STATISTICAL REPORT:

Soil Elemental Composition Statistical Similarity Assessment

SUBMITTED TO:

Soil Forensics Group, James Hutton Institute

AUTHORED BY:

Dr Nicholas Schurch

Principal Statistician for Environmental Science & Ecology

1 Declaration summary

This digital notebook generates a report consisting of 31 (thirty one) pages, each dated, and is digitally signed. I confirm that the contents of this report are true to the best of my knowledge and belief.

Should my opinion change in relation to any material issue, I will inform the authorities as soon as reasonably practical and give reasons.

[redacted]

2 Report

2.1 Introduction

This report describes the statistical analysis of soil samples from a forensic inquiry undertaken to quantify the evidential strength for questioned soil samples being consistent with a common origin with any of a set of reference soil samples. The raw data were provided by Prof Lorna Dawson. No discussion took place with Prof Dawson regarding her evaluation until after this statistical evaluation was carried out. The statistical analysis involves an initial exploration of the data to assess the characteristics of the individual dataset, the identification of potential data problems, data transformation and imputation, hierarchical sample clustering using an appropriate distance metric, Principal Component Analysis and Fasano & Franceschini 2D Kolmogorov-Smirnov testing for the questioned samples. The latter is a threshold-based null hypothesis test that categorizes whether the questioned soils are consistent with, or inconsistent with, a common origin with the references soil sample. For maximum transparency, the analysis is conducted with the open source, freely available, statistical software package R, using the following open source, freely available statistical and visualization libraries:

```
library(tidyverse)
library(readxl)
library(pvclust)
library(pcaMethods)
library(RColorBrewer)
library(latex2exp)
library(knitr)
library(kableExtra)
library(zCompositions)
library(ggiraphExtra)
library(tidyr)
library(fasano.franceschini.test)
```

2.2 Data

The data for this report are a questioned soil sample and soil samples taken from each foot of two pairs of boots:

1. JM019/Area 1: Soil from silver strip From vest belonging to PC Short (labcode: 1367045)
2. JM019/Area 2: Soil from silver strip near edge of yellow fabric From vest belonging to PC Short (labcode: 1367046)
3. JM019/Area 3: Soil from yellow fabric From vest belonging to PC Short (labcode: 1367047)
4. GAY016/Area 1: Soil to toe at welt From right boot belonging to Mr Bayoh (labcode: 1367041)
5. GAY016/Area 2: Soil at heal of sole From right boot belonging to Mr Bayoh (labcode: 1367042)
6. GAY017/Area 1: Welt inner aspect mid-section From left boot belonging to Mr Bayoh (labcode: 1367043)
7. GAY017/Area 2: Sole towards inner aspect of toe area From left boot belonging to Mr Bayoh (labcode: 1367044)
8. AM001/Area 3: Sole at toe area From left boot belonging to PC Walker (labcode: 1367039)
9. AM002/Area 1: Soil at heel From right boot belonging to PC Walker (labcode: 1367040)

Soil samples were subjected to Energy Dispersive X-ray Analysis (EDXA), with six replicate measurements per sample. The instrument used was a Zeiss EVO LS10 Scanning electron Microscope and the Inca System Energy Dispersive X-ray Analyser, from Oxford Instruments. EDXA produces spectra, with peaks corresponding to each of the elements in the sample, and peak heights that correspond to the abundance of the element. The data provided here has measurements for 11 elements; Sodium, Magnesium, Aluminium, Silicon, Sulphur, Potassium, Calcium, Titanium, Copper, Phosphorus & Magnesium.

The data are relative composition measurements that sum to 100 (i.e. percentages). The data are quoted to 2 decimal places, however a conversation with the technical helpdesk of Oxford Instruments[1], who made the system, places the instrument accuracy at approximately +/-1%). Relative composition data have some important statistical characteristics, including:

1. they are non-normally distributed and are bounded (in this case by the values 0 & 100), and;
2. the values are strongly correlated and thus the strength of one category directly impacts the values of the other categories. For example, when comparing two samples which have the same absolute elemental composition across elements except for Sodium levels, the relative composition measurements will have different values across **all** the elements, even those whose absolute abundance has not changed.

These characteristics make the use of most standard statistical approaches unsuitable for use directly with these data, including hierarchical clustering and dimensionality reduction techniques (such as Principal component analysis or Non-Metric Multidimensional Scaling). A common approach for mitigating these limitations is to define one of the categories as a reference, and transforming the rest of the data into log-ratios to that reference (see Sections 2.2 & 2.3).

First I load the data, renaming *Sample identifier* to *sample.id*, making the column names lower case, making *labcode* non-numeric since it's actually a label), and splitting the sample and area details out of *sample.id*. I then rounded the data to ~%1 percent, ensuring that the rounded percentages sum to 100% (using a standard rounding function for percentages that gets the floor integer value for each number, calculates each residuals part, and then distributes the sum of the residuals according their size ordering).

```r
round_percent <- function(percents.dataframe) {

  # Rounds a dataframe of percentages, ensuring that the sum of each row is 100%
  # Find integer components of the values
  # Find out how much we are missing
  # Distribute missing points according to remainders with random tie breaker

  ivals <- floor(percents.dataframe)
  isums <- rowSums(ivals)

  for (i in 1:length(isums)) {
    row.data <- percents.dataframe[i,]
    if(isums[i] < 100) {
      o <- order(row.data %% 1, sample(length(row.data)), decreasing=TRUE)
      ivals[i,][o[1:(100-isums[i])]] <- ivals[i,][o[1:(100-isums[i])]] + 1
```

[1] Oxford Instruments Helpdesk - tel: 0144479222, email: customer.support@oxinst.com

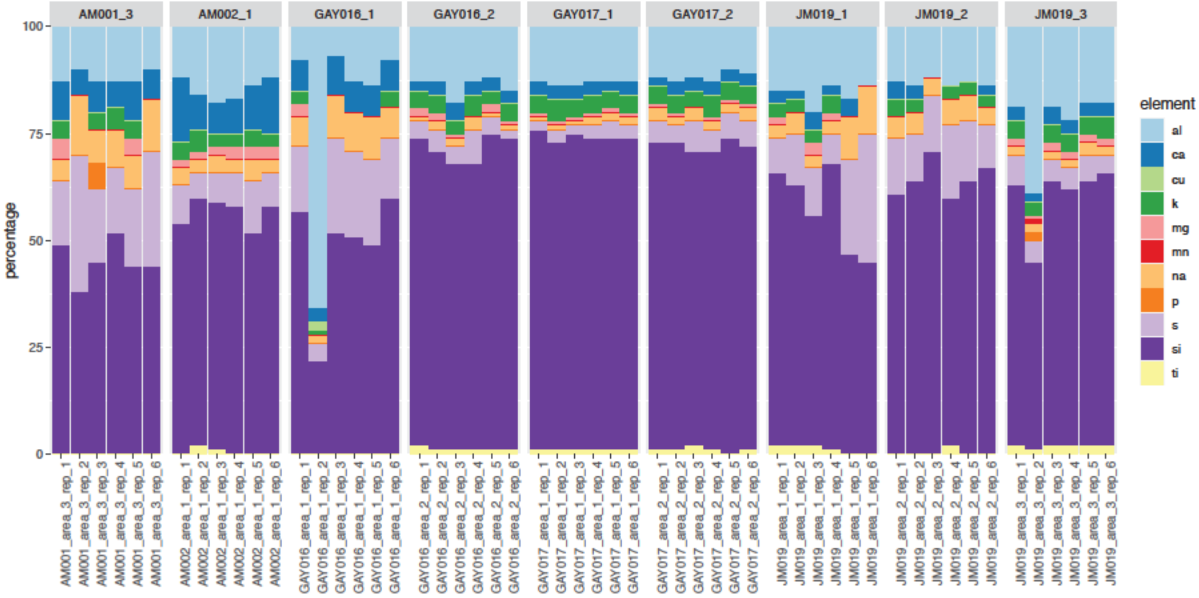
```
    }
  }

  return(ivals)
}

filename <- "data/SO Inquiry verified data for NS final.xlsx"

# load and manipulate the data, renaming sample identifier column, making
# column names lower case, labcode's non-numeric, and separating out sample
# locations and areas
soils.data <- read_excel(filename, sheet = "verified data elemental comp",
                         range = "A1:N55") %>%
  rename(sample.id = "Sample identifier") %>%
  rename_all(., .funs = tolower) %>%
  mutate(labcode = as.character(labcode)) %>%
  separate(sample.id, c("location", "x","area","y","rep")) %>%
  dplyr::select(-x, -y) %>%
  mutate(sample.id = paste0(location, "_area_", area, "_rep_", rep))

# round the data to 1%
round.data <- soils.data %>%
  select_if(is.numeric) %>%
  dplyr::select(-sum)

rounded.soils.data <- round_percent(round.data) %>%
  rowwise() %>%
  mutate(sum = sum(c_across(where(is.numeric)))) %>%
  bind_cols(soils.data %>% select_if(negate(is.numeric)))

# remove intervening variables
rm(round.data)
```

2.2.1 Data visualization

A key step in statistical analysis is to visualize the raw data. I used two types of visualization which are good for use with relative compositional data; stacked bar plots and radar charts.

2.2.1.1 Stacked Bar PLot Stacked bar plots show the relative composition data for each sample as a bar running from 0-100%, with the elemental percentage breakdowns shown within each bar shows as different colours. The stacked bar plot can summarise all the data in a single plot, which is useful, but it can be hard to make comparisons within and between groups, so here I split the plot into panels separating out the areas within each location.

