:2)          {
                                        p2[o] <- (S[o+2] - S[o+1]) / (S[4]-S[2])
                                        mu3[o] <- a0 + a1 * t3[o] + o * muc
                                        t3[o] ~ dunif(S[o+1], S[o+2])
                                        mu4[o] <- a0 + a1 * t4[o] + o * muc
                                        t4[o] ~ dunif(S[o+1], S[o+2])
                                        mu5[o] <- a0 + a1 * t5[o] + o * muc
                                        t5[o] ~ dunif(S[o+1], S[o+2])    }
                                        for (p in 1:N1)         {
                                                  for (q in 1:3)          {
                                                            can1[p, q] ~ dnorm(mu1[q], tau1[q]         }              }
                                        for (a in 1:N2)         {
                                                  for (b in 1:3)          {
                                                            can2[a, b] ~ dnorm(mu2[b], tau1[b])        }              }
                                        for (c in 1:N3)         {
                                                  for (d in 1:2)          {
                                                            can3[c, d] ~ dnorm(mu3[d], tau1[d+1])    }              }
                                        for (e in 1:N4)         {
                                                  for (f in 1:2)          {
                                                            can4[e, f] ~ dnorm(mu4[f], tau1[f+1])     }              }
                                        for (g in 1:N5)         {
                                                  for (h in 1:2)          {
                                                            can5[g, h] ~ dnorm(mu5[h], tau1[h+1])    }              }

                                        sn1 ~ dcat(p1[])
                                        st1 <- t1[sn1]
                                        sn2 ~ dcat(p1[])
                                        st2 <- t2[sn2]
                                        sn3 ~ dcat(p2[])
                                        st3 <- t3[sn3]
                                        sn4 ~ dcat(p2[])
                                        st4 <- t4[sn4]
                                        sn5 ~ dcat(p2[])
                                        st5 <- t5[sn5]          }
list(mul=0.4, taul=1, a0=3, a1=-6, muc=4, tauc=1, taue=1, r=c(0.3, 0.3))
list(Ns = 2, Nt = 10, N1=14, N2 =5, N3=17, N4=25, N5=23)
```

The gamma distribution is parameterized in terms of the mean length of time for each segment $\mu_l = \alpha / \beta$ and precision of the length of time for each segment $\tau_l^2 = \beta^2 / \alpha$ to ensure convergences of those parameters.

References

Alfassi, Z. B. Roger, Z. & Ronen, Y. (2005). *Statistical treatment of analytical data.* Boca Raton: CRC Press.

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika 70,* 57-65.

Altchison, J. (1986). *The statistical analysis of compositional data.* London: Chapman & Hall.

Aitchison, J. (2003). Compostional data analysis: where are we and where should we be heading? In *proceedings of Compositional Data Analysis Workshop, 15-17 October 2003.*

Aitchison, J. & Kay, J. W. (2003). Possible soluctions of some esstial zero problems in compostional data analysis In *proceedings of Compositional Data Analysis Workshop, 15-17 October 2003.*

Bacon-Shone, J. (2003). Modelling structural zeros in compostional data In *proceedings of Compositional Data Analysis Workshop, 15-17 October 2003.*

Borchardt, G. A. Aruscavage, P. J. & Millard, H. T. (1972). Correlation of the Bishop Ash, a Pleistocene marker bed, using instrumental neutron activation analysis. *Journal of Sedimentary Petrology 42(2),* 301-306.

Charman, D. J. & Gratten, J. (1999). An assessment of discriminant function analysis in the identification and correlation of distal Icelandic tephras in the British Isles. In C. R. Firth and W. J. McGuire (Eds.), *Volcanoes in the Quaternary*, pp. 147–160. London: Geological Society.

Chen, J. & Gupta A. K. (2000). *Parametric statistical change point analysis*. Boston: Birkhäuser.

Cronin S. J. Allace, R. C. & Neall, V. E. (1996) Sourcing and identifying andesitic tephras using major oxide titanomagnetite and hornblende chemistry, Egmont volcano and Torgariro Volcanic Centre, New Zealand. *Bulletin of Volcanonology 58,* 33-40.

Forggatt, P. C. (1992). Standardization of the chemical analysis of tephra deposits. Report of the ICCT working Group. *Quaternary International 13/14*, 93–96.

Fry, J. M. Fry, T. R. L. & MaLaren, K. R. (2000). Compositional data analysis and zeros in micro data. *Applied Economics 2,* 953-959.

Hunt, J. & Hill, P. (1993). Tephra geochemistry: a discussion of some persistent analytical problems. *The Holocene 3,* 271-278.

Khattree, R. & Nair, D. N. (2000). *Multivariate data reduction and discrimination with SAS software*. Cary, NC: SAS Institute.

Mahmoud M. A. (2004). *The Monitoring of linear profiles and the inertial properties of control charts.* Doctoral dissertation. The Faculty of the Virginia Polytechnic Institute and State University, Virginia.

Martin-Fernández, J. A. Palarea-Albaladejo, J. & Gómez-García, J. (2003). Markov chain monte carlo method applied to rounding zeros of compositional data: first approach In *proceedings of Compositional Data Analysis Workshop, 15-17 October 2003.*

Lee, P. M. (2004). *Bayesian statistics: an introduction.* London: Arnold.

Stokes, S. & Lowe, D. J. (1988). Discriminant function analysis of the late Quaternary tephras from the five volcanoes in New Zealand using glass shard major element chemistry. *Quaternary Research 30*, 270–283.

Timm, N. H. (2002). *Applied multivariate analysis*. New York: Springer.