An alternative method for dating unknown tephras based on a segmented regression model

B. W. Lee[1] and J. Bacon-Shone[2]
[1]Quality Evaluation Centre, The City University of Hong Kong, HK; *bikwalee@yahoo.com.hk*
[2]Social Sciences Research Centre, The University of Hong Kong, Hong Kong

Abstract

In CoDaWork'05, we presented an application of discriminant function analysis (DFA) to 4 different compositional datasets and modelled the first canonical variable using a segmented regression model solely based on an observation about the scatter plots. In this paper, multiple linear regressions are applied to different datasets to confirm the validity of our proposed model. In addition to dating the unknown tephras by calibration as discussed previously, another method of mapping the unknown tephras into samples of the reference set or missing samples in between consecutive reference samples is proposed. The application of these methodologies is demonstrated with both simulated and real datasets. This new proposed methodology provides an alternative, more acceptable approach for geologists as their focus is on mapping the unknown tephra with relevant eruptive events rather than estimating the age of unknown tephra.

Kew words: Tephrochronology; Segmented regression.

1 Introduction

In the past, tephrochronology, 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, we presented an application of discriminant function analysis (DFA) to 4 compositional datasets in the CoDaWork'05 (Lee & Bacon-Shone, 2005). The DFA performed quite well in all these datasets, in that the first two canonical variables could explain up to nearly 80% of the variation in the compositional pattern. Moreover, it seemed to have moderate changing patterns on the first canonical variable for all the 4 cases. It appeared that the first canonical variable decreased linearly with time, jumped abruptly at some time points, and then decreased again. As a result, we proposed a segmented regression to model for this changing pattern. In fact, Westgate and Evans (1978) had mentioned that chemical data showed a systematic and unidirectional trend in tephra composition with time, in which earlier eruptions produced slightly more acidic tephra, but there was no follow-up for such a composition-age relationship. For this particular reason, it was very meaningful for us to try to study the relationship between first canonical variable and time and find a suitable model.

However, even though the estimated segmented regression line based on Bayesian approach seemed to fit the dataset quite well, two critical problems existed in the modeling procedure. Firstly, the proposal of segmented regression model was solely based on observation without any model testing procedure, so the validity of the segmented regression might be in doubt. Secondly, consistency in changing pattern was only suggested for the first canonical variable, but not for the second canonical variable. Due to the unit-sum constraint in compositional data, change in one component affects all the other components. Thus, it seems inconsistent to allow different change points or change patterns for different canonical variables. The real issue should be rephrased as whether the jump or drop at the change point for the latter canonical variable, which explains very little variation, is significant enough to be detected.

To address these two flaws, the 4 datasets are re-studied in this paper. Multiple linear regressions are applied on canonical variables from the 4 datasets. The results are shown in the coming section. It shows how the segmented regression could apply not only the first canonical variable, but also for the other canonical variables. This further affirms our proposed segmented regression model for estimating the age of unknown tephras in the previous paper. With the estimated age of the unknown tephras, it is able to map the unknown tephra with the relevant eruptive event. This induced an alternative method of dating the unknown tephra by mapping them directly into samples of the reference set or missing samples in between consecutive reference samples. The application of these methodologies is demonstrated with simulated and real datasets in the third section. This new proposed methodology is more acceptable to geologists as their focus is on mapping the unknown tephra with relevant eruptive events rather than gauging the age of unknown tephra. Finally, a conclusion for the whole paper is given in the fourth section.

2 Validity of segmented regression model

As mentioned in the previous paper, all four datasets seemed to have moderate changing patterns on the first canonical variable across age. The overall pattern could be divided into several sections of parallel straight-lines. The four datasets are re-studied here using more than two canonical variables. Dummy variable determining the number of segment has been included based on the age of tephras in the

reference set. Multiple regressions are performed to identify the relationship between the first few canonical variables with the two factors of the age and the new-formed segment variable, $can_{ij} = a_{0i} + a_{1i} age_j + a_{2i} segment_j + \varepsilon_{ij}$, where i and j identify the canonical variable and the observation respectively with the assumption that $\varepsilon_{ij} \sim N(0, \sigma_i^2)$. Each time, only one canonical variable was substituted as the dependent variable, starting from the first variable that explained the greatest variation and then moved to the second, third canonical variable etc, until both factors were not significant in the analysis.

2.1 Black ashes from New Zealand (Dataset 1)

The patterns for the four canonical variables are shown in Figure 1. At first sight, it seems as if the first canonical variable decreases from age = 0 ka to age = 0.83 ka, and then increases again to the same level as the beginning at age = 1.8 ka, but the problem is that no data is actually available between age = 0.83 ka and age = 1.8 ka. If we just focus on the pattern up to age = 0.83 ka, the first canonical variable is of slightly different level at age = 0.01 ka and age = 0.4 ka, and then keeps decreasing from age = 0.4 ka to age = 0.83 ka, with a small jump at age around 0.6 ka, but the jump at this point is less obvious than other suspected jumps. Similarly, an increasing pattern can be observed in the same interval between age = 0.4 ka and age = 0.83 ka with a slight deviation at age around 0.6 ka for the second and third variables. Therefore, we tried to divide the whole age range into four segments with the cut-points at 0.3 ka, 0.9 ka and 1.5 ka. The dummy variable set for each tephra and the result for the four multiple regression analyses can be found in Table 1 and Table 2 respectively. Both factors are significant for modeling in the first two variables, but the segment becomes only marginally significant for modeling the third variable with p-value of 0.033 and both factors are not significant for modeling the fourth variable. Figure 1 also shows that the fourth canonical variable is quite stable around the zero line. The fitted line for the first three variables is also shown in the figure. The lines fit particularly well for the period between age = 0.4 ka and age = 0.83 ka, but quite poorly at age = 0.01 ka and age = 1.8 ka. This is because most of the data are condensed in this range and there are two large gaps presenting the data between age = 0.01 ka and age = 0.4 ka as well as between age = 0.83 ka and age = 1.8 ka.