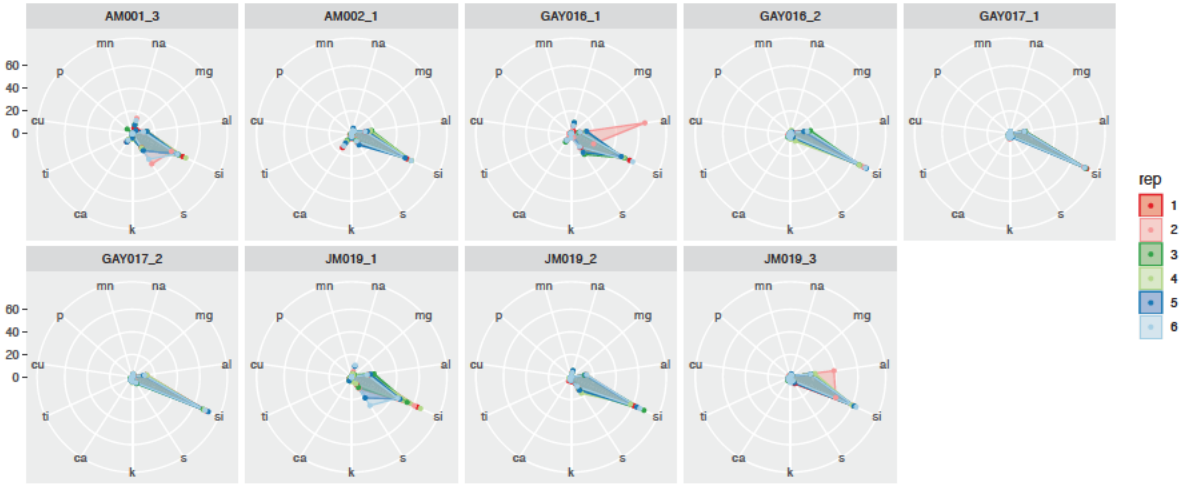
```
rounded.soils.data %>%
  mutate(group = paste0(location, "_", area)) %>%
  dplyr::select(-labcode, -sum, -location, -area, -rep) %>%
  pivot_longer(-c(sample.id, group), names_to="element",
               values_to="percentage") %>%
  ggplot(aes(x = sample.id, y = percentage, fill = element)) +
    geom_bar(stat = "identity") +
    ylim(c(0, 100)) +
    labs(y = "percentage") +
    scale_y_continuous(expand = c(0, 0), limits = c(0, NA)) +
    theme(axis.text.x = element_text(angle = 90, vjust = 1, hjust=1)) +
    facet_wrap(~group, scales = "free_x", nrow=1) +
    scale_fill_brewer(palette = "Paired") +
    theme(axis.title.x=element_blank())
```



2.2.1.2 Radar charts

Radar charts show each of the elements as a different axis, evenly distributed around a circle, with points plotted on each axis for the elemental composition percentage and lines connecting the points form a shaded polygon. The polygons shapes indicate similar/different sets of values. Again we break the plot into panels for the each locations and area to visualize the consistency of the replicates within each area and contrast the polygon shapes across areas.

```
rounded.soils.data %>%
  mutate(group = paste0(location, "_", area)) %>%
  dplyr::select(-labcode, -sum, -location, -area) %>%
  relocate(sample.id) %>%
  ggRadar(mapping = aes(colour = rep, facet = group),
          rescale = FALSE, interactive = FALSE, legend.position = "right",
          size = 1, alpha = 0.4) +
    facet_wrap(~group, ncol=5) +
    scale_fill_brewer(palette = "Paired", direction = -1) +
    scale_colour_brewer(palette = "Paired", direction = -1)
```



The stacked bar plots, and the radar charts show that silicon is the dominant signal in all the samples. The plots also suggest some potential similarities and differences between the samples:

1. AM001 and AM002 appear similar but have some consistent differences; AM001 has larger fractions of Sulphur and Sodium, while AM002 has larger fraction of Calcium.

2. GAY016 and GAY017 appear similar, however there is a consistent difference between GAY016 area 1 which has no Titanium and increased fractions of Sulphur, Sodium and Calcium, compared to the other samples. Replicate 2 of GAY016 area 1 is significantly different from all the other samples and looks like it may be a failed sample.

3. JM019 areas 1 & 2 appear similar, although the variability between the replicates within area 1 is large.

4. JM019 area 3 has good consistency between replicates, except for replicate 2 which is significantly different from the others and looks like it may also be a failed sample. JM019 area 3 appears to have a consistently higher aluminium fraction than the other JM019 areas.

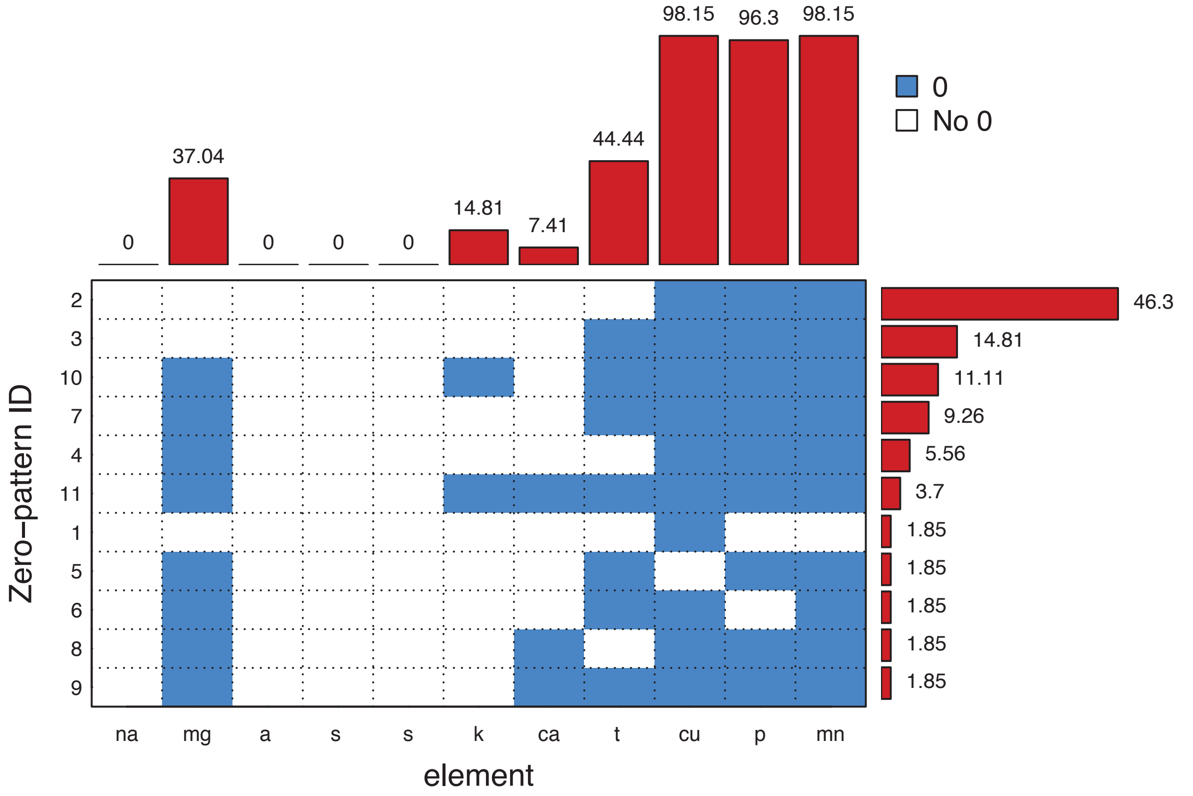
2.2.2 Data imputation

The characteristics of relative compositional data makes the use of most standard statistical approaches unsuitable for use directly with these data. This limitation is frequently mitigated by transforming the data into log-ratios where, for a set of relative composition data $x = (x_1, ..x_n)$, we define one value as a reference value (say, x_1) and then transform the rest of the data so that $y_{1..n-1} = log_{10}(x_{2..n}/x_1)$). The resulting log-ratios are typically normally distributed and are not bounded, removing the primary statistical constraints of working with relative composition data.

An additional problem with this data is the presence of zero values for some elements which prevent us from transforming the data into log-ratios. From this data, we are not able to distinguish between true zeros and zeros that originate from trace concentrations below the detection threshold of the EDXA method. In order to construct the log ratios, I have made the assumption that none of the zeros are true zeros and that all the elements exist at trace levels in all the samples. Using the detection limit for the technique, we can then impute non-zero values that are below the detection limit for the zeros values. The trace element detection limit for the EDXA system is likely to be element-dependent (because the background signal from elastic X-ray scattering is not uniform over all energies) and dependent of the energy used for the electron beam, but discussion with Oxford Instruments indicated that these were relatively small effects that a trace element detection threshold of approximately 0.1% is an absolute best case scenario. Discussions with the Soil Forensics team at the James Hutton Institute (JHI), and with the team at Robert Gordon University that generated the data, suggested a detection threshold limit of ~1% is more realistic for complex soils, so this is the detection threshold I use here.

To impute the data below this value we'll use the R package zCompositions, which is specifically designed for data imputation of left-censored compositional data. We'll use the `lrEM` function from zCompositions, which uses an expectation maximization approach to produce conditional estimates based on the patterns of missing data. It assumes multivariate normality and we use the default value of 0.65 of the detection limit as the mean for the simulations (recommended by Martin-Fernandez et. al. 2003, doi: 10.1023/A:1023866030544).

```r
# set the random seed so the results are reproducible
set.seed(379456)

# get just the numeric data
impute.data <- rounded.soils.data %>%
  select_if(is.numeric) %>%
  dplyr::select(-sum)

# visualize the patterns of zeros
zero.patterns <- zPatterns(impute.data, label = 0, bar.labels = TRUE,
                           cex.axis = 0.8, bar.ordered = c(TRUE, FALSE),
                           plot=TRUE,
                           axis.labels = c("element", "Zero-pattern ID"))
```



```
# impute the data. The low number of complete cases makes initial estimation of
# the covariance matrix difficult, so here we use the multRepl option to replace
# the zeroes with dl*0.65 initially, and then compute the covariance matrix.
imputed.soils.data <- lrEM(impute.data, label = 0, ini.cov = "multRepl",
                          dl = rep(1, ncol(impute.data)))

rm(impute.data)
```

Looking at the imputed data, the phosphorus values look very different from the other imputed data and one of the two phosphorus signals is in an outlier, potentially failed, replicate. I'm not convinced that the Phosphorus data is informative so I remove this element from the subsequent analysis.

```
# remove phosphorus
imputed.soils.data <- imputed.soils.data %>%
  dplyr::select(-p)

# rescale the percentages back to 100% total after removing phosphorus
row.sums <- rowSums(imputed.soils.data)
imputed.soils.data <- imputed.soils.data * (100/row.sums)

# rebind the metadata columns and make the sum column as a sanity check
imputed.soils.data <- imputed.soils.data %>%
  rowwise() %>%
  mutate(sum = sum(c_across(where(is.numeric)))) %>%
  bind_cols(soils.data %>% select_if(negate(is.numeric)))

rm(row.sums)
```