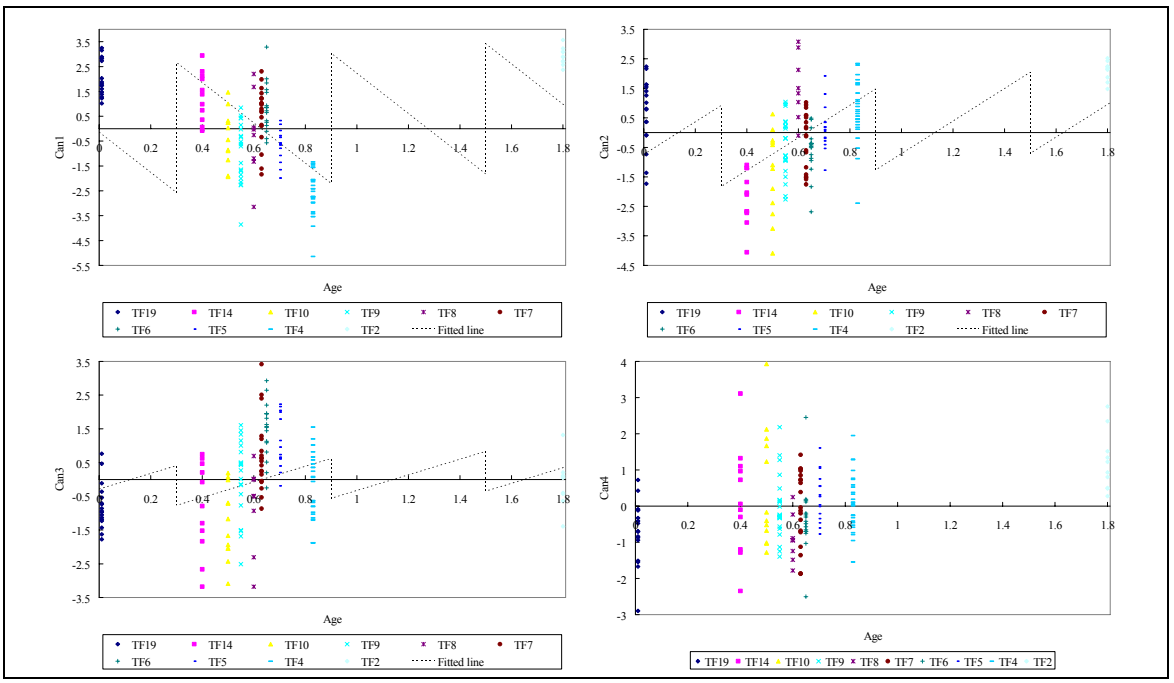

Figure 1: Scatter plots of the first four canonical variables against age (ka) for the New Zealand black ashes, where the fitted lines are estimated using the multiple regressions with significant results for both factors of the age and the newly formed segment variable.

Table 1. Dummy variable of segment set for New Zealand black ashes.

Independent Variables		Existing
Original	Dummy	tephra samples
Age (ka)	Segment	in reference set
<0.3	0	TF19
0.3 - 0.9	1	TF14 TF10 TF9 TF8 TF7 TF6 TF5 TF4
0.9 – 1.5	2	-
≥1.5	3	TF2

Table 2. Result of multiple regressions for New Zealand black ashes

Can1	Estimated Coefficients	Standard Error	P -value for equality to 0	Can2	Estimated Coefficients	Standard Error	P -value for equality to 0
Intercept	-0.14	0.26	0.591	Intercept	-0.74	0.20	<0.001
Age (ka)	-8.10	1.13	<0.001	Age (ka)	5.52	0.88	<0.001
Segment	5.24	0.71	<0.001	Segment	-2.75	0.55	<0.001
Can3	Estimated Coefficients	Standard Error	P -value for equality to 0	Can4	Estimated Coefficients	Standard Error	P -value for equality to 0
Intercept	-0.28	0.20	0.156	Intercept	-0.65	0.16	<0.001
Age (ka)	2.32	0.87	0.009	Age (ka)	-0.13	0.71	0.850
Segment	-1.18	0.55	0.033	Segment	0.72	0.44	0.108

2.2 Pumice layers from New Zealand (Dataset 2)

As shown in Figure 2, for the first canonical variable, three obvious change-points could be noted, the first one is between the age = 10.8 ka and age = 12 ka, the second one is between age = 13 ka and age = 14 ka, and the last one is between age = 17.9 ka and age = 19 ka. Although there is no data between age = 14 ka and age = 16 ka, it seems as if there might be one more jump point between this interval. A similar pattern can be noticed for the second and third variable. However, for the fourth variable, the patterns are less clear. The fourth variable drops slowly respectively from age = 10 ka to age = 12 ka, becomes stabilized around the value of zero between age = 12 ka and age = 16 ka, and rises slowly after age = 16 ka. Hence, we tried to divide the whole age range into six segments with change point at 11 ka, 13 ka, 15 ka, 17 ka and 19 ka. The dummy variable of segment set for different sets of variable could be found in Table 3. The fitted line could be seen in Figure 2 and the result for the analysis could be found in Table 4. The two factors are very significant for the first three canonical variables and are not significant for both the fourth variables. This agrees with the observation from the scatter plots.

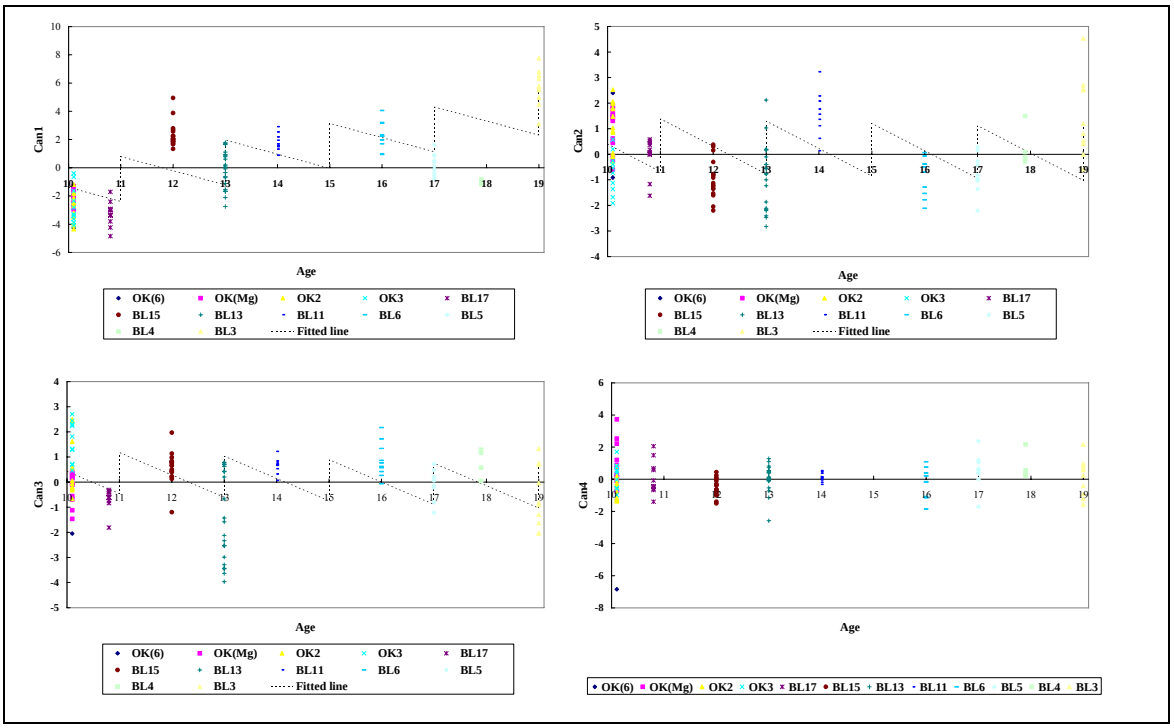

Figure 2: Scatter plots of the first four canonical variables against age (ka) for the New Zealand pumice layers, where the fitted lines are estimated using the multiple regressions with significant results for both factors of the age and the newly formed segment variable.

Table 3. Dummy variable of segment set for New Zealand pumice layers.