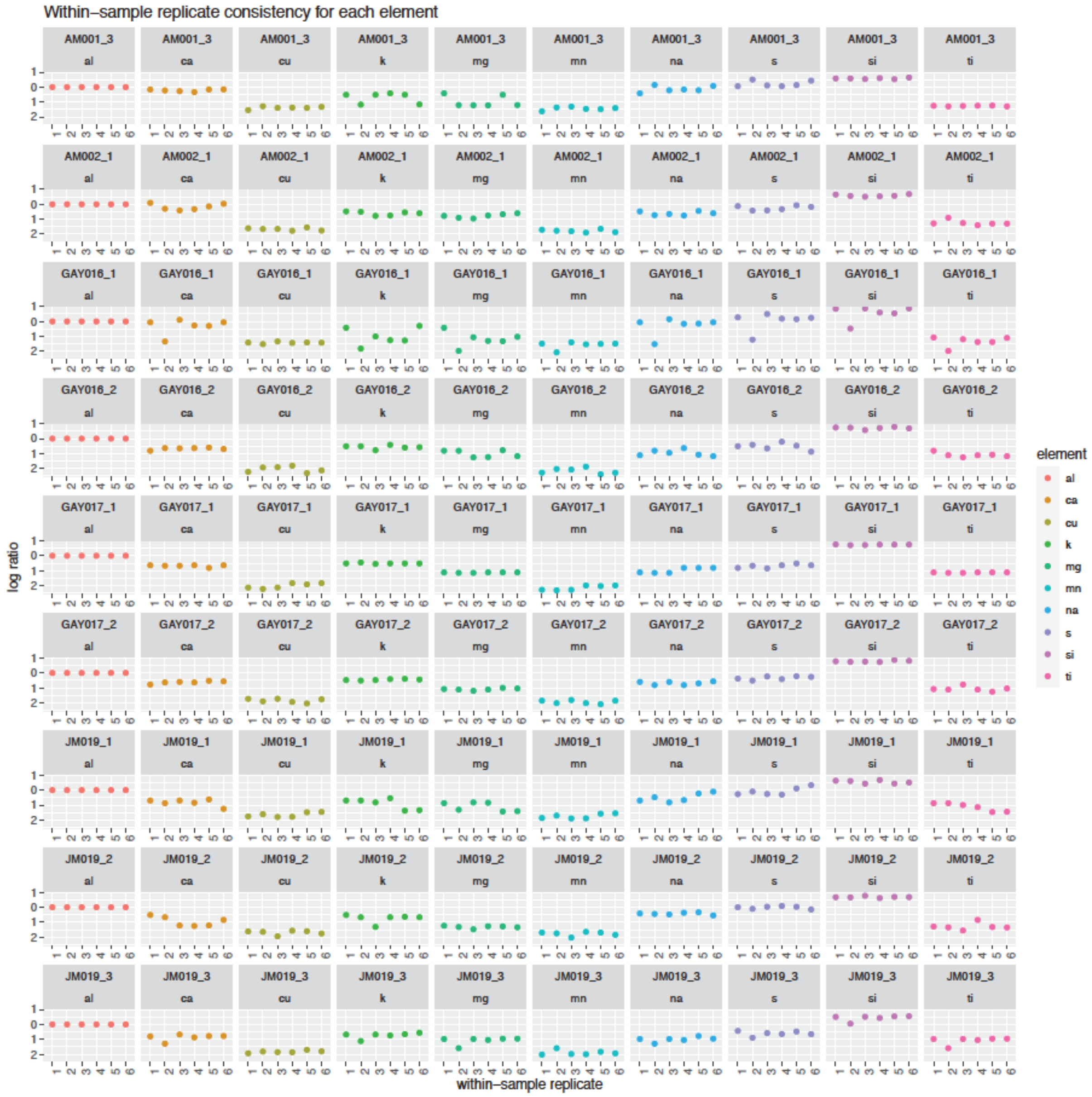
2.2.3 log-ratio transformation

Next, I construct the log ratios. Ideally the reference element for this would be present in all the samples at relatively consistent levels. Initially silicon looked like a good choice, but discussion with the Soil Forensics team at JHI suggested that this element was indicative of the sand content of the soils and so was particularly important as a distinguishing contrast and might not make the best reference point. Aluminium was identified as the sensible alternative and is used as the reference element here because it's present in all the samples with a relatively consistent fraction, and is commonly found in a wide variety of soils.

```
lr.data <- imputed.soils.data %>%
  select_if(is.numeric) %>%
  dplyr::select(-sum)

lr.data <- log10(lr.data/lr.data$al)

lr.soils.data <- lr.data %>%
  bind_cols(soils.data %>% select_if(negate(is.numeric)))

rm(lr.data)
```

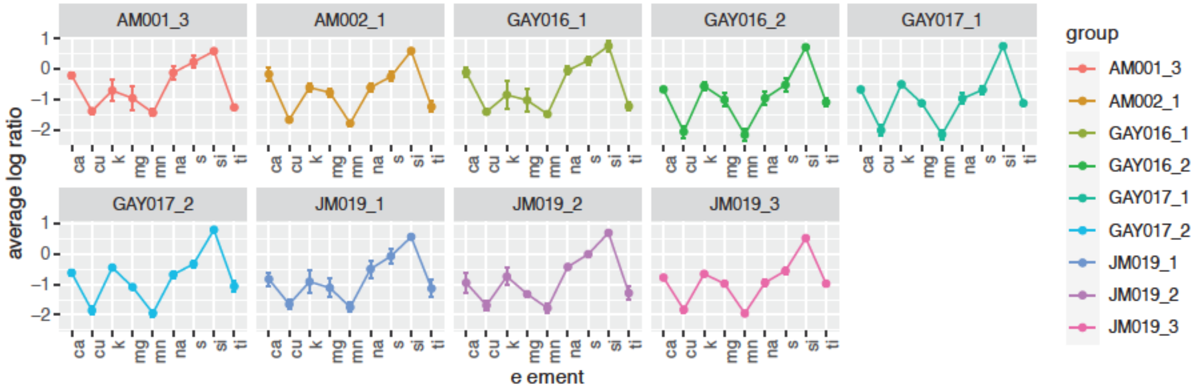
Visualizing the within-replicate stability of the log-ratios for each element, for each sample, gives us an indication of how consistent the log-ratio values are between the replicate samples within each area.

```
lr.soils.data %>%
  mutate(group = paste0(location, "_", area)) %>%
  dplyr::select(-labcode, -location, -area) %>%
  pivot_longer(-c(sample.id, group, rep), names_to="element",
               values_to="log.ratio") %>%
  ggplot(aes(x = rep, y = log.ratio, colour = element)) +
    geom_point() +
    labs(y = "log ratio", x= "within-sample replicate",
         title = "Within-sample replicate consistency for each element") +
    theme(axis.text.x = element_text(angle = 90, vjust = 1, hjust=1)) +
    facet_wrap(~group+element, scales = "free_x", ncol=10)
```



The within-replicate consistency of the log-ratios looks excellent in most cases, with only the potentially failed outlier replicates (GAY016 area 1 replicate 2 & JM019 area 3 replicate 2) showing significant deviation. I have therefore removed these replicates before summarizing the log-ratio data across the replicates within each location area with a mean log-ratio and it's standard deviation. The plots below show how this elemental signature pattern differs between samples and areas.

```
mean.lr.soils.data <- lr.soils.data %>%
  filter(!(location == "GAY016" & area == 1 & rep == 2)) %>%
  filter(!(location == "JM019" & area == 3 & rep == 2)) %>%
  dplyr::select(-al) %>%
  pivot_longer(-c(labcode, location, area, rep, sample.id),
               names_to="element", values_to="log.ratio") %>%
  group_by(labcode, location, area, element) %>%
  summarize(mean.lr = mean(log.ratio),
            sd.lr = sd(log.ratio))

mean.lr.soils.data %>%
  mutate(group = paste0(location, "_", area)) %>%
  ggplot(aes(x = element, y = mean.lr, ymin = mean.lr-sd.lr,
             ymax = mean.lr+sd.lr, colour = group, group = group)) +
    geom_errorbar(width=0.2) +
    geom_point() +
    geom_line() +
    labs(y = "average log ratio") +
    theme(axis.text.x = element_text(angle = 90, vjust = 1, hjust=1)) +
    facet_wrap(~group, scales = "free_x", ncol=5)
```



2.3 Sample comparison

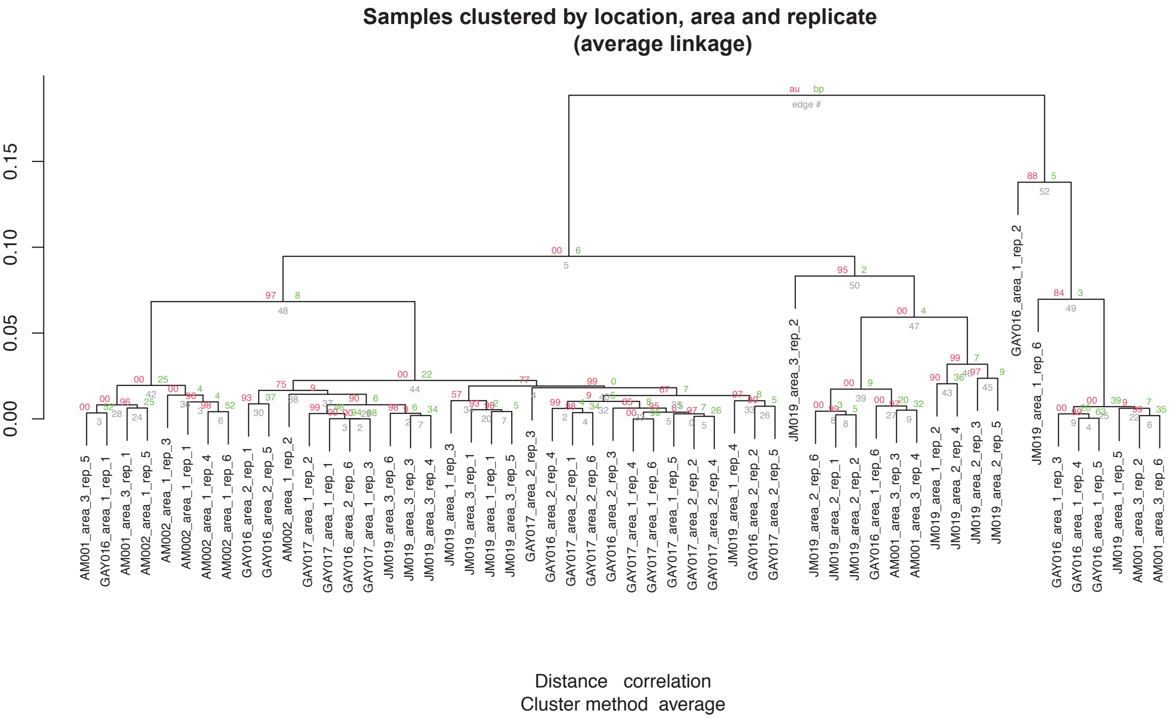
Now we can assess the sample similarity, beginning with how the samples cluster using the bootstrapped hierarchical clustering routine `pvclust`, and then using an unsupervised Principal Components Analysis.

2.3.1 Hierarchical clustering

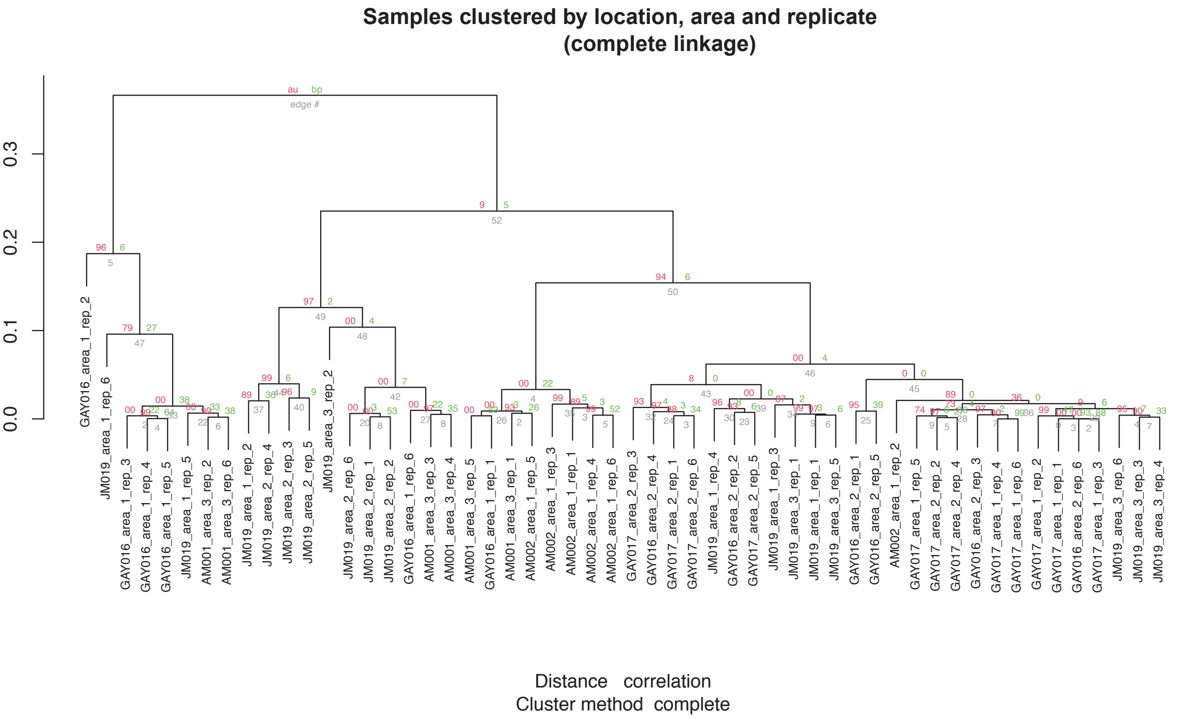
The hierarchical clustering performed here uses correlation distance (the Pearson's correlation between the sample data vectors) as the clustering distance metric, and uses two different clustering linkages - complete (the most conservative) and average (less conservative), in order to get an idea of how robust the patterns of clustering are to changes in the clustering parameters. Note that `pvclust` uses multi-scale bootstrapping to re-sample the clustering trees and assign a p-value to each vertex in the tree. These come in two varieties - the bootstrapping p-value and the corrected Approximately Unbiased p-value. The latter of these is a robust measure of the support for a given clustering vertex in the dataset.

```
# grab the numerical data only and convert it to the right format
lr.soils.clustering.data <- lr.soils.data %>%
  dplyr::select(-c(labcode, location, area, rep, sample.id, al)) %>%
  t()
colnames(lr.soils.clustering.data) <- lr.soils.data$sample.id

# do the clustering
lr.soils.clustering.average <- pvclust(lr.soils.clustering.data,
                                       method.hclust="average",
                                       method.dist = "correlation",
                                       quiet=TRUE)

lr.soils.clustering.complete <- pvclust(lr.soils.clustering.data,
                                       method.hclust="complete",
                                       method.dist = "correlation",
                                       quiet=TRUE)
```

```
plot(lr.soils.clustering.average,
     main = "Samples clustered by location, area and replicate \
            (average linkage)", cex=0.7, cex.pv=0.5)
```



```
plot(lr.soils.clustering.complete,
     main = "Samples clustered by location, area and replicate \
            (complete linkage)", cex=0.7, cex.pv=0.5)
```

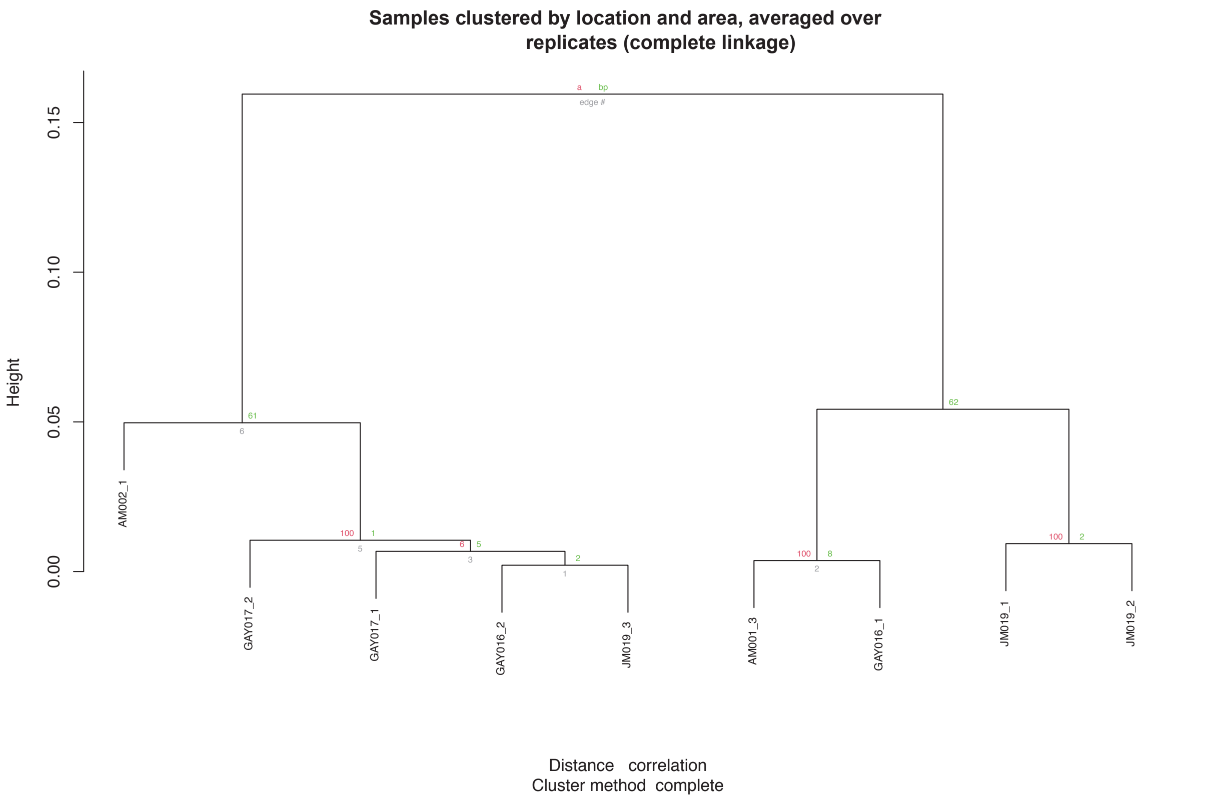


These clustering plots show the complexity of the relationships between the samples, the variability between replicates and the p-values show how reproducible the clustering is. The patterns of clustering are stable to changes in the clustering parameters, except at low grouping levels where some changes are seen (as expected). The clustering shows that the JM019 area 3 replicates cluster most tightly with GAY017 areas 1 & 2 and GAY016 area 2. JM019 area 2 replicates cluster most tightly with AM001 area 3 replicates although some AM001 area three replicates cluster elsewhere, and JM019 area 1 replicates are noisy and are spread across the clustering. The clustering also highlights how different the outlier replicates (GAY016 area 1 replicate 2 & JM019 area 3 replicate 2) are from the other replicates.

We can also use the mean log-ratio values (with the outlier replicates removed) to perform the clustering to get a simpler picture, although this analysis does not then encompass the range of the within-sample replicate variation so care should be taken not to over interpret the strength of the results from this. Only one clustering is shown here because the clustering is totally insensitive to changes in clustering parameters.

```
# grab the numerical data only and convert it to the right format
lr.soils.clustering.data <- mean.lr.soils.data %>%
  pivot_wider(-sd.lr, names_from=element, values_from = mean.lr) %>%
  mutate(group = paste0(location, "_", area))

clustering.colnames <- lr.soils.clustering.data$group

lr.soils.clustering.data <- lr.soils.clustering.data %>%
  dplyr::select(-group) %>%
  ungroup() %>%
  dplyr::select(-c(labcode, location, area)) %>%
  t()

colnames(lr.soils.clustering.data) <- clustering.colnames

# do the clustering
lr.soils.clustering.average <- pvclust(lr.soils.clustering.data,
                                       method.hclust="average",
                                       method.dist = "correlation",
                                       quiet=TRUE)

lr.soils.clustering.complete <- pvclust(lr.soils.clustering.data,
                                       method.hclust="complete",
                                       method.dist = "correlation",
                                       quiet=TRUE)
```

```
# plot the results
plot(lr.soils.clustering.complete,
     main = "Samples clustered by location and area, averaged over \
            replicates (complete linkage)", cex=0.7, cex.pv=0.5)
```



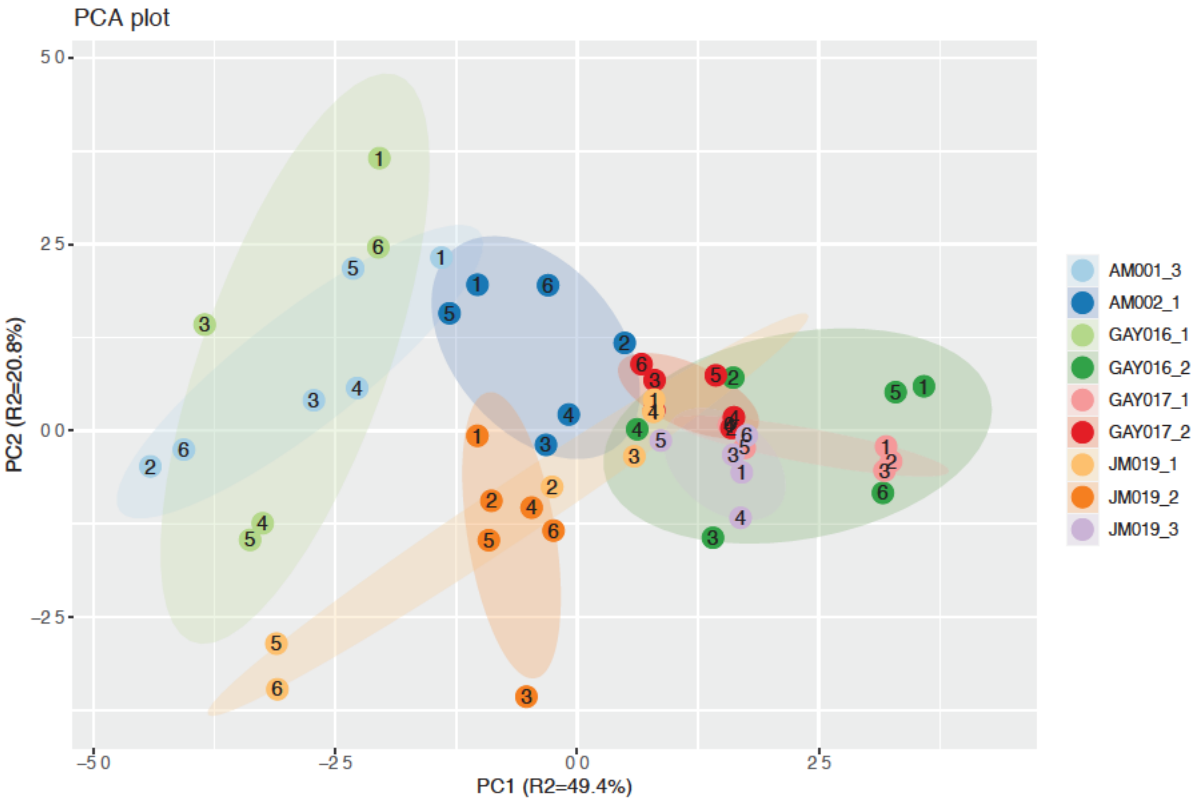
2.3.2 PCA

Principal Component Analysis is a dimensionality reduction method that identifies sets of eigenvectors that are linear combinations of the dimensions of the original data (in this case the elements) and which preserve the maximum amount of information (and thus explain the maximum about of variation) from the original data. For the PCA calculation here we'll use two Principal Components (PCs, because we're primarily interested in making a clustering plot for the data) and we'll center and scale the data for unit variance first. PCs are sensitive to outliers so the two outlier replicates identified previously (GAY016 area 1 replicate 2 and JM019 area 3 replicate 2) are removed before building the PCs.