Independent Variables		Existing
Original	Dummy	tephra samples
Age (ka)	Segment	in reference set
<11	0	OK(6) OK(Mg) OK2 OK3 BL17
11 - 13	1	BL15 BL13
13 - 15	2	BL11
15 - 17	3	BL6 BL5
17 - 19	4	BL4
≥19	5	BL3

Table 4. Result of multiple regressions for New Zealand pumice layers

Can1	Estimated Coefficients	Standard Error	P -value for equality to 0	Can2	Estimated Coefficients	Standard Error	P -value for equality to 0
Intercept	8.62	3.26	0.009	Intercept	11.13	2.30	<0.001
Age (ka)	-1.00	0.31	0.001	Age (ka)	-1.07	0.22	<0.001
segment	3.18	0.58	<0.001	segment	2.06	0.41	<0.001
Can3	**Estimated Coefficients**	**Standard Error**	**P -value for equality to 0**	**Can4**	**Estimated Coefficients**	**Standard Error**	**P -value for equality to 0**
Intercept	9.34	2.25	<0.001	Intercept	-3.12	2.07	0.135
Age (ka)	-0.89	0.21	<0.001	Age (ka)	0.29	0.20	0.149
segment	1.64	0.40	<0.001	segment	-0.45	0.37	0.230

2.3 Black ashes from U. S. A. (Dataset 3)

The patterns for these canonical variables (Fig. 3) are not as clear as those for New Zealand datasets. It may be because the available tephra units are unevenly distributed over time. Little geochemical information is available after age = 3 ka. For the first canonical variable, the changing pattern could be seen from age = 0.4 ka to age = 3.4 ka, there should be a change point between age = 1.1 ka to age = 2.3 ka. After age = 3.4 ka, the value for the first canonical variable seems to remain at the same level, while for the other variables, the values seem to stay at similar level for all time points, so regression analysis has been done to see whether there is really any change across time. We divided the whole time range into five segments with change points at 1.5 ka, 3 ka, 4.5 ka, 6 ka and 7.5 ka (Table 5). The two factors are extremely significant for modeling the first and second variables, quite significant for the third variable and also slightly significant for the fourth variable, but both are insignificant for the fifth variable (Table 6). Therefore, regression could depict changes in the second, third and fourth variables that are impossible to indicate by observation.

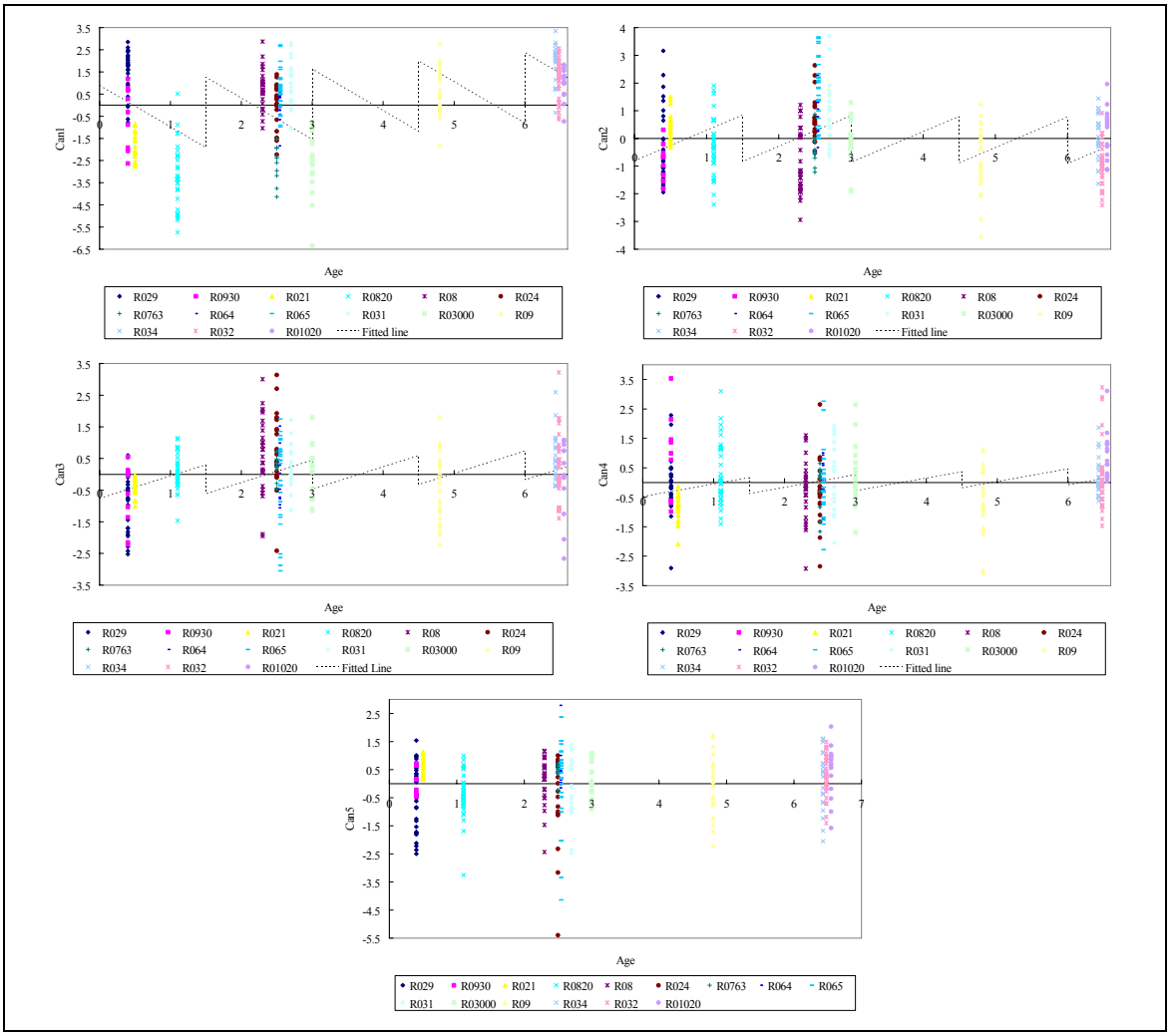

Figure 3: Scatter plots of the first five canonical variables against age (ka) for the U. S. A. black ashes, where the fitted lines are estimated using the multiple regressions with significant results for both factors of the age and the newly formed segment variable.

Table 5. Dummy variable of segment set for U. S. A. black ashes.

Independent Variables		Existing
Original	Dummy	tephra samples
Age (ka)	Segment	in reference set
<1.5	0	R029 R0930 R021 R0820
1.5 – 3	1	R08 R024 R0763 R064 R065 R031 R03000
3 - 4.5	2	-
4.5 – 6	3	R029
6 - 7.5	4	R034 R032
≥7.5	5	R02000

Table 6. Result of multiple regressions for U. S. A. black ashes

Can1	Estimated Coefficients	Standard Error	P -value for equality to 0	Can2	Estimated Coefficients	Standard Error	P -value for equality to 0
Intercept	0.89	0.29	0.002	Intercept	-0.81	0.21	<0.001
Age (ka)	-1.87	0.28	<0.000	Age (ka)	1.10	0.21	<0.001
segment	3.18	0.40	<0.001	segment	-1.68	0.30	<0.001
Can3	Estimated Coefficients	Standard Error	P -value for equality to 0	Can4	Estimated Coefficients	Standard Error	P -value for equality to 0
Intercept	-0.76	0.19	<0.001	Intercept	-0.48	0.19	0.011
Age (ka)	0.71	0.18	<0.001	Age (ka)	0.43	0.18	0.019
segment	-0.92	0.26	<0.001	segment	-0.55	0.26	0.035
Can5	Estimated Coefficients	Standard Error	P -value for equality to 0				
Intercept	-0.15	0.18	0.409				
Age (ka)	0.02	0.18	0.913				
segment	0.06	0.25	0.804				

2.4 Pumice layers from U. S. A. (Dataset 4)

As shown in Figure 4, for the first canonical variable, it seems to decrease continuously from the beginning, jump up suddenly at age = 6 ka and decrease continuously afterward; since there is no data point until age = 8.75 ka, it is inconclusive whether there is any change point within this interval. The patterns for the second and third variables are similar to the pattern of the fourth variable in dataset 2, the values drop slowly to a certain extent and stabilize at the end. For the fourth variable, the value fluctuates around zero. As the age range for this dataset overlaps that for the dataset 3 and the two datasets are from the same volcano, we tried to test the relationship with the same change points of 4.5 ka, 6 ka and a new change point 7.5 ka. We assign those samples to the suitable segment as shown in Table 7. Although segment is highly significant in modeling the first canonical variable, age is marginally insignificant to model the variable. However, both factors are highly significant for modeling the second variable, slightly significant for the third and not significant for the fourth. The result is summarized in Table 8. The lines do not fit the first two canonical variable of R02000 well, but much better for the third canonical variable. This is again due to due to great gap after age = 6.5 ka.

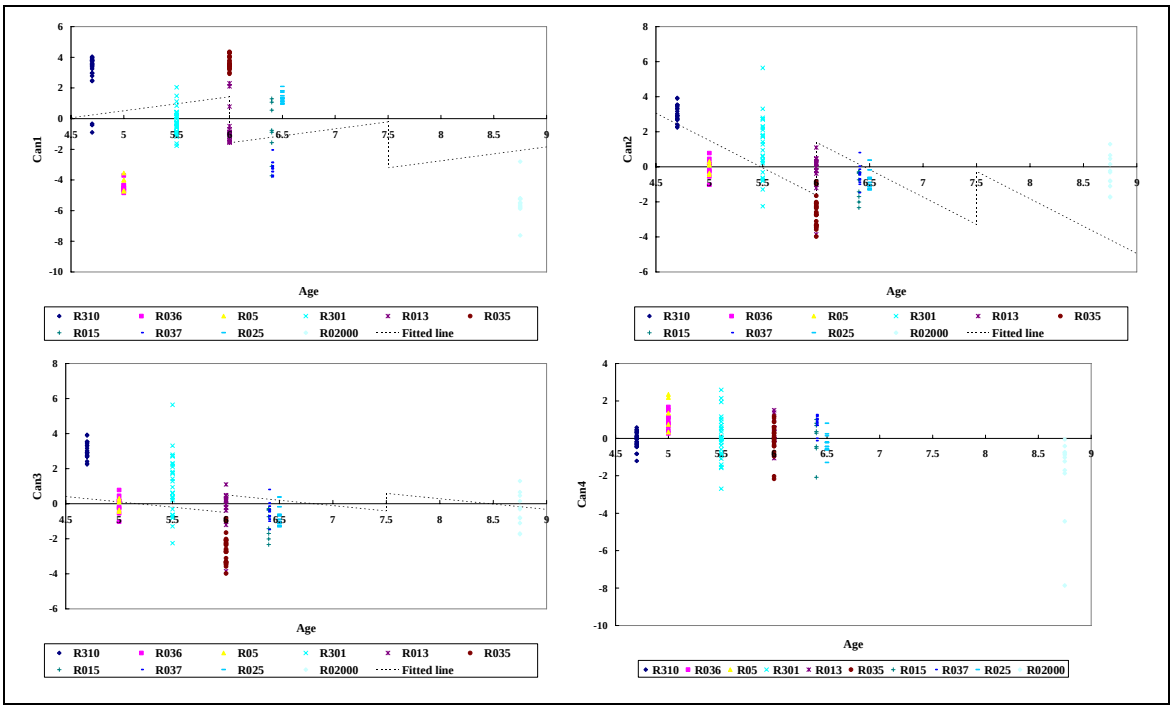

Figure 4: Scatter plots of the first four canonical variables against age for the U. S. A. pumice layers, where the fitted lines are estimated using the multiple regressions with significant results for both factors of the age and the newly formed segment variable.

Table 7. Dummy variable of segment set for U. S. A. pumice layers

Independent Variables		Existing
Original	Dummy	tephra samples
Age (ka)	Segment	in reference set
<4.5	0	R310 R036 R05 R301 R013 R035
4.5 – 6	1	R015 R037 R025
6 - 7.5	2	-
≥7.5	3	R02000

Table 8. Result of multiple regressions for U. S. A. pumice layers

Can1	Estimated Coefficients	Standard Error	*P* -value for equality to 0	Can2	Estimated Coefficients	Standard Error	*P* -value for equality to 0
Intercept	-4.04	2.70	0.137	Intercept	17.12	1.47	<0.001
Age (ka)	0.91	0.49	0.068	Age (ka)	-3.12	0.27	<0.001
Segment	-3.00	0.60	<0.001	segment	3.02	0.33	<0.001
Can3	**Estimated Coefficients**	**Standard Error**	***P* -value for equality to 0**	**Can4**	**Estimated Coefficients**	**Standard Error**	***P* -value for equality to 0**
Intercept	3.15	1.41	0.028	Intercept	1.25	1.12	0.266
Age (ka)	-0.61	0.26	0.020	Age (ka)	-0.18	0.21	0.373
segment	1.00	0.31	0.002	segment	-0.37	0.25	0.139

2.5 Combining the two datasets from the U. S. A. (Dataset 5)

There are large age gaps in the two datasets from U. S. A and thus major information gaps within each dataset, but the age ranges for the two data sets from U. S. A. overlap; the black ashes from U. S. A. are in the age range of 0.4 ka to 6.55 ka, while the pumice layers are is from age = 4.7 ka to age = 8.75 ka. Moreover, the two datasets are from the same volcano, so it is logical to combine the two datasets to see

whether this will give a clearer picture for the relationship between age and composition. The value of similarity coefficient (SC) as defined by Borchardt and others (1971), as well as the value of Mahalanobis distance squared statistics (D^2) from DISCRIM procedure of SAS for comparing the similarity between the tephra samples from the two datasets are also found. A value of SC equalling 0 means wholly dissimilar and 1 means identical. As shown in Appendix, most of the values of similarity coefficient are greater than 0.5 and only two are less than 0.5, but quite close to 0.5, so the composition of the two types of tephra sample from the same vent are in fact quite close to each other.

Combining the two datasets provides a fuller picture for investigation (Fig. 5). Based on the observations and conclusions from the previous two sections, we divided the whole age range into six segments with change points at 1.5 ka, 3 ka, 4.5 ka, 6 ka and 7.5 ka. The assignation of different tephra samples could be found in Table 9. Five canonical variables have been studied in dataset 4; we include results of regression analyses up to the fifth canonical variable in Table 10 even though both factors are insignificant from the third variable. It could pick up changes for the first two variables, the regression lines are also shown in Figure 5. The lines fitted all the reference tephras in this combined dataset better than the fitting in the two separate datasets.