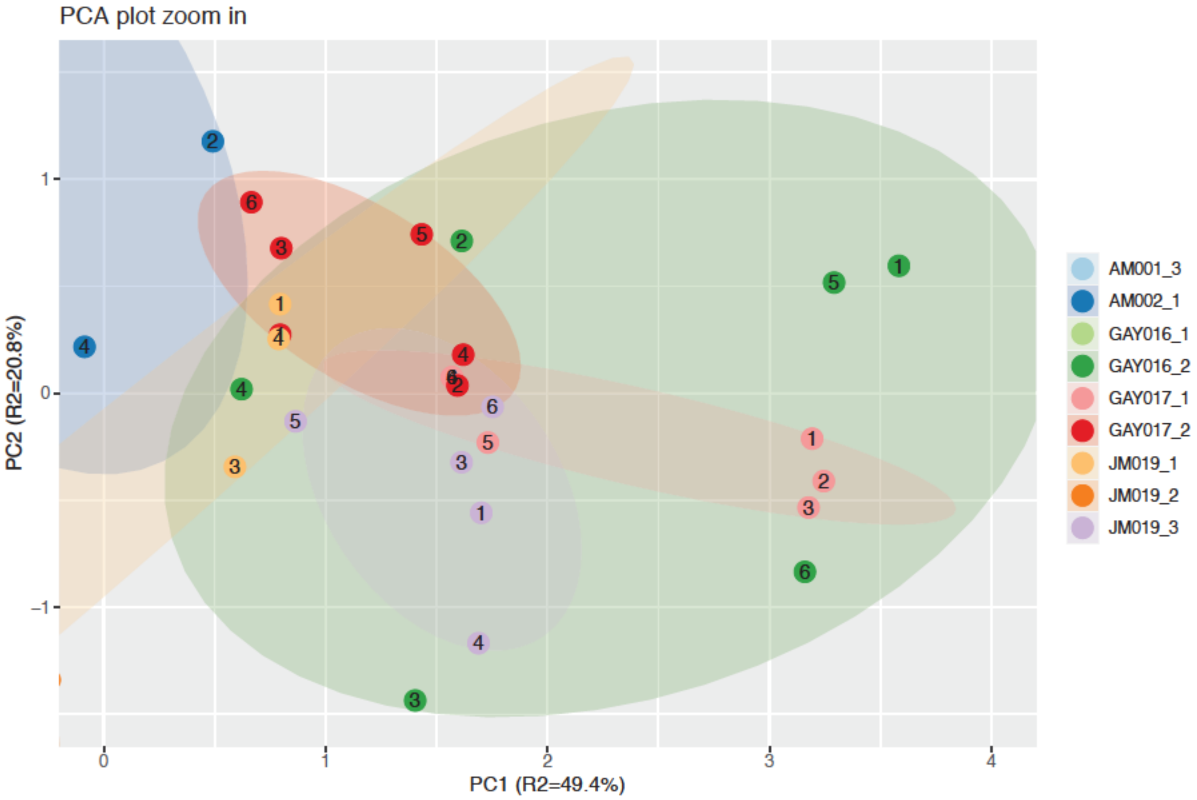
```
lr.soils.pca.data <- lr.soils.data %>%
  filter(!(location=="GAY016" & area == "1" & rep == "2")) %>%
  filter(!(location=="JM019" & area == "3" & rep == "2")) %>%
  dplyr::select(-al)

lr.soils.pca <- lr.soils.pca.data %>%
  pca(nPcs=2, scale="uv", center=TRUE)
```

Plotting each of the samples using their new Principal Component coordinates shows how the sample replicates cluster, and how the samples relate to each other is shown with a shaded ellipse that identifies the indicative region of the PC-space that is occupied by each of the samples (specifically, this ellipse is defined by one standard deviation of the multivariate normal distribution for each group).

```
cbind(lr.soils.pca.data, scores(lr.soils.pca)) %>%
  mutate(pca.label = paste0(location, "_", area)) %>%
  ggplot(aes(x=PC1, y=PC2)) +
    stat_ellipse(aes(fill=pca.label), type = "norm", level = 0.66,
                 geom="polygon", alpha=0.2) +
    geom_point(aes(colour=pca.label), size=5) +
    geom_text(aes(label=rep), check_overlap = FALSE, hjust = "centre",
               vjust = "centre", nudge_y = 0) +
    theme(legend.title = element_blank()) +
    labs(title = "PCA plot",
         x = sprintf("PC1 (R2=%.1f%s)", lr.soils.pca@R2[1]*100, "%"),
         y = sprintf("PC2 (R2=%.1f%s)", lr.soils.pca@R2[2]*100, "%")) +
    scale_fill_brewer(palette = "Paired") +
    scale_colour_brewer(palette = "Paired")
```



```
cbind(lr.soils.pca.data, scores(lr.soils.pca)) %>%
  mutate(pca.label = paste0(location, "_", area)) %>%
  ggplot(aes(x=PC1, y=PC2)) +
    stat_ellipse(aes(fill=pca.label), type = "norm", level = 0.66,
                 geom="polygon", alpha=0.2) +
    geom_point(aes(colour=pca.label), size=5) +
    geom_text(aes(label=rep), check_overlap = FALSE, hjust = "centre",
               vjust = "centre", nudge_y = 0) +
    theme(legend.title = element_blank()) +
  labs(title = "PCA plot zoom in",
     x = sprintf("PC1 (R2=%.1f%s)", lr.soils.pca@R2[1]*100, "%"),
     y = sprintf("PC2 (R2=%.1f%s)", lr.soils.pca@R2[2]*100, "%")) +
  coord_cartesian(xlim=c(0, 4), ylim=c(-1.5, 1.5)) +
  scale_fill_brewer(palette = "Paired") +
  scale_colour_brewer(palette = "Paired")
```



The first two PCs capture the majority of the variation in the original sample (70%), after removal of the two outlier replicates. For at least some of the samples, the replicates cluster tightly (e.g. JM019 area 3 and GAY017 area 2), suggesting that these samples are homogeneous and the sample collection and analysis is robust and accurate, and can be trusted. Some of the samples (e.g. JM019 area 1) are not tightly clustered, suggesting that these samples are heterogeneous either because the soil in the sample is from a single source that is naturally heterogeneous or because the soil is from mixed origins. Examining where the questioned samples sit relative to the other samples we see that:

- The JM019 area 3 replicate cluster overlaps strongly with the clustering of GAY016 area 2, GAY017 area 1 and, to a lesser degree, GAY017 area 2. This indicates that these samples are consistent with sharing a common origin.
- The GAY016 area 2 and GAY017 area 1 replicate clusters overlap strongly, and both show two distinct subsets of replicates, well separated by PC1. These samples are likely to be related.
- The JM019 area 1 replicates are not tightly clustered, and do not consistently cluster with a distinct set of the other samples, suggesting this sample is highly heterogeneous and may not be from a single origin.

Table 1: Principal Component loadings

	PC1	PC2
na	-0.46	0.06
mg	0.07	0.58
si	0.11	0.26
s	-0.44	0.05
k	0.26	0.49
ca	-0.21	0.52
ti	0.25	0.24
cu	-0.45	0.10
mn	-0.45	0.11

- The JM019 area 2 replicates cluster tightly, and are well separated from the other samples, suggesting a separate origin for this soil.
- The AM001 replicates and the GAY016 area 1 replicates are not tightly clustered, but they are well separated from other samples and overlap strongly, suggesting a separate origin for these two samples.

2.3.2.1 PCA loadings Examining the weighting of the element contributions to each of PCs is important to get an idea of whether they are dominated with one or two elements, or whether all the elements are contributing. The loadings here show that the principal components are contributed to by a wide range of elements, with titanium and silicon levels contributing the least information.

```
kable(lr.soils.pca@loadings, digits=2,
      caption = "Principal Component loadings") %>%
  kable_styling(full_width = FALSE)
```

2.4 Quantifying the probability of soil sample relationships

The above exploratory analysis presents some context and evidence for possible relationships between the questioned samples from the vest and the other samples from the footwear, and presents a quantification for how distinct the samples are, but does not directly address the question of how likely it is that data from a questioned sample is drawn from the same parent population as a known sample.

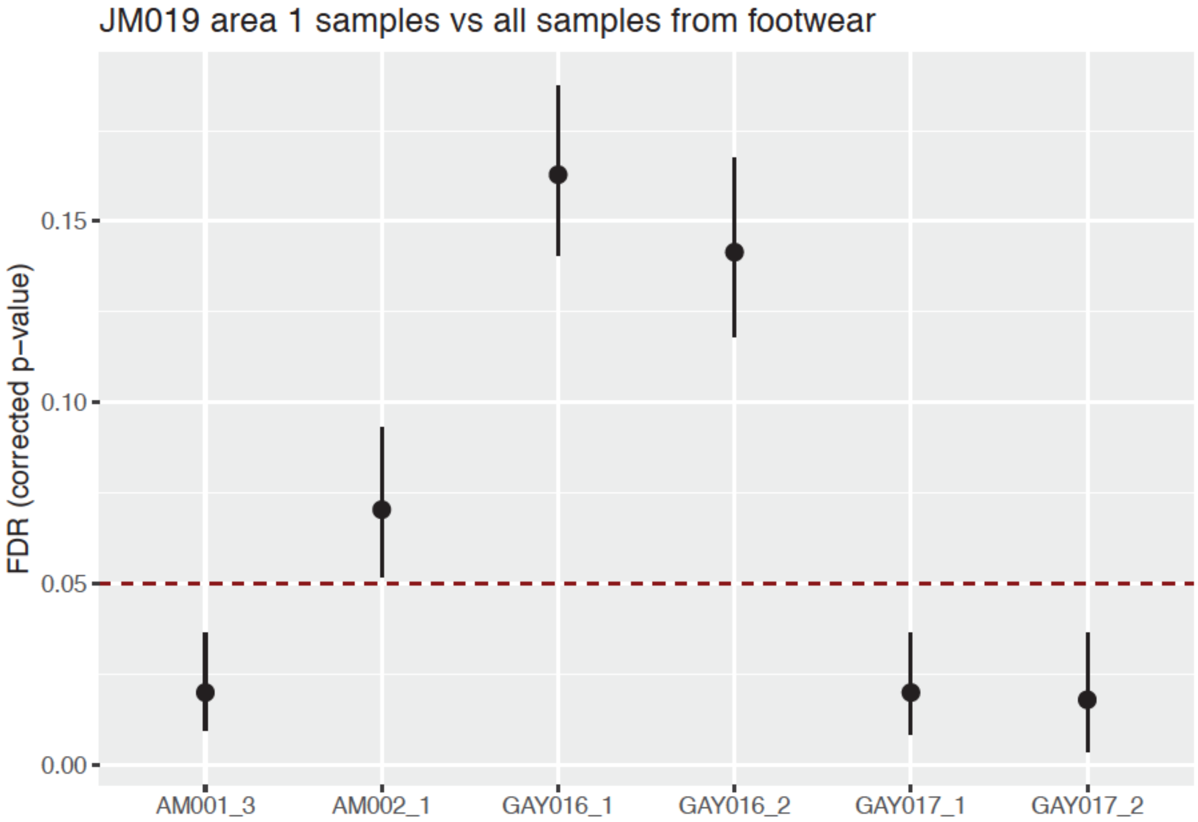
To quantify this I use an extension of the non-parametric two-sample Kolmogorov-Smirnov statistical test for two-dimensional datasets (Fasano & Franceschini, 1987, MNRAS, doi: 10.1093/mnras/225.1.155). This widely used test quantifies this probability based on the point of greatest distance between the empirical distribution functions (EDF) of the multivariate datasets. I used principal components identified in the previous section to compare the samples, which integrate information from across the elemental profiles of the soils, to maximize the sensitivity of the test. Importantly, each questioned sample is compared against six known samples (six tests), necessitating a multiple testing correction for the resulting p-values (the standard False Discovery Rate multiple testing correction method from Benjamini, Hochberg, and Yekutieli is used, which controls the expected proportion of false discoveries among the rejected hypotheses to less than 5%).

It is important to recall that this is a hypothesis test that looks for the strength of evidence to reject a null hypothesis. In this case, **the null hypothesis for these tests is that the data from the questioned and known sample being tested were drawn from the same underlying parent distribution. Low FDRs (< 0.05) mean we can reject this null hypothesis and can conclude that the data come from different parent distributions, implying a different soil origin. High FDRs (> 0.05) mean that we cannot reject the null hypothesis and we cannot rule out that the data come from the same parent distributions, indicating that the samples are consistent with having come from the same soil origin.** These tests explicitly do not reject the possibility an alternate untested source for either conclusion or conclusively identify the soils as definitely originating from the same location.

```
pval.cor.method = "fdr"

JM019_1 <- cbind(lr.soils.pca.data, scores(lr.soils.pca)) %>%
  filter(location == "JM019" & area == "1") %>%
  dplyr::select(PC1, PC2)

JM019_2 <- cbind(lr.soils.pca.data, scores(lr.soils.pca)) %>%
  filter(location == "JM019" & area == "2") %>%
  dplyr::select(PC1, PC2)

JM019_3 <- cbind(lr.soils.pca.data, scores(lr.soils.pca)) %>%
  filter(location == "JM019" & area == "3") %>%
  filter(!(rep == "2")) %>% # exclude outlier rep
  dplyr::select(PC1, PC2)

GAY016_1 <- cbind(lr.soils.pca.data, scores(lr.soils.pca)) %>%
  filter(location == "GAY016" & area == "1") %>%
  filter(!(rep == "2")) %>% # exclude outlier rep
  dplyr::select(PC1, PC2)

GAY016_2 <- cbind(lr.soils.pca.data, scores(lr.soils.pca)) %>%
  filter(location == "GAY016" & area == "2") %>%
  dplyr::select(PC1, PC2)

GAY017_1 <- cbind(lr.soils.pca.data, scores(lr.soils.pca)) %>%
  filter(location == "GAY017" & area == "1") %>%
  dplyr::select(PC1, PC2)

GAY017_2 <- cbind(lr.soils.pca.data, scores(lr.soils.pca)) %>%
  filter(location == "GAY017" & area == "2") %>%
  dplyr::select(PC1, PC2)

AM001_3 <- cbind(lr.soils.pca.data, scores(lr.soils.pca)) %>%
  filter(location == "AM001" & area == "3") %>%
  dplyr::select(PC1, PC2)

AM002_1 <- cbind(lr.soils.pca.data, scores(lr.soils.pca)) %>%
  filter(location == "AM002" & area == "1") %>%
  dplyr::select(PC1, PC2)
```

2.4.1 JM019 Area 1

```
vsGAY016_1 <- fasano.franceschini.test(JM019_1, GAY016_1, threads = 4,
                                       seed = 273329, nPermute = 1000)
vsGAY016_2 <- fasano.franceschini.test(JM019_1, GAY016_2, threads = 4,
                                       seed = 273329, nPermute = 1000)
vsGAY017_1 <- fasano.franceschini.test(JM019_1, GAY017_1, threads = 4,
                                       seed = 273329, nPermute = 1000)
vsGAY017_2 <- fasano.franceschini.test(JM019_1, GAY017_2, threads = 4,
                                       seed = 273329, nPermute = 1000)
vsAM001_3 <- fasano.franceschini.test(JM019_1, AM001_3, threads = 4,
                                      seed = 273329, nPermute = 1000)
vsAM002_1 <- fasano.franceschini.test(JM019_1, AM002_1, threads = 4,
                                      seed = 273329, nPermute = 1000)


JM019_1vs <- data.frame(
  name = c("GAY016_1", "GAY016_2", "GAY017_1", "GAY017_2",
           "AM001_3", "AM002_1"),
  corrected.pval = p.adjust(c(vsGAY016_1$p.value, vsGAY016_2$p.value,
                              vsGAY017_1$p.value, vsGAY017_2$p.value,
                              vsAM001_3$p.value, vsAM002_1$p.value),
                            method = pval.cor.method),
  low = p.adjust(c(vsGAY016_1$conf.int[1], vsGAY016_2$conf.int[1],
                   vsGAY017_1$conf.int[1], vsGAY017_2$conf.int[1],
                   vsAM001_3$conf.int[1], vsAM002_1$conf.int[1]),
                 method = pval.cor.method),
  high = p.adjust(c(vsGAY016_1$conf.int[2], vsGAY016_2$conf.int[2],
                    vsGAY017_1$conf.int[2], vsGAY017_2$conf.int[2],
                    vsAM001_3$conf.int[2], vsAM002_1$conf.int[2]),
                  method = pval.cor.method))

JM019_1vs %>%
  ggplot(aes(x = name, y = corrected.pval, ymin = low, ymax = high)) +
  geom_pointrange() +
  labs(title = "JM019 area 1 samples vs all samples from footwear",
       y = "FDR (corrected p-value)") +
  geom_hline(yintercept=0.05, linetype = 2, colour = "darkred") +
  theme(axis.title.x=element_blank())
```



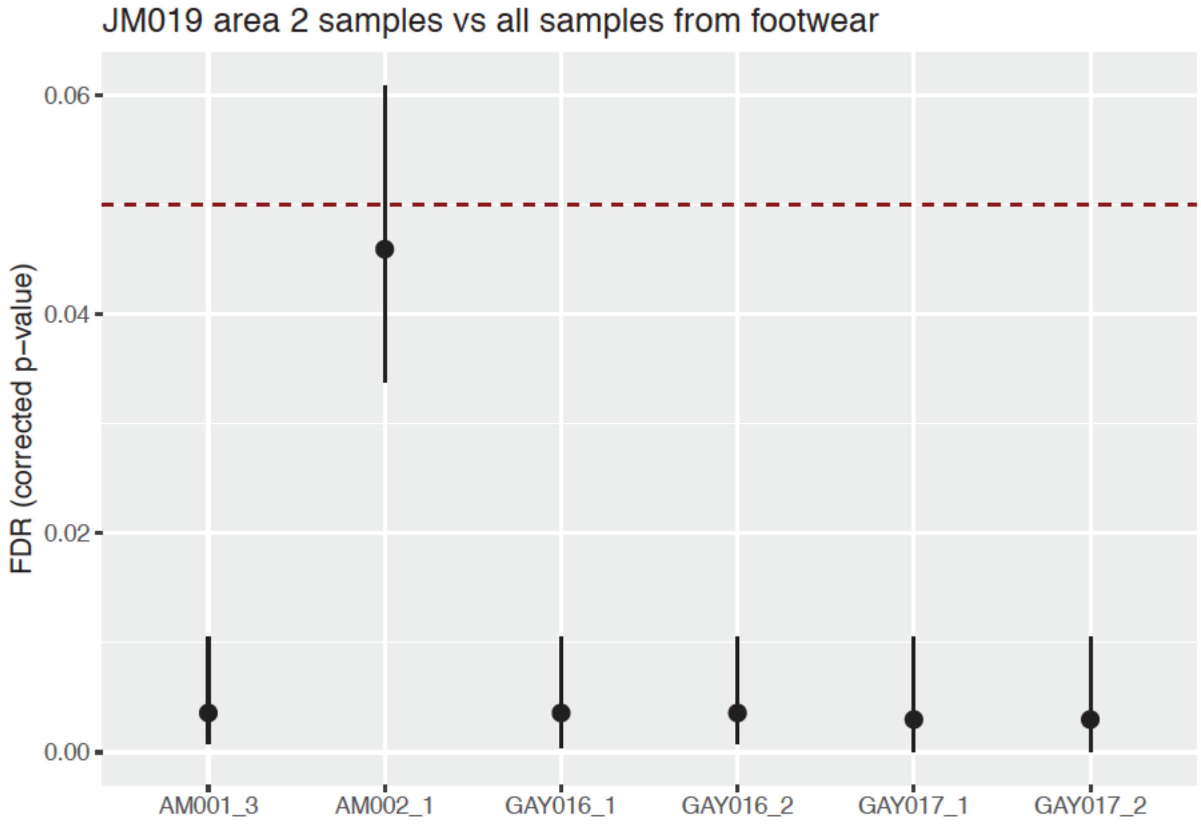
These results suggest that we can reject the null hypothesis for AM001 area 3 and both GAY017 areas and thus conclude that:

- the JM019 area 1 sample is not consistent with sharing a common origin with the AM001 area 3 and GAY017 soils.
- the JM019 area 1 sample is consistent with sharing a common origin with the AM002 area 1 and GAY016 soils.