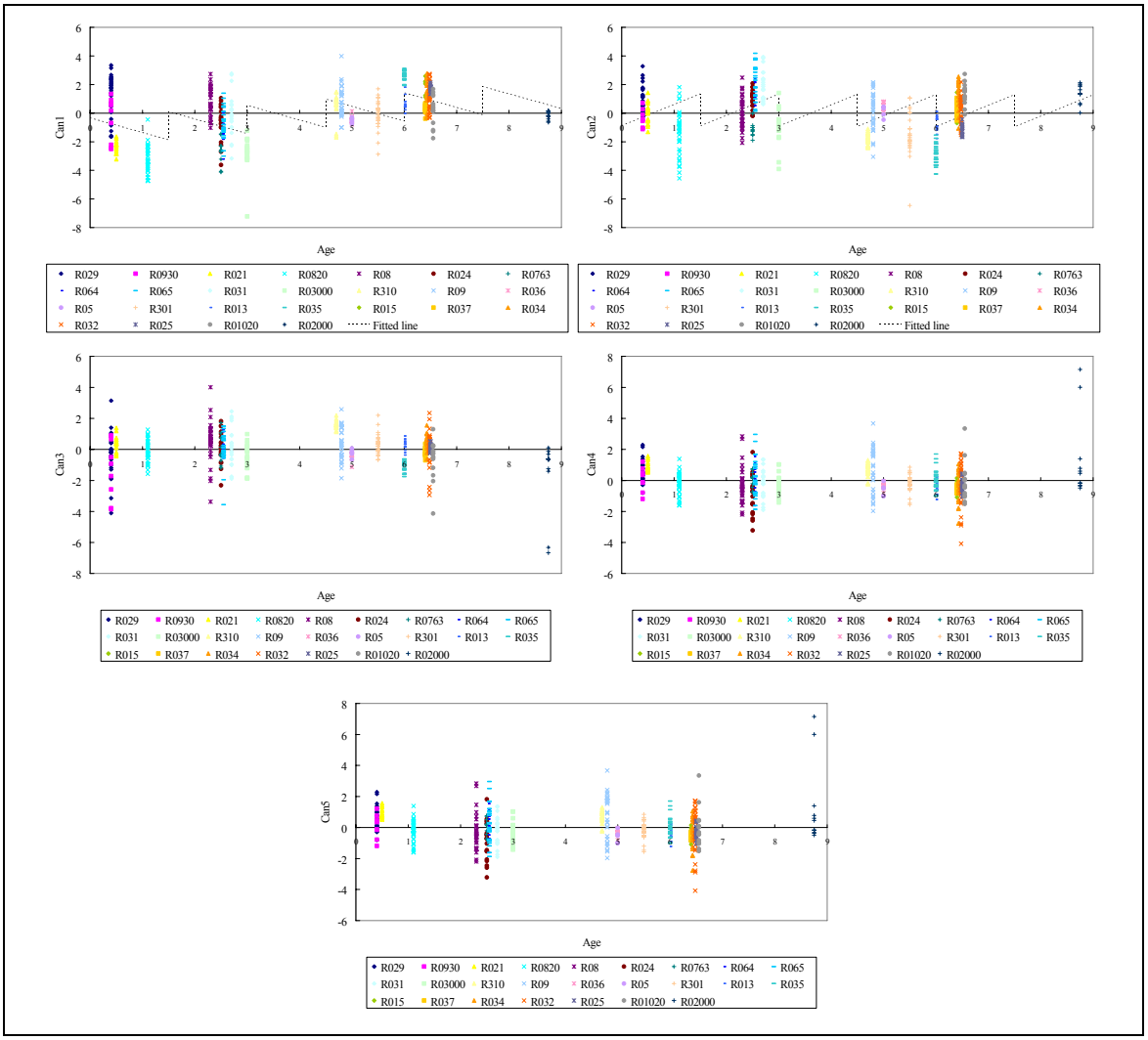

Figure 5: Scatter plots of the first five canonical variables against age for combining the two U. S. A. datasets, where the fitted lines are estimated using the multiple regressions with significant results for both factors of the age and the newly formed segment variable.

Table 9 Dummy variable of segment set for combining the two U. S. A. datasets

Independent Variables		Existing
Original	Dummy	tephra samples
Age (ka)	Segment	in reference set
<1.5	0	R029 R0930 R021 R0820
1.5 - 3	1	R08 R024 R0763 R064 R065 R031
3 - 4.5	2	R03000
4.5 – 6	3	R310 R0930 R036 R05 R301
6 - 7.5	4	R013 R035 R015 R037 R034 R032 R025 R01020
≥7.5	5	R02000

Table 10. Result of multiple regressions for combining the two U. S. A. datasets

Can1	Estimated Coefficients	Standard Error	P -value for equality to 0	Can2	Estimated Coefficients	Standard Error	P -value for equality to 0
Intercept	-0.33	0.21	0.116	Intercept	-0.88	0.21	<0.001
Age (ka)	-1.02	0.21	<0.001	Age (ka)	1.48	0.21	<0.001
Segment	1.97	0.29	<0.001	segment	-2.24	0.29	<0.001
Can3	Estimated Coefficients	Standard Error	P -value for equality to 0	Can4	Estimated Coefficients	Standard Error	P -value for equality to 0
Intercept	0.09	0.15	0.55	Intercept	0.28	0.15	0.054
Age (ka)	-0.04	0.15	0.82	Age (ka)	-0.25	0.14	0.078
Segment	0.02	0.22	0.92	segment	0.32	0.20	0.113
Can5	Estimated Coefficients	Standard Error	P -value for equality to 0				
Intercept	-0.17	0.14	0.239				
Age (ka)	0.08	0.14	0.587				
Segment	-0.06	0.20	0.768				

3 Newly proposed method for dating unknown tephra

The analyses in the previous section provide a better insight into the changing pattern of the changing pattern of the composition with the use of canonical variables. Although the assignation of change point and dummy variable of segment factor is still subject to question, this could still further confirm the feasibility of the proposed segmented regression model and understanding of the model in the previous paper. It is reasonable to have changes in distributions of all canonical variables at same time points, but the changes might not be easily distinguishable in latter variables. It appears that these canonical variables move linearly with time, shift down or up abruptly at some time points, and then move linearly with a similar slope again. As similar moving trends are observed for all the variables, it is reasonable to simplify the whole problem by studying just the first variable.

Based on this observed relationship between first canonical variable and age, we further proposed to estimate the age of unknown tephras by calibration in the Second Compositional Data Analysis Workshop CoDaWork'05. While this method was useful in studying the changing pattern and dating the unknown tephras, it was subject to two major criticisms. First, it gave an equivocal answer for the age of unknown tephra as multiple solutions were produced in the calibration, this could be dealt with by setting a suitable informative prior. In fact, the prior could be set based on the approximate ages of unknown tephras. Even though the approximate age was not precise, the ordering of the estimates might also be useful and should not be ignored. Therefore, the calibrated ages for the unknown samples should be increasing from NZ500 to NZ526. The ordinal information is included in the new method proposed in this paper. The second limitation was that the method proposed in previous paper did not really address the problem of geologists, whose primary concern is to identify which reference sample each of the unknown

tephra could belong to, rather than the age of the unknown tephra. Therefore, our newly proposed method is focused on the mapping problem. However, it should be noticed that our proposed second method is broader than DFA in that we do not just consider the tephras in reference set, but also include the possibility of missing intermediate groups between each pair of consecutive reference samples.

3.1 Methodology

The newly proposed method is still based on the framework of a specific segmented regression mentioned in previous paper; based on the simulated age of unknowns, we assign the unknowns to a suitable subgroup. To put the unknowns into subgroup, we have to inquire the range for the real age of each tephra sample in the reference set. It has to consider the greatest measured error, which is reflected by the value of the measurement error. The independent variable age should not be counted as a fixed variable, because the aging of the tephra with the use of EMA involves measurement error, which is random and is assumed to follow a uniform distribution over the range of the value of the measured age plus or minus its greatest measured error, i.e. $Age_i \sim Unif\,(RAge_i - \xi_i, RAge_i + \xi_i)$, where $RAge_i$ is the given age for the i th tephra sample in the reference set and ξ_i is the corresponding measurement error.

As the ordinal information of the given approximate age is also considered in this newly proposed measure, we first have to arrange the tephra samples in increasing order of age in the reference set and of given approximate age in the unknown set, i.e. $RAge_1 < RAge_2 < \cdots < RAge_{N_r-1} < RAge_{N_r}$, where N_r is the number of tephra samples in the reference set, so the range for the segmented regression is between the range of $RAge_1 - \xi_1$ and $RAge_{N_r} + \xi_{N_r}$. This range could be divided into $2N_r - 1$ sub-segments according to the values of the ages plus or minus the corresponding measurement error, that is the upper and lower boundaries for the uniform distribution of age. Therefore, the principle for mapping the unknown is based on the sub-segment in which the age for the unknown tephra falls into; if the age of the unknown is inside the range of the age for one of the reference sample plus or minus the measurement error, this unknown is assigned to that reference sample. Conversely, if the unknown could not be mapped to any of the reference sample, it is considered to be a new-found tephra. The allocation rule is summarized in the following bracket:

$$\left\{\begin{array}{cc}
RAge_1 - \xi_1 < Age_i < RAge_1 + \xi_1 & Sample_1 \\
RAge_1 + \xi_1 < Age_i < RAge_2 - \xi_2 & Miss_1 \\
RAge_2 - \xi_2 < Age_i < RAge_2 + \xi_2 & Sample_2 \\
RAge_2 + \xi_2 < Age_i < RAge_3 - \xi_3 & Miss_2 \\
\vdots & \Rightarrow \quad \vdots \\
RAge_{N_r-2} + \xi_{N_r-2} < Age_i < RAge_{N_r-1} - \xi_{N_r-1} & Miss_{N_r-2} \\
RAge_{N_r-1} - \xi_{N_r-1} < Age_i < RAge_{N_r-1} + \xi_{N_r-1} & Sample_{N_r-1} \\
RAge_{N_r-1} + \xi_{N_r-1} < Age_i < RAge_{N_r} - \xi_{N_r} & Miss_{N_r-1} \\
RAge_{N_r} - \xi_{N_r} < Age_i < RAge_{N_r} + \xi_{N_r} & Sample_{N_r}
\end{array}\right\}$$

For simplicity, assume these $2N_r - 1$ subgroups are non-overlapping, i.e. $\xi_{i+1} + \xi_i \le RAge_{i+1} - RAge_i$ for all i from 1 to $N_r - 1$.

The ordinal information for the approximate age is used in this method. Therefore, in each MCMC simulation for the Bayesian analysis, it maps the unknown with smallest approximate age into a subgroup (u_1); based on the new found information u_1, the unknown with the second smallest approximate age was assigned to the relevant subgroup. The process repeats until reaching the last unknown with the greatest approximate age. The mechanism for assigning the first unknown to the suitable subgroup is as follows:

$$
\begin{aligned}
u_1 &\sim Cat(p_1 \quad p_2 \quad \cdots \quad p_{2N_r-2} \quad p_{2N_r-1}) \\
&\Downarrow \\
u_2 &\sim Cat(0 \quad \cdots \quad 0 \quad p_{u_1} \quad p_{u_1+1} \quad \cdots \quad p_{2N_r-2} \quad p_{2N_r-1}) \\
&\Downarrow \\
&\vdots \\
&\Downarrow \\
u_{N_u-1} &\sim Cat(0 \quad \cdots \quad 0 \quad p_{u_{N_u-2}} \quad p_{u_{N_u-2}+1} \quad \cdots \quad p_{2N_r-2} \quad p_{2N_r-1}) \\
&\Downarrow \\
u_{N_u} &\sim Cat(0 \quad \cdots \quad 0 \quad p_{u_{N_u-1}} \quad p_{u_{N_u-1}+1} \quad \cdots \quad p_{2N_r-2} \quad p_{2N_r-1})
\end{aligned}
,
$$

whereas N_r is the number of unknown tephras; p_i corresponds to the percentage of assigning each unknown to i th sub-group; $\mathbf{p}$ is given with a prior of Dirichlet distribution, $[p_1 \quad p_2 \quad \cdots \quad p_{2N_r-2} \quad p_{2N_r-1}] \sim Dirch(1-p_{miss} \quad p_{miss} \quad \cdots \quad p_{miss} \quad 1-p_{miss})$, where p_{miss} stands for the probability that the unknown tephra is a new found one and thus does not belong to any tephra in the reference set; Thus, it is assumed there is p_{miss} probability for each unknown tephra to be a new found one and $1-p_{miss}$ probability that the unknown belongs to the reference set. For simplicity, p_{miss} is given with a prior of uniform distribution in between 0 and 1 instead of the $Beta$ distribution, which can be used to study the posterior for the probability to find a missing tephra in the reference set.

Therefore, the $2N_r-1$ subgroups can be divided into 2 subsets; a set of N_r subgroups, which belong to the reference sets and another set of N_r-1 subgroups, which cover ages missed between the reference samples; within each set, equal probability is assigned to each of the subset by giving parameters in the Dirichlet distribution as it is assumed there is not much information about the real age. Besides, as we have no information about how possible it is to find a new found tephra to fill in the gap in reference set and how many missing values could exist in the unknown set, a non informative prior has been set to estimate the value of p_{miss} . u_i is restricted not to be less than u_{i-1} . Such censoring utilizes the order information based on the approximate age and may eliminate those redundant solutions. The application of these methodologies is demonstrated with a simulated dataset and a real dataset as follow.

3.1.1 Simulated dataset

30 datasets with combinations of different common levels of measurement error (ζ = 0.1, 0.2, 0.3, 0.4, 0.5) and different missing levels (No. of missing = 0, 1, 2, 3, 4, 5) have been simulated based on the segmented regression model mentioned in the previous paper with $A_{01}=16$, $A_1=-2$, $\mu_c=6$, $\sigma_c^2=0.001$, $\sigma_r^2=0.1$, $\alpha=5$, $\beta=1$. In all these datasets, each of the unknowns could be assigned to the correct subgroup with a single solution. The misclassification probability has been calculated for each of the simulated dataset and is reported in Table 11. All values are less than 0.5, which is much less than the naïve estimate = 18/19. It is expected that the misclassification probability will increase with the common error level and with the number of missing groups.