2.4.2 JM019 Area 2

```
vsGAY016_1 <- fasano.franceschini.test(JM019_2, GAY016_1, threads = 4,
                                       seed = 273329, nPermute = 1000)
vsGAY016_2 <- fasano.franceschini.test(JM019_2, GAY016_2, threads = 4,
                                       seed = 273329, nPermute = 1000)
vsGAY017_1 <- fasano.franceschini.test(JM019_2, GAY017_1, threads = 4,
                                       seed = 273329, nPermute = 1000)
vsGAY017_2 <- fasano.franceschini.test(JM019_2, GAY017_2, threads = 4,
                                       seed = 273329, nPermute = 1000)
vsAM001_3 <- fasano.franceschini.test(JM019_2, AM001_3, threads = 4,
                                      seed = 273329, nPermute = 1000)
vsAM002_1 <- fasano.franceschini.test(JM019_2, AM002_1, threads = 4,
                                      seed = 273329, nPermute = 1000)
```

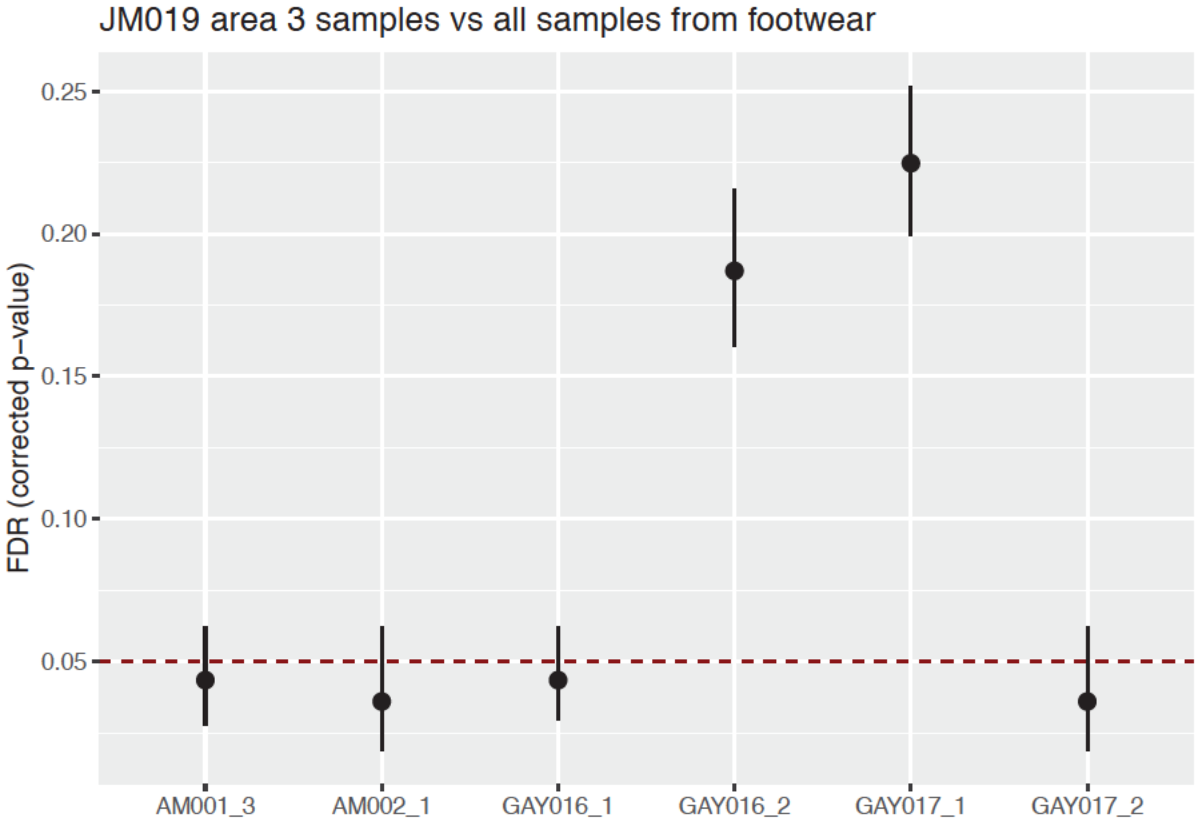
```r
JM019_2vs <- data.frame(
  name = c("GAY016_1", "GAY016_2", "GAY017_1", "GAY017_2",
           "AM001_3", "AM002_1"),
  corrected.pval = p.adjust(c(vsGAY016_1$p.value, vsGAY016_2$p.value,
                              vsGAY017_1$p.value, vsGAY017_2$p.value,
                              vsAM001_3$p.value, vsAM002_1$p.value),
                            method = pval.cor.method),
  low = p.adjust(c(vsGAY016_1$conf.int[1], vsGAY016_2$conf.int[1],
                   vsGAY017_1$conf.int[1], vsGAY017_2$conf.int[1],
                   vsAM001_3$conf.int[1], vsAM002_1$conf.int[1]),
                 method = pval.cor.method),
  high = p.adjust(c(vsGAY016_1$conf.int[2], vsGAY016_2$conf.int[2],
                    vsGAY017_1$conf.int[2], vsGAY017_2$conf.int[2],
                    vsAM001_3$conf.int[2], vsAM002_1$conf.int[2]),
                  method = pval.cor.method))

JM019_2vs %>%
  ggplot(aes(x = name, y = corrected.pval, ymin = low, ymax = high)) +
  geom_pointrange() +
  labs(title = "JM019 area 2 samples vs all samples from footwear",
       y = "FDR (corrected p-value)") +
  geom_hline(yintercept=0.05, linetype = 2, colour = "darkred") +
  theme(axis.title.x=element_blank())
```



These results suggest that we can reject the null hypothesis for all the soils and conclude that the JM019 area 2 soil is not consistent with sharing a common origin with any of the other soils examined here.

2.4.3 JM019 Area 3

```
vsGAY016_1 <- fasano.franceschini.test(JM019_3, GAY016_1, threads = 4,
                                        seed = 273329, nPermute = 1000)
vsGAY016_2 <- fasano.franceschini.test(JM019_3, GAY016_2, threads = 4,
                                        seed = 273329, nPermute = 1000)
vsGAY017_1 <- fasano.franceschini.test(JM019_3, GAY017_1, threads = 4,
                                        seed = 273329, nPermute = 1000)
vsGAY017_2 <- fasano.franceschini.test(JM019_3, GAY017_2, threads = 4,
                                        seed = 273329, nPermute = 1000)
vsAM001_3 <- fasano.franceschini.test(JM019_3, AM001_3, threads = 4,
                                       seed = 273329, nPermute = 1000)
vsAM002_1 <- fasano.franceschini.test(JM019_3, AM002_1, threads = 4,
                                       seed = 273329, nPermute = 1000)

JM019_3vs <- data.frame(
  name = c("GAY016_1", "GAY016_2", "GAY017_1", "GAY017_2",
           "AM001_3", "AM002_1"),
  corrected.pval = p.adjust(c(vsGAY016_1$p.value, vsGAY016_2$p.value,
                              vsGAY017_1$p.value, vsGAY017_2$p.value,
                              vsAM001_3$p.value, vsAM002_1$p.value),
                            method = pval.cor.method),
  low = p.adjust(c(vsGAY016_1$conf.int[1], vsGAY016_2$conf.int[1],
                   vsGAY017_1$conf.int[1], vsGAY017_2$conf.int[1],
                   vsAM001_3$conf.int[1], vsAM002_1$conf.int[1]),
                 method = pval.cor.method),
  high = p.adjust(c(vsGAY016_1$conf.int[2], vsGAY016_2$conf.int[2],
                    vsGAY017_1$conf.int[2], vsGAY017_2$conf.int[2],
                    vsAM001_3$conf.int[2], vsAM002_1$conf.int[2]),
                  method = pval.cor.method))

JM019_3vs %>%
  ggplot(aes(x = name, y = corrected.pval, ymin = low, ymax = high)) +
  geom_pointrange() +
  labs(title = "JM019 area 3 samples vs all samples from footwear",
       y = "FDR (corrected p-value)") +
  geom_hline(yintercept=0.05, linetype = 2, colour = "darkred") +
  theme(axis.title.x=element_blank())
```



These results suggest that we can reject the null hypothesis for AM001 area 3, AM002 area 1, GAY016 area 1 and GAY017 area 2 and thus conclude that:

- the JM019 area 3 sample is not consistent with sharing a common origin with the AM001 area 3, AM002 area 1, GAY016 area 1 and GAY017 area 2 soils.
- the JM019 area 3 sample is consistent with sharing a common origin with the GAY016 area 2 and GAY017 area 1 soils.

2.5 Conclusions & Caveats

Integrating the conclusions from the various analyses presented here, I would conclude that:

- The JM019 area 1 replicates are from a heterogeneous sample with similarities to several of the known samples in this examination. This soil sample is most similar to GAY016 area 1 and GAY016 area 2, which are both soils from a similar location (the right boot of Mr Bayoh), and is consistent with sharing a common origin with soils from these locations.
- The JM019 area 2 replicates are from a relatively homogeneous sample that appears to be distinct from the other samples in this examination.
- The JM019 area 3 replicates are from a homogeneous sample with similarities to several of the known samples in this examination. This is most similar to GAY016 area 2 and GAY017 area 1, which are soils from the left and right boot of Mr Bayoh, and it is consistent with sharing a common origin with these soils. It is also similar to some GAY017 area 2 samples.

2.6 Commissioning

This independent analysis was commissioned by Professor Lorna Dawson, Head of Forensics at the James Hutton Institute, on 24th Sept 2022, on behalf of the Sheku Bayoh Public Inquiry. The report has been independently reviewed by Prof. Mark Brewer, Director of Biomathematics and Statistics Scotland.

2.7 Software package details

```
sessionInfo()
```

```
## R version 4.1.3 (2022-03-10)
## Platform: x86_64-conda-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.5 LTS
##
## Matrix products: default
## BLAS/LAPACK: /home/nick/anaconda3/envs/forensics_base/lib/libopenblasp-r0.3.21.so
##
## locale:
##  [1] LC_CTYPE=C.UTF-8       LC_NUMERIC=C           LC_TIME=C.UTF-8
##  [4] LC_COLLATE=C.UTF-8     LC_MONETARY=C.UTF-8    LC_MESSAGES=C.UTF-8
##  [7] LC_PAPER=C.UTF-8       LC_NAME=C              LC_ADDRESS=C
## [10] LC_TELEPHONE=C         LC_MEASUREMENT=C.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
```

```
##  [1] fasano.franceschini.test_2.1.0 ggiraphExtra_0.3.0
##  [3] zCompositions_1.4.0-1          truncnorm_1.0-8
##  [5] NADA_1.6-1.1                   survival_3.4-0
##  [7] MASS_7.3-58.1                  kableExtra_1.3.4
##  [9] knitr_1.40                     latex2exp_0.9.5
## [11] RColorBrewer_1.1-3             pcaMethods_1.86.0
## [13] Biobase_2.54.0                 BiocGenerics_0.40.0
## [15] pvclust_2.2-0                  readxl_1.4.1
## [17] forcats_0.5.2                  stringr_1.4.1
## [19] dplyr_1.0.10                   purrr_0.3.4
## [21] readr_2.1.2                    tidyr_1.2.1
## [23] tibble_3.1.8                   ggplot2_3.3.6
## [25] tidyverse_1.3.2
##
## loaded via a namespace (and not attached):
##  [1] nlme_3.1-159        fs_1.5.2            lubridate_1.8.0
##  [4] insight_0.18.4      webshot_0.5.4       httr_1.4.4
##  [7] tools_4.1.3         backports_1.4.1     sjlabelled_1.2.0
## [10] utf8_1.2.2          R6_2.5.1            DBI_1.1.3
## [13] mgcv_1.8-40         colorspace_2.0-3    withr_2.5.0
## [16] tidyselect_1.1.2    rematch_1.0.1       compiler_4.1.3
## [19] cli_3.4.1           rvest_1.0.3         xml2_1.3.3
## [22] labeling_0.4.2      scales_1.2.1        systemfonts_1.0.4
## [25] digest_0.6.29       rmarkdown_2.16      svglite_2.1.0
## [28] pkgconfig_2.0.3     htmltools_0.5.3     highr_0.9
## [31] dbplyr_2.2.1        fastmap_1.1.0       htmlwidgets_1.5.4
## [34] rlang_1.0.6         rstudioapi_0.14     farver_2.1.1
## [37] generics_0.1.3      jsonlite_1.8.0      googlesheets4_1.0.1
## [40] magrittr_2.0.3      Matrix_1.4-1        Rcpp_1.0.9
## [43] munsell_0.5.0       fansi_1.0.3         lifecycle_1.0.2
## [46] stringi_1.7.8       yaml_2.3.5          plyr_1.8.7
## [49] grid_4.1.3          sjmisc_2.8.9        ppcor_1.1
## [52] crayon_1.5.2        lattice_0.20-45     haven_2.5.0
## [55] splines_4.1.3       hms_1.1.2           pillar_1.8.1
## [58] uuid_1.1-0          reshape2_1.4.4      reprex_2.0.2
## [61] glue_1.6.2          ggiraph_0.8.2       evaluate_0.16
## [64] mycor_0.1.1         RcppParallel_5.1.5  modelr_0.1.9
## [67] vctrs_0.4.2         tzdb_0.3.0          cellranger_1.1.0
## [70] gtable_0.3.1        assertthat_0.2.1    xfun_0.33
## [73] broom_1.0.1         googledrive_2.0.0   viridisLite_0.4.1
## [76] gargle_1.2.1        ellipsis_0.3.2
```