Table 11. Misclassification probabilities for the 30 simulated datasets

No. of Missing groups	Common measurement error for age variable				
	0.1	0.2	0.3	0.4	0.5
0	0.015	0.047	0.180	0.194	0.190
1	0.026	0.085	0.295	0.204	0.214
2	0.052	0.189	0.247	0.288	0.310
3	0.069	0.155	0.220	0.311	0.259
4	0.075	0.063	0.125	0.351	0.296
5	0.085	0.071	0.201	0.260	0.166

3.1.2 Real dataset

As in the previous paper, we choose to demonstrate the methodology using the New Zealand black ashes. We have tried to fit the reference set with number of change points = 2, 3, 4, 5. However, similar to the result in previous paper, all these four cases do not seem to fit very well, but could roughly follow the trend. The four fitted lines are shown in Figure 6. It may be because of the two great gaps from age = 0.01 ka to age = 0.4 ka and from age = 0.83 ka to age = 1.8 ka. To determine which fits best, each of the reference samples has been re-substituted into the fitted segmented regression, mapped based on the model and the misclassification rate is calculated. The average misclassification rate for each of the fitted segmented regression models and MSE are shown in Table 12 and this gives insight into whether the segmented regression fits the dataset well and can be used to assess whether the segmented regression is useful for assigning the unknown tephras into a suitable sub-group.

Both statistics for checking for the accuracy of the model are the smallest with five change points. Therefore, our mapping was done based on the segmented regression model with 5 change points. The kernel density for the mapped subgroup for each of the unknown tephra in the real dataset is shown in Figure 5.8. It shows that NZ359, NZ361, NZ374 and NZ526 should be assigned to 18^{th} subgroup, which is the missing group between age = 0.83 ka and 1.8 ka. The solution for NZ500 is not very clear, P(mg = 14) is slightly greater than P(mg = 2); the 14^{th} subgroup is the missing group between age = 0.6 ka and age = 0.7 ka and 2^{nd} subgroup is the missing group between age = 0.01 ka and 0.4 ka. With the approximate age of 0.5 ka of NZ500, we tend to accept that its age is between 0.6 ka and 0.7 ka. For NZ08, the kernel density shows that it is from the 16^{th} subgroup that is the missing group aged between 0.7 ka and 0.8 ka.

Table 12. Detected change points, M. S. E. and the average misclassification rate of the estimated segmented regressions

No. of change points	Detected change points (ka)					M.S.E.	Average misclassification rate
2	0.9643	1.764				2.7921	0.7702
3	0.6108	1.323	1.737			4.3798	0.7461
4	0.3863	0.8838	1.296	1.71		2.9711	0.7468
5	0.323	0.955	1.193	1.436	1.677	2.6373	0.7303

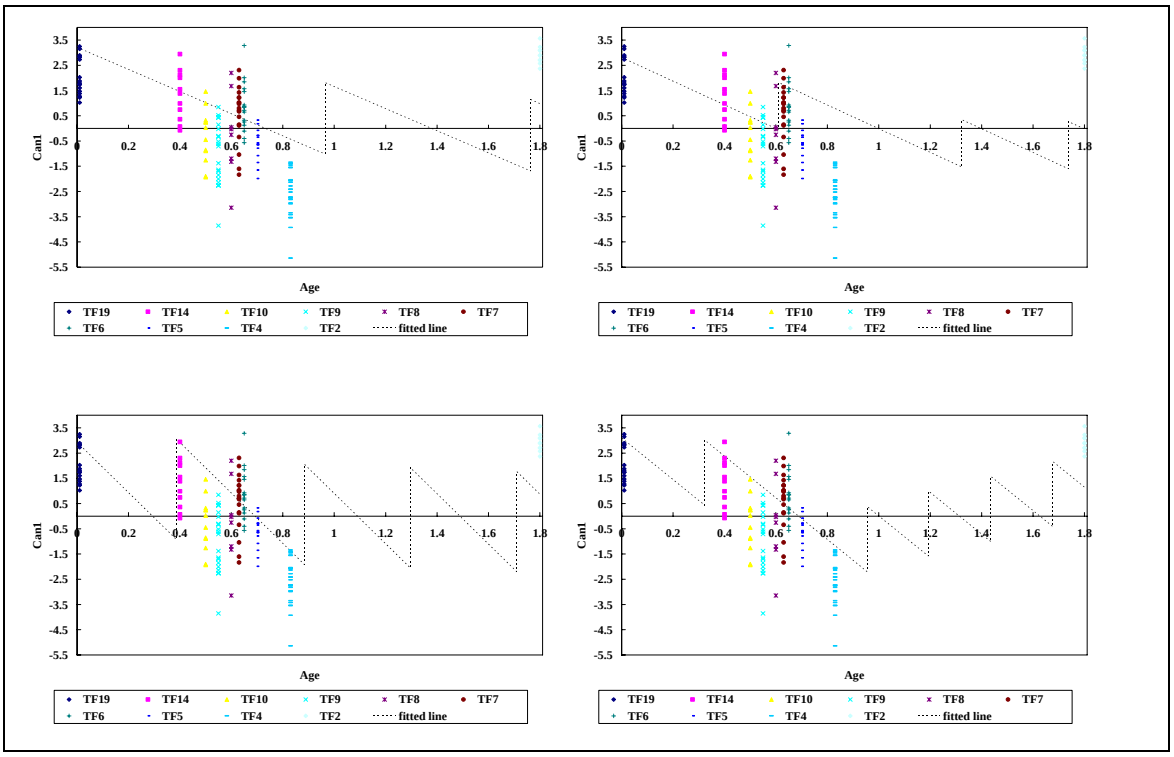

Figure 6: Scatter plot of the first canonical variables for New Zealand black ashes with the segmented regression line (No. of change points = 2, 3, 4, 5)

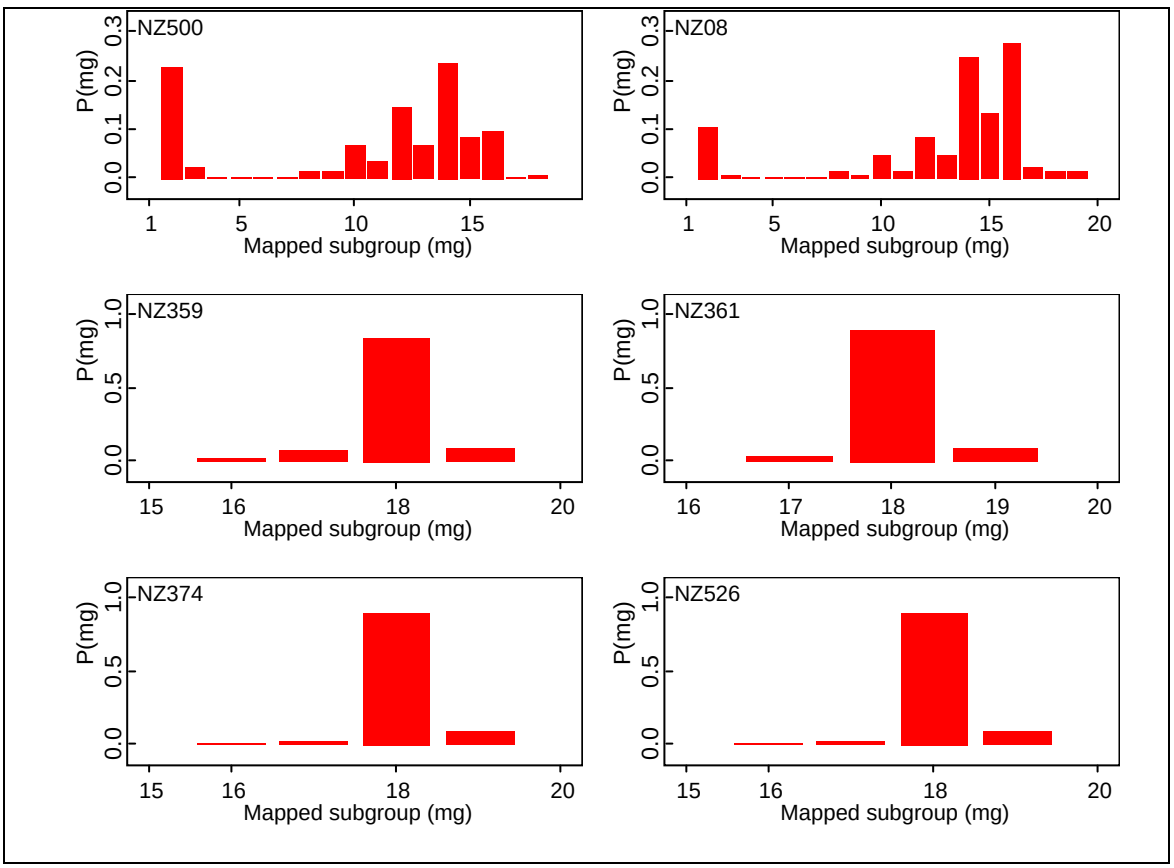

Figure 7: Kernel densities for the mapped subgroup for the unknown tephras in the New Zealand black ashes dataset (with five change points)

4 Conclusion

Obviously, this newly proposed measure makes better use of the information from the unknown tephras, so that it can detect some of the change points that are not detected in the calibration method proposed in previously, such as the change point between age = 0.01 ka and age =0.4 ka. The analysis in Section 2.1 suggests that there is a change point in this range; however, this change point cannot be detected in the analysis in previous paper even though it can be detected with this newly proposed measure in the trials with four and five change points. Besides, undoubtedly, the use of ordinal information eliminates unrealistic solutions.

The performance of applying the proposed method to the simulated datasets is quite good, suggesting the feasibility of our proposed methodology. However, the performance of the method in real dataset of New Zealand black ashes is not very good. The line does not fit the dataset very well and a double solution is arrived for NZ500. The critical problem is the sparseness of the reference dataset. By sparseness, we do not mean the number of tephra samples in the dataset, the number of individual shards in each sample or the problem of outliers; the previous two problems are typical in geological analyses, but are not very serious in our chosen dataset, in which all, except for TF8, have more than 10 individual shards. The real concern here is the uneven distribution of the tephra samples across age; 8 out of 10 samples are condensed between age from 0.4 ka to 0.83 ka, while there are two great gaps from age = 0.01 ka to age = 0.4 ka and from age = 0.83 ka to age = 1.8 ka. Therefore, the goodness of fit of the proposed segmented regression is really in doubt. The poor performance may also be due to too much simplification in the proposed model, use of non-informative prior for some of the estimators such as p_{miss} or not utilizing the information of estimated age of unknown tephra in the mapping procedure. After all, the first canonical variable can just explain part of the variance in the major-oxide composition and there may exist some variation in the first canonical variable that could not be explained by our proposed model. Regardless of these poor results, we believe that it is a very good starting point to try to model the changing pattern with our proposed segmented regression. With more input from the geologists, we hope to build a better model by adjusting some of the assumptions and adding more information into the Bayesian hierarchy.

Acknowledgement

We thank to Dr. Sue Donoghue for bringing in this interesting problem and supplying us with the four datasets in the analysis. We should also thank Professor John Aitchison for providing valuable advice in the whole research.

Appendix

		R029	R0930	R021	R0820	R08	R024	R0763	R064	R065	R031	R03000	R09	R034	R032	R01020
R310	*SC*	0.76	0.62	0.60	0.54	0.74	0.61	0.62	0.60	0.58	0.60	0.54	0.77	0.69	0.72	0.61
	D^2	27.56	220786.00	185.55	78.51	11.25	114.03	198.68	64.78	31.75	22.90	108.35	4.94	17.18	11.39	135.76
R036	*SC*	0.71	0.93	0.86	0.79	0.67	0.87	0.78	0.90	0.85	0.91	0.81	0.73	0.78	0.77	0.91
	D^2	15.61	153813.00	136.15	11.63	4.89	2.95	318.07	181.84	3.72	32.94	9.88	4.25	18.56	7.21	57.96
R05	*SC*	0.72	0.95	0.87	0.80	0.67	0.85	0.79	0.90	0.84	0.93	0.79	0.74	0.78	0.78	0.91
	D^2	21.50	12317.00	139.13	12.17	3.36	4.75	259.50	308.49	5.03	43.67	10.59	2.95	16.95	8.39	107.69
R301	*SC*	0.89	0.75	0.72	0.63	0.89	0.73	0.73	0.72	0.69	0.73	0.63	0.84	0.86	0.90	0.73
	D^2	45.66	1206045.00	139.78	31.64	14.77	58.59	256.04	168.24	15.50	34.42	40.43	5.14	19.50	7.47	198.76
R013	*SC*	0.80	0.71	0.69	0.60	0.90	0.71	0.60	0.71	0.68	0.71	0.60	0.90	0.87	0.87	0.72
	D^2	19.86	1142819.00	288.11	43.60	3.10	41.00	210.97	442.07	11.32	20.11	78.71	3.46	13.29	5.52	91.40
R035	*SC*	0.71	0.54	0.53	0.47	0.76	0.55	0.56	0.56	0.53	0.55	0.47	0.69	0.64	0.65	0.56
	D^2	44.55	270213.00	677.38	144.11	24.96	132.91	1781.00	1902.00	29.00	63.93	517.50	9.57	60.87	29.37	183.40
R015	*SC*	0.85	0.76	0.74	0.65	0.85	0.77	0.65	0.77	0.75	0.78	0.65	0.94	0.88	0.93	0.71
	D^2	14.24	2811496.00	357.85	39.47	3.94	21.79	317.98	177.34	8.59	5.40	68.12	3.51	6.38	3.19	15.55
R037	*SC*	0.75	0.83	0.78	0.69	0.80	0.81	0.68	0.81	0.76	0.83	0.68	0.76	0.82	0.81	0.83
	D^2	13.65	558181.00	199.21	20.54	0.99	5.21	251.12	309.75	4.29	17.51	21.04	1.93	12.30	4.16	49.53
R025	*SC*	0.72	0.65	0.63	0.56	0.83	0.66	0.55	0.66	0.63	0.66	0.56	0.85	0.78	0.81	0.66
	D^2	20.64	1672759.00	402.44	66.13	5.22	67.58	266.81	423.06	15.53	15.31	140.78	4.16	12.24	5.40	63.61
R02000	*SC*	0.73	0.64	0.62	0.55	0.82	0.65	0.54	0.65	0.63	0.65	0.55	0.85	0.78	0.80	0.66
	D^2	12.04	243542.00	139.65	29.26	8.18	20.17	506.40	871.44	4.82	58.95	86.50	13.38	115.59	16.58	17.35

References

Borchardt, G.A. Harward, M.E. & Schmitt, R.A. (1971). Correlation of volcanic ash deposits by activation analysis of glass separates. *Quaternary Research* 1(2): 247-60.

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.

Lee, B. W. & Bacon-Shone, J. (2005). Application of discriminant function analysis and change-point problem in dating volcanic ashes In *proceedings of Compositional Data Analysis Workshop, 19-21 October 2005.*

Westgate, J. A. & Evans, M. E. (1978). Compositional variability of Glacier Peak tephra and its stratigraphic significance. *Canadian Journal of Earth Sciences* 15: 1554-67.