3 CV

CV: Dr Nicholas Schurch **Principal Statistician for Environmental Science and Ecology**
Biomathematics and Statistics Scotland
[redacted] **https://www.bioss.ac.uk/people/nschurch.html**

Qualifications & Career:

2019-Present Principal Statistician for Environmental Science, BioSS.
2014-2019 Senior Bioinformatician (BBSRC). Dundee University, UK.
2012-2014 Research bioinformatician (BBSRC). Dundee University, UK.
2009-2012 Data Analysis Group bioinformatician (SULSA). Dundee University, UK.
2007-2008 UK-China Research Fellowship for Excellence Fellow (DIUS). Institute of High Energy Physics, Chinese Academy of Sciences, Beijing, China.
2004-2007 Research Assistant (PPARC). Durham University, UK.
2003-2004 Postdoctoral Research Scholar (NASA). Carnegie Mellon University, USA.
1999-2002 PhD (2002) Physics & Astronomy, University of Leicester, UK.
1995-1999 MSci (1999) Physics with Astrophysics, University of Bristol, UK.

Role and responsibilities:

- Leader of BioSS statistical consultancy in the area of Ecology & Environmental Science with responsibility for six experienced BioSS statisticians and co-supervising three PhD students, collaborating on projects across ecology, environmental and hydrological science.
- Senior research scientist with >20 years of experience applying robust statistical experimental design, statistical modelling, and modern data analysis and visualisation techniques to complex domain-specific datasets, spanning astrophysics to forensic science, and plant RNA molecular biology to environmental science.
- Member of BioSS Management Group, Horizon Scanning Group, Income Generation Group, and Research Scotland Repository JHI/BioSS open science committees.

Funding Track Record (Last 5 Years):

2022 UKCEH/Welsh Government ERAMMP “Methodological development for integrating Ground Survey and Satellite Habitat monitoring data” (£25k) .
2022 UKCEH/BioSS Framework for Statistical Methodology (flexible funding).
2022 SEPA/Marine Scotland Seafloor Biodiversity MeioMetBar technical report (£8k).
2021 Forensics statistics analysis report (£5k).
2021 NatureScot/BioSS Framework for Statistical Methodology (flexible funding).
2021 Scottish Government Crop Map technical report (£10k).
2020 Landmark, Leopard translocation technical review (£8k).
2019 SEFARI Gateway Responsive Opportunity Fund, “Waterwalls: Citizen stories of Scotland’s waters.” (£12k).

Research Interests: My work spans a wide range of scientific domains, with a correspondingly wide variety of methodologies including multivariate statistical modelling, latent variable analyses, structural equation modelling, Bayesian hierarchical models, deep learning neural networks, and time-series analysis. Recent statistical consultancy work includes technical reports for the Scottish Government on machine learning models for improving crop maps and yield estimates, for Marine Scotland on predicting ocean floor biodiversity from eDNA sediment composition analysis, modelling the environmental drivers of annual early cherry drop, and statistical analyses of soil physicochemical properties for forensics. I am also involved in statistical methodological and computational development work, including R package for Markovian mixture models for environmental time series and analysing 4th generation long-read DNA sequencing technology for eDNA, soil, and signal species metagenomics, and statistical methods for improving joint species distribution modelling approaches for biodiversity assessment. I have published 36 scientific papers, including highly-cited and high-profile papers (e.g., Schurch et. al. 2015, 641 citations to date) and 2 open-source bioinformatics tools (RATs, Froussios et. al. 2019; RoSA, Mourão et al 2019). I have presented at 6 international conferences, 6 national conferences, and invited seminars at more than 10 scientific institutions across the world. I am strongly committed to, and have a track record in publishing, Open Science and I am leading efforts to firmly embed this culture within BioSS and its parent organisation, the James Hutton Institute.

Selected Refereed Publications [Total publications: 36; 1st author:18; Total Citations: 2,761; h-index: 26; *joint first author, ^ joint corresponding author]:

1. Brooker, Brown, George, Pakeman, Palmer, Ramsay, Schöb, **Schurch**, Wilkinson, "Active and adaptive plasticity in a changing climate", 2022, Trends in Plant Science, doi:10.1016/j.tplants.2022.02.004
2. Parker* M., Knop K.*, Sherwood A.*, **Schurch, N. J.***, Mackinnon K., Gould P.D., Hall A, Barton G. J.^ & Simpson G. G.^, "Nanopore direct RNA sequencing maps an Arabidopsis N6 methyladenosine epitranscriptome", 2020, eLife, doi:10.7554/eLife.49658
3. Mourão K.*, **Schurch N. J.***, Lucoszek R., Froussios K., MacKinnon K., Duc C, Simpson G. G. and Barton G. J., "Detection and Mitigation of Spurious Antisense RNA-seq Reads with RoSA", 2019, F1000Research, doi:10.12688/f1000research.18952.1
4. Froussios K., Mourão K., Barton G. J.^, **Schurch N. J.^**, "Differential isoform abundance with RATs: a universal tool and a warning", F1000Research 2019, doi: 10.12688/f1000research.17916.1
5. Froussios K.*, **Schurch N. J.***, Mackinnon K., Gierlinski M., Duc C., Simpson G. G.^, Barton G. J.^, 2017, "How well do RNA-Seq differential gene expression tools perform in a complex eukaryote? A case study in Arabidopsis thaliana", Bioinformatics, doi: 10.1093/bioinformatics/btz089
6. **Schurch, N. J.***, Schofield P.*, Gierliński M.*, Cole C.*, Sherstnev A., Singh V., Wrobel N., Gharbi K., Simpson G. G.^, Owen-Hughes T.^, Blaxter M.^, Barton G. J.^, 2016, "How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?", RNA, 22, 6, 839-851, doi:10.1261/rna.053959.115.
7. Gierliński M.*, Cole C.*, Schofield P.*, **Schurch N. J.***, Sherstnev A., Singh, V., Wrobel N., Gharbi K., Simpson G. G.^, Owen-Hughes T.^, Blaxter M.^, Barton G. J.^, 2015, "Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment", Bioinformatics, 31, 22, 3625-3630, doi: 10.1093/bioinformatics/btv425.

Other Measures of esteem:

2022 Invited Talk: 9th Animal, Plant and Soil Traces (APST) Working Group of the European Network of Forensic Science Institutes (ENFSI): "How analyses and replication affect Likelihood Ratios"

2019-22: NERC & EPSRC grant reviewer.

2018 Invited talk, Nottingham University: "Rapid, cost-effective, data-driven, transcriptome annotation".

2018 Invited Talk: 'London Calling' International Oxford Nanopore Conference, London: "Redefining the Arabidopsis thaliana transcriptome with DRS sequencing".

2017 Invited talk, European Bioinformatics Institute, Cambridge: "Pinning down de-novo transcriptome assembly".

Public Engagement & Scientific Communication Track Record:

- 'Stutterer' Art Installation, Thompson & Craighead - LifeSpace (Dundee, 2014), The Lowry (Salford, 2016), Party Booby Trap (Caroll/Fletcher, 2016) & Festival of Ideas (Cambridge, 2016).
- ImpactFactor, Schurch & Cole - Symbiosis Art Exhibition (Dundee, 2015), ISMB/ECCB (Prague, 2017 - 1st prize for Art in Science).

4 Declaration

I, Dr Nicholas J E Schurch declare that:

1. I understand that my duty is to help the inquiry to achieve the overriding objective by giving independent assistance by way of objective, unbiased opinion on matters within my expertise, both in preparing reports and giving oral evidence. I understand that this duty overrides any obligation to the party by whom I am engaged or the person who has paid or is liable to pay me. I confirm that I have complied with and will continue to comply with that duty.
2. I confirm that I have not entered into any arrangement where the amount or payment of my fees is in any way dependent on the outcome of the report.
3. I know of no conflict of interest of any kind, other than any which I have disclosed in my report.
4. I do not consider that any interest which I have disclosed affects my suitability as an expert witness on any issues on which I have given evidence.
5. I will advise the party by whom I am instructed if, between the date of my report and the trial, there is any change in circumstances which affects my answers to points 3 and 4 above.
6. I have shown the sources of all information I have used.
7. I have exercised reasonable care and skill in order to be accurate and complete in preparing this report.
8. I have endeavoured to include in my report those matters, of which I have knowledge or of which I have been made aware, that might adversely affect the validity of my opinion. I have clearly stated any qualification to my opinion.
9. I have not, without forming an independent view, included or excluded anything which has been suggested to use by others including our instructing lawyers.
10. I will notify those instructing use immediately and confirm in writing if for any reason my existing report requires any correction or qualification.
11. I understand that: 11.1 my report will form the evidence to be given under oath or affirmation; 11.2 the inquiry may at any stage direct a discussion to take place between experts; 11.3 the inquiry may direct that, following a discussion between the experts, a statement should be prepared showing those issues which are agreed and those issues which are not agreed, together with the reasons; 11.4 I may be required to attend an inquiry hearing to give evidence to my report;

Dr Nicholas
Schurch

Digitally signed by Dr Nicholas Schurch
DN: cn=Dr Nicholas Schurch gn=Dr
Nicholas Schurch c=GB United Kingdom
l=GB United Kingdom o=Biomathematics
and Statistics Scotland ou=Principal
Statistician for Environmental Science and
Ecology [redacted]
Reason: I am the author of this document
Loca ion:
Date: 2022-11-03 10:23Z

[redacted